The perception of scatterplots
Doherty, Michael E;Anderson, Richard B;Angott, Andrea M;Klopfer, Dale S
Perception and Psychophysics; Oct 2007; 69, 7; ProQuest
pg. 1261

Perception & Psychophysics
2007, 69 (7), 1261-1272

The perception of scatterplots

MICHAEL E. DOHERTY AND RICHARD B. ANDERSON
Bowling Green State University, Bowling Green, Ohio

ANDREA M. ANGOTT
University of Michigan, Ann Arbor, Michigan

AND

DALE S. KLOPFER
Bowling Green State University, Bowling Green, Ohio

Four experiments investigated the perception of correlations from scatterplots. All graphic properties, other than error variance, that have been shown to affect subjective but not objective correlation (r) were held constant. Participants in Experiment 1 ranked 21 scatterplots according to the magnitude of r. In Experiments 2 and 3, participants made yes/no judgments to indicate whether a scatterplot was high (signal) or low (noise). Values of r for signal and noise scatterplots varied across participants. Differences between correlations for signal and for noise scatterplots were constant in r in Experiment 2, and constant in r^2 in Experiment 3. Standard deviations of the ranks in Experiment 1 and d' values in Experiments 2 and 3 showed that discriminability increased with the magnitude of r. In Experiment 4, faculty and graduate students in psychology and sociology made point estimates of r for single scatterplots. Estimates were negatively accelerated functions of objective correlation.

The research literature on the perception of scatterplots is reviewed below. However, it is first necessary to distinguish between the two fundamentally different forms of judgments made in the research, discriminative judgments and absolute judgments, or point estimates. They can be operationally distinguished in a straightforward way. Discriminative judgments are those that call for the organism to make greater than/less than judgments, whereas absolute judgments call for participants to make point estimates on a scale.

Discriminative Judgments

In a study commonly cited as the first investigation of scatterplot perception, Pollack (1960) had participants make discriminative judgments about scatterplots presented on an oscilloscope. He described his scatterplots as a "dazzling dancing parade of correlation patterns" (p. 352). Using the classical psychophysical method of constant stimulus differences (Guilford, 1936, pp. 186ff), Pollack assessed two-choice difference thresholds in eight different experiments, or, in his terms, eight series, that varied the duration of the display, frequency range, size of the scatterplot, sign of the correlation represented, and characteristics of the pulses that generated the points. In each experiment, participants judged whether a scatterplot represented a higher or lower correlation than a reference scatterplot by pressing a lever up or down, and were provided immediate confirmation of the correctness or incorrectness of their judgments. The number of standard stimuli, or reference correlations, varied in given experiments from five to ten, with from four to six variable stimuli associated with each one. Threshold levels for 60%, 75%, and 90% correct were assessed for each reference correlation. The results were presented as the minimum difference between r^2 for the reference scatterplots and r^2 for the judged scatterplots required to achieve the desired level of discriminability.

Pollack (1960) interpreted the participant's task "as that of testing alternative statistical hypotheses" (p. 359), and found that the difference thresholds diminished markedly as the magnitude of the correlations increased. He suggested that "we might profitably search for an appropriate metric of confusability such as the d' measure of the theory of signal detectability" (pp. 359–360). He went on to speculate that r^2 might be the appropriate metric for the "representation of the visual scatterplot in discrimination space."

Pollack's suggestion that d' be considered as a possible metric for the confusability of scatterplots has not been pursued, nor have there been further investigations of the discriminability, as such, of scatterplots. There has, however, been considerable research, reviewed below, on point estimates of correlations exhibited in scatterplots. The first three experiments in the present paper were designed primarily to pursue Pollack's suggestion to investigate scatterplot confusability. Point estimates were collected in these three experiments, but were ancillary to the major aim of the experiments. The discriminability of scatterplots was assessed using both classical and contemporary methods, specifically, the methods of rank order (Guilford, 1936) in

M. E. Doherty, mdoher2@bgnet.bgsu.edu

Copyright 2007 Psychonomic Society, Inc.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Experiment 1, and of the theory of signal detectability, or TSD (Wickens, 2002), in Experiments 2 and 3.

Experiment 4 constituted a fundamental departure from the others in that participants made absolute judgments, that is, point estimates of the correlation coefficients represented on the scatterplots rather than discriminative responses. As noted above, Pollack's (1960) study constitutes, so far as we can discern, the sole literature on scatterplot discrimination per se. Hence the remainder of the introduction focuses on investigations that use a variety of direct estimates of the degree of relatedness shown in scatterplots.

Absolute Judgments

Early research on the perception of scatterplots, subsequent to Pollack, explored the function relating point estimates of the subjective correlation (denoted r_{est}) to the objective correlation (denoted r), and revealed a number of influences on the shape of that function. The first investigation to assess the direct estimation of correlation from scatterplots was that of Strahan and Hansen (1978), who found that people underestimated the magnitude of r, except when r equaled 0 or 1.0. Strahan and Hansen's scatterplots were such that the axes were unmarked and of equal length, and the variances of the two variables represented were equal. The spacing of the scatterplots was manipulated between participants; half seeing 13 scatterplots equally spaced in r and half seeing them equally spaced in r^2. Their participants were statistically sophisticated, and responded with direct, two decimal estimates of r. Spacing by r^2 diminished but did not eliminate underestimation. The degree of underestimation was related to the magnitude of r, such that the function relating r_{est} to r was positively accelerated. Strahan and Hansen speculated about possible models of scatterplot perception, a topic taken up by later investigators.

Wainer and Thissen (1979) presented participants with normally distributed scatterplot correlations and did not find the consistent underestimation that Strahan and Hansen (1978) reported. However, Wainer and Thissen's results are not comparable to other reports in the literature, as Wainer and Thissen provided their participants with a context for judgment—namely, four bivariate normal, prototype scatterplots with r values of 0, .25, .50, and .75—while the participants judged the experimental scatterplots. No description of the response required was provided, beyond the instructions to look at the four scatterplots as "prototypes and then estimate the correlation in each of the scatterplots in their packet" (p. 549).

Concerns about the possible practical implications of scatterplot perception led Bobko and Karren (1979) to pursue the issue of the misperception of scatterplots. Their statistically sophisticated participants provided direct numerical estimates of the correlation coefficients on scatterplots varying in range, apparent slope, and shape of the envelope of points (e.g., with and without outliers, approximately bivariate normal vs. distributions with the middle third of the data present or absent, and the presence or absence of a "twisted pear" characteristic). The term "twisted pear" was used to describe a distribution that was both radically nonlinear and heteroscedastic. There was a general tendency for participants to underestimate r that was most pronounced when $.20 \leq r \leq .60$, and this tendency was exacerbated when unequal axes produced unusual slopes.

Cleveland, Diaconis, and McGill (1982) manipulated the degree of association and the scale of the ordinate (hence, the size of the point cloud), and had statistically sophisticated participants judge the strength of linear X, Y relationship on a scale ranging from 0 (*no association*) to 100 (*perfect linear association*). The degree of association in the scatterplots was indexed as $w(r) = 1 - (1 - r^2)^{1/2}$, or one minus the coefficient of alienation. Significant and substantial underestimation was found. In addition, as the scale of the ordinate increased (which entailed a decrease in the size of the point cloud), the judged association also increased. Estimates were better fit by $w(r)$ than by r, but the investigators concluded that neither adequately describe judged association.

Lane, Anderson, and Kellam (1985) explored what they termed "the psychophysics of covariation detection" (p. 640). Noting that the Pearson r is based on three components—slope, the variance of X, and the variance of Y—they manipulated those components in a series of studies, using undergraduates and doctoral degree holders as participants. The stimuli were scatterplots with nine data points in each plot. Participants made subjective estimates of the degree of relationship by marking a point on a scale that ranged from 0 (*no relationship*) to 100 (*perfect relationship*). The slope, the variance of X, and the variance of Y significantly influenced judgments of relatedness, with the variance of Y having the largest effect.

Lane, Anderson, and Kellam (1985) did not explore in detail the issue of underestimation, but their participants' estimates of relatedness tended to be lower than the Pearson r and r^2 values for the scatterplots. For example a number of their experimental conditions contained scatterplots with $r = .78$, but the median of the mean estimates in those conditions was only .48. The finding that the different combinations of the slope, the variance of X, and the variance of Y yielded different estimates of covariation led the authors to express pessimism about the effort "to determine the psychophysical function relating values of Pearson's correlation to subjects' judgments of relatedness" (p. 648).

A study by Lauer and Post (1989), run on microcomputers, was similar in concept and results to that of Lane et al. (1985). The participants, who were statistically sophisticated, made estimates by using a mouse to select a point on a number line that was displayed below each scatterplot. The estimates were influenced by a number of factors including the objective r, the variance of X, the variance of Y, and the number of data points per scatterplot (200 or 400). There was significant underestimation of relatedness with respect to both r and of r^2, with the mean estimate being a sharply positively accelerated function of r. Lauer and Post's conclusion echoed that of Lane et al. in that the former investigators warned of "a host of factors which influence estimated correlation" (p. 244).

Yet another demonstration of underestimation was provided by Collyer, Stanley, and Bowater (1990), but their

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

procedure was sufficiently different from other investigators that meaningful comparisons are precluded. Their statistically naive participants fitted a line to the scatterplot points, by eye, and then estimated the *X*, *Y* relationship on a scale ranging from 0 to +1.0. The extent of underestimation in Collyer et al. can be gauged from the fact that 70% of the participants' estimates were lower than the objective r^2.

Two experiments relevant to the issues of underestimation and the influence of visual aspects of the scatterplot, as well as the effects of statistical training, were reported by Meyer and Shinar (1992). In addition to level of correlation, Meyer and Shinar manipulated scedasticity and the presence or absence of a regression line, with both novices and statistically sophisticated participants. In Experiment 1, participants assessed scatterplots, each of which contained 21 data points, by marking a point on a scale ranging from 0 to 100. There was significant and substantial underestimation, as well as effects of type of dispersion and the presence or absence of a regression line. These effects did not vary with the participants' level of prior statistical training, though experts tended to give higher estimates than did novices. A second experiment replicated the findings of Experiment 1 and also showed that the slope of the regression line influenced the judgments of sophisticated participants but not novices.

In interpreting their findings, Meyer and Shinar (1992) speculated that underestimation may reflect a bias in the mapping of numerical values of r and r^2 to the level of correlation exhibited in a scatterplot. They went on to suggest that one practical implication of such a bias is that if people are given a correlation coefficient from which they generate an imagined point cloud (which, conceptually, is the converse of what their study participants were asked to do), they may tend to imagine a scatterplot showing a stronger relationship than that indicated by the coefficient.

Two recent articles (Boynton, 2000; Meyer, Taieb, & Flascher, 1997) focused primarily on mathematical modeling of the perception of scatterplots. Their reviews of prior experimental findings led them to theorize that scatterplot perception might be influenced by the geometric characteristics of the displays. A unique aspect of the approach by Meyer et al. involved fitting functions to a variety of possible metrics for the regression residuals.

In Experiment 1, Meyer et al. (1997) varied r across the positive range, manipulated the presence/absence of a regression line, and used both novice and sophisticated participants. Each scatterplot stimulus contained 28 data points, and participants estimated the degree of relationship by assigning a two-digit number between 0 (*no correlation*) and 100 (*perfect correlation*). Meyer et al. concluded that their participants' data were best fit by an exponential function of the form $r_{est} = 1 - aX^b$, where X denotes the average distance from the least squares line of best fit, and a and b are free parameters. Their Experiments 2 and 3 were devoted to showing that defining X as perpendicular rather than the vertical residuals on which the Pearson is based led to better fits, and to showing the superiority of absolute differences rather than squared differences as inputs to the model. Two aspects of the results are of special relevance to the present review. One is that the participants' mean r_{est} values in Experiment 1 (see their Figure 2) were significantly *negatively* accelerated functions of r^2, whether regression lines were present or not, with r_{est} values *overestimating* r^2. This contrasts with the studies reviewed above that tended to demonstrate positively accelerated functions with substantial underestimation. However, the overestimation and negative acceleration did not appear in a similar condition in their Experiment 2. A second aspect of the results is the demonstration of the very large impact of the slope of the regression line on the level of estimation. When r^2 was, for example, .50, mean values of r_{est} varied from a high of .45 to a low of .17 as the slope varied from 25° to 55°.

Like Meyer et al. (1997), Boynton (2000) set out to model the perceptual dimensions of covariation estimation from scatterplots. Boynton approached the modeling task differently in that he tested the hypothesis that point cloud elongation was critical. He addressed the problem both by comparing r_{est} to the elongation ratio and by means of multidimensional scaling. Sophisticated participants estimated the degree of relationship on a 0 (*no relationship*) to 100 (*perfect linear relationship*) scale in each of a number of scatterplots. Each scatterplot contained 50 data points. The results showed a positively accelerated function relating r_{est} to r such that r_{est} consistently underestimated r. Boynton concluded that observers focus on the elongation ratio and on the standard error.

Limits on the Method of Absolute Judgment

The literature reviewed in this paper focuses on the relationship between actual and numerical estimates of correlation. But as noted above, such estimates vary widely across studies, and across conditions that are unrelated to the true magnitude of correlation (see, e.g., Bobko & Karren, 1979; Boynton, 2000; Cleveland et al., 1982; Lane et al., 1985; Lauer & Post, 1989). Moreover, as implied by Meyer and Shinar (1992), estimated magnitudes might be impacted not only by the accuracy of perception, but also by a scale bias in the mapping of graphical representations of correlation to numerical values of the correlation coefficient, or to the particular response scale used.

The Present Study

A fundamental purpose for studying the relationship between subjective and objective correlation is to determine the relative distances between representations of correlation in psychological space. Specifically, we want to ascertain whether representations that are equally spaced with respect to some objective measure of relationship give rise to equally spaced mental representations of relationship. In so doing, we wish to avoid the possibility that response biases could contaminate the measure of representational distance. Therefore, in Experiments 1, 2, and 3, we pursue Pollack's (1960) suggestion to use discriminability as means of studying the relation between subjective and objective correlation. The strategy for the present studies was to hold constant those factors required to test the psychophysical models described above, and to

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

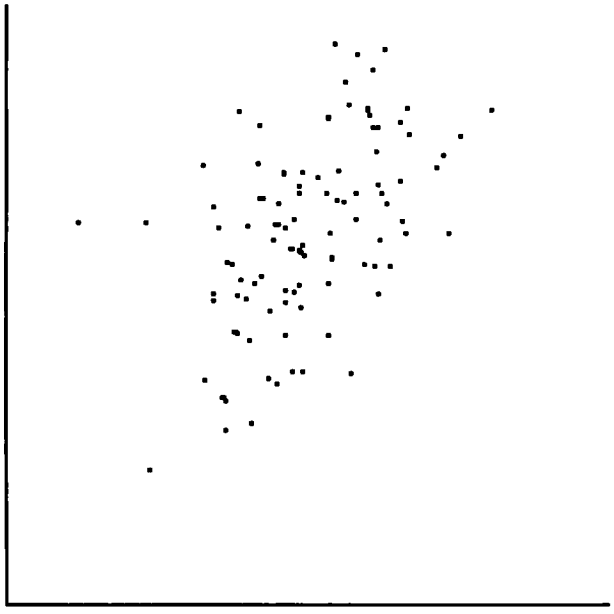

Figure 1. A sample scatterplot with $n = 100$ and $r = .50$.

manipulate only the level of correlation displayed in the scatterplot.

All participants in Experiments 1, 2, and 3 were graduate students in psychology and thus relatively sophisticated in the use of statistics. Experiment 1 used a classical psychophysical method. Specifically it involved the ranking of scatterplots representing different degrees of positive linear correlation, r. Experiments 2 and 3, in line with Pollack's (1960) recommendation, assessed discriminability using the d' measure of TSD. Experiments 2 and 3 used a yes/no paradigm, and were identical except for the spacing between signal, S, and noise, N, scatterplots. In Experiment 2, the differences between S and N scatterplots were in equal units of r, whereas in Experiment 3, the differences between S and N scatterplots were in equal units of r^2. An ancillary part of Experiments 1, 2, and 3 had participants make point estimates of correlation based on either one or two scatterplots. The fundamental prediction concerning discriminability for all three experiments was that that scatterplots corresponding to high values of r would be more discriminable than those corresponding to low values of r.

There are two bases for this prediction. First, the predicted pattern is an empirical generalization from Pollack's (1960) results, though his stimuli were not static scatterplots printed on paper, and his methods did not include TSD. Second, the same prediction can be derived from elementary psychophysical laws. A highly salient visual feature of a scatterplot is the variability reflected in the point cloud. If it is assumed that observers are judging that variability, then the difference in the amount of variation required for an observer to perceive a difference between two scatterplots is, by Weber's law, a direct function of the magnitude of the variation. Point clouds for low correlations show much more variability than point clouds for high correlations. Hence low correlations, when equally different in terms of the magnitude of r, should be more confusable, and thus less discriminable, than high correlations. This is essentially the argument advanced formally by Meyer et al. (1997), but the data on which they predicated their model were point estimates, not discriminative judgments.

EXPERIMENT 1

In the major part of Experiment 1, participants were given a randomly ordered set of scatterplots and instructed to rank them according to the magnitudes of the correlations represented. Thus, the standard deviations of the ranks provided a measure of plot discriminability.

Method

Participants. Twenty graduate students in psychology served as participants. All participants had taken courses in statistics or methodology as part of their undergraduate and graduate programs (the range was one to nine courses, and the median was five). An experimenter's error in handling the scatterplots made the data of 1 additional participant unusable.

Materials. The scatterplots were printed on standard 8½ × 11 in. paper. The axes were unlabeled, and were 135 mm in length. The number (m) of data points per scatterplot was either 9 or 100. Figure 1 shows a sample scatterplot with $m = 100$ and $r = .50$. The scatterplots were generated via an algorithm for generating a data set with a specified X, Y correlation. This algorithm involved correlating a standardized variable, X, with a second variable which was a composite of itself and a weighted, standardized residual from a regression of another variable, Y, on X. The scatterplots were printed on a laser printer, and had round data points about 1 mm in diameter. The generation algorithm was quite precise—producing, for example, for a scatterplot for $r = 0$, a value of -1.31×10^{-20}. This precision would not, of course, be possible had the stimuli been rounded numerical values rather than scatterplots. The only variability for scatterplots with the same value of r was due to the selection of points required to satisfy the algorithm and to whatever variability might have been introduced by the imprecision of the printer. Each scatterplot was unique.

Procedure. Each participant individually met the experimenter in his office. The participant first signed an informed consent sheet, then was shown two scatterplots, one with $m = 100$ and one with $m = 9$ data points. The participant was informed in writing that the two scatterplots were produced randomly such that the means and standard deviations of the X and Y variables were the same as each other and were the same on both scatterplots, and that the two correlations might be the same or different. There were four possible values of r: .30, .50, .70, and .90. The two scatterplots presented to each participant were selected randomly, with the restriction that the four values of r were equally represented across participants. The left–right spatial ordering of the two plots was randomly determined. The participant estimated the Pearson correlation by writing a decimal number on each scatterplot. Hence, in this ancillary, point estimation, phase of the experiment, each participant made just a single judgment on each of two scatterplots.

The participant was then taken into a seminar room in which four tables, each 0.61 m wide by 2.44 m long, were arranged in an approximately square configuration. On each of two opposing tables there were 21 scatterplots, with either $m = 100$ or $m = 9$ data points, arranged randomly in two parallel rows of 10 and 11, with the X-axes oriented toward the interior of the square. On the adjacent tables were 21 sheets of 8½ × 11 paper in two parallel rows, prominently numbered from 1 to 21. The r values represented by the scatterplots with both $m = 100$ and $m = 9$ ranged from 0.00 to 1.00, in steps of .05. Thus, this study employed the classical psychological scaling method of rank order (Guilford, 1936), which involves a set of serial discriminations.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 1
Individual and Mean r_{est} Values at Each Level of m and r, Experiment 1

	r (m = 9)				r (m = 100)			
	.30	.50	.70	.90	.30	.50	.70	.90
	.30	.50	.75	.90	.50	.45	.89	.65
	.72	.60	.55	.70	.67	.45	.78	.85
	.55	.30	.89	.90	.40	.70	.65	.69
	.30	.60	.70	.60	.40	.60	.70	.70
	.20	.20	.70	.70	.09	.65	.75	.90
Mean	.44	.45	.74	.79	.43	.58	.77	.78

Note—Each of the 20 entries for individuals for m = 9 indicates the judgment of 1 participant, as does each entry for m = 100. Each of 20 participants produced two point estimates, one for each of the two levels of m.

The participant was informed that the scatterplots all represented nonnegative correlations, that each correlation on a given table was different from the others, that the X and Y variables had equal means and equal standard deviations within a given scatterplot, and that means and standard deviations of X and of Y were equal across all scatterplots. No regression lines were shown. The participant was then told that the task was to rank the scatterplots from highest to lowest by transferring the scatterplots to the adjacent table, and placing the scatterplot with the highest correlation on the page numbered "1," the next highest on "2," and so forth. The participant was further informed that the scatterplots could be picked up and sorted through in any fashion. The participant, working from inside the square of tables, ranked the m = 100 scatterplots first, then ranked those with m = 9.

Having all participants rank the m = 100 scatterplots first was done with two considerations in mind: (1) We considered the m = 100 scatterplots to be much more representative of actual research practice, hence we wanted a purer assessment of discriminability for m = 100, and (2) differential discriminability at different levels of m was not a major goal of the study (if it had been then more than two levels of m would have been required). When the second task was completed, the participant was debriefed, paid $10, and thanked. Finally, the experimenter picked up the scatterplots in order, so that the participant's rankings could be subsequently recorded.

Results

The r_{est} values corresponding to the objective r values of .30, .50, .70, and .90 are presented in Table 1. Further discussion of these results will be deferred until the introduction to Experiment 4. The means and standard deviations of the ranks across participants are presented in Figures 2 and 3.[1] These discrimination data show considerable regularity. The correlation between r and the mean ranks for the scatterplots with m = 100, was −.99. In spite of the strong linear component represented by that correlation, there was a statistically significant quadratic component ($p < .01$), clearly visible in Figure 2. The same is true of the m = 9 data, with the Pearson r = −.98, but again with a significant quadratic component ($p < .001$) clearly visible in Figure 3.

The relation between the magnitude of r and the confusability of the scatterplots can be seen easily in the systematic variation of the standard deviations of the ranks in Figures 4 and 5, which are zero with high values of r but quite substantial with moderate and low values. This relation is indexed by the correlation between r and the standard deviation of the ranks. For the m = 100 scatterplots that correlation was −.87, and for those with m = 9, it was −.71. Significant quadratic components characterize both the m = 100 and m = 9 regressions of the standard deviations on r. Significance tests between statistical indices for the m = 9 and m = 100 scatterplots are not reported because the m = 100 scatterplots were ranked before the m = 9 scatterplots.

The data were also analyzed by participant. The correlations between the participant's ranks and the true ranks were high, with median correlations of .96 and .83 for m = 100 and m = 9, respectively. There was no correlation between the two sets of Fisher's Z values for the m = 100 and m = 9 judgments, presumably because of the relative lack of variability for the m = 100 data (r = .20, p = .39). That is, there was no evidence of a tendency for participants who performed well on the m = 9 scatterplots to perform well on the m = 100 scatterplots.

Some informal observations were made. For both conditions of m, every participant first moved the scatterplot representing r = 1.0 to the location on the adjacent table marked "1." About half the participants went in a strictly high-r to low-r order; about half went from r = 1 to approximately the middle of the r range, then skipped to r = 0 and worked up from there. The time for the ranking task, exclusive of the time it took to read the instructions,

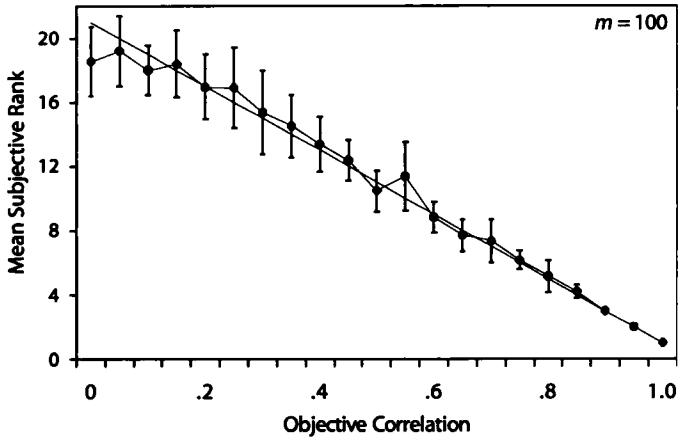

Figure 2. The relation between r and the means of the ranks in Experiment 1 for m = 100. Error bars are standard deviations.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

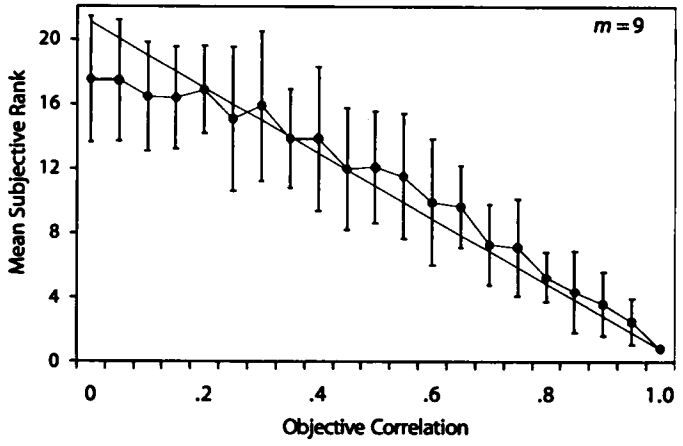

Figure 3. The relation between *r* and the means of the ranks in Experiment 1 for *m* = 9. Error bars are standard deviations.

was recorded for 17 of the participants. It ranged from 6–30 min, with a median of 15 min. There was no relation between time to completion and accuracy, as defined by the correlation between a participant's ranks and the true ranks. It was evident from the participants' behavior that they were making a series of paired comparisons. Almost all participants did some rearranging and subranking of the scatterplots on the table before moving them, often holding up and inspecting two at a time.

Discussion

The data indicate that scatterplots of low correlations were substantially more confusable than scatterplots of high correlations. The decision to have all participants rank the $m = 100$ scatterplots first makes formal significance tests inappropriate, but the mean of the standard deviations of the ranks for $m = 9$ is more than twice the corresponding value for $m = 100$. The variability in responding to the $m = 9$ scatterplots was so substantial as to call for some speculation. This result might be predicted qualitatively from an extension of Meyer et al.'s (1997) psychophysical model. We have removed the issue of the sampling variability of *r* by our procedure, but not the sampling variability of the data points. While the mean squared perpendicular distances from the regression line should be the same for different scatterplots of a given sample size, given that we have fixed the value of *r*, the mean absolute perpendicular difference scores will not be constant. The mean absolute perpendicular difference scores will vary much more within the $m = 9$ condition than within the $m = 100$, simply due to sample size differences.

In any actual usage of correlational statistics, the sampling variability of *r*, which was eliminated in this study, increases dramatically as the magnitude of the correlation diminishes. Hence, in practical contexts in which scatterplots representing sample *r* values are involved in making inferences about the population ρ value, the consequences of the confusability found with low correlations would be greatly exacerbated.

With regard to the psychological issue concerning the relation between the statistical relation displayed on the scatterplot and the mental representation of relationship, the manifest differential discriminability implies that high correlations, as represented in scatterplots, are further apart in psychological space than are low correlations. (Note: We make no claim concerning the question of whether the internal representations are abstract or graph-like in nature.) Although this investigation was concerned with confusability, the results also show the perception of correlation in a positive light. The very high correlations between the individual participant's ranks and the true ranks, especially for $m = 100$, reflect impressive performance with respect to the participants' discriminative abilities.

In light of the desirability of converging operations in support of theoretical claims (Garner, Hake, & Eriksen,

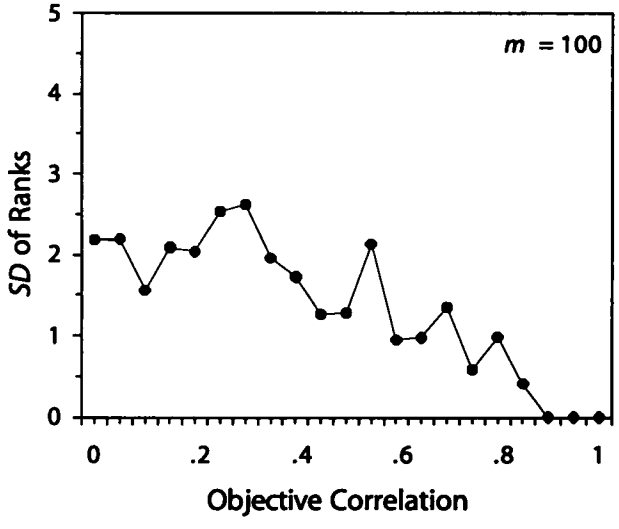

Figure 4. The relation between *r* and the standard deviations of the ranks in Experiment 1, *m* = 100. Each point is the standard deviation of 5 participants.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

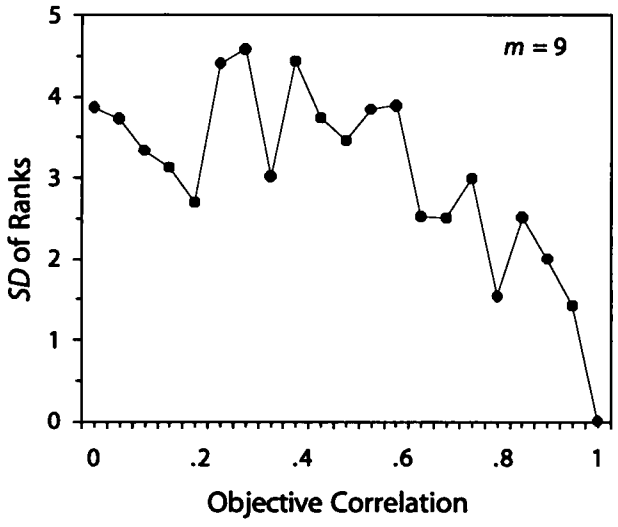

Figure 5. The relation between *r* and the standard deviations of the ranks in Experiment 1, *m* = 9. Each point is the standard deviation of 5 participants.

1956), a second experiment was performed using a very different set of operations, specifically the yes/no method of the theory of signal detectability.

EXPERIMENT 2

A more contemporary measure of confusability is the d' index of TSD. As noted above, in the earliest investigation of the perception of scatterplots, Pollack (1960) suggested using TSD to assess the confusability of scatterplots. Experiment 2 was designed to provide d' values across a range of objective correlations.

Method

Participants. The participants were 21 graduate students in psychology. As many as possible of the same students who participated in Experiment 1 were solicited to serve in Experiment 2, of whom 18 did so. Additional graduate students from the same participant pool were recruited to complete the design, and an extra participant was run as the call for participants was such that 21 names were solicited to give all 21 participants from Experiment 1 the opportunity to serve.

Materials and Design. Participants made yes/no judgments to discriminate between *S* and *N* scatterplots. The *r* values for the *S* plots were .34, .54, .74, and .94, and the *r* values for the corresponding *N* plots were .04 lower than those of the *S* plots. Hence, a constant Δr value of .04 was used for all four levels of correlation. There were 55 unique *N* scatterplots and 55 unique *S* scatterplots at each of the four levels. Each scatterplot was reproduced a sufficient number of times to yield material for 5 or 6 participants at each level of correlation. Each participant experienced just one of the four *S* versus *N* combinations. The algorithm for generating the scatterplots was the same as that used in Experiment 1. However, because it was necessary for the experimenter to be able to identify the individual scatterplots for scoring purposes, a code number was buried in a string of 40 random numbers, varying on each page, in small print at the bottom of the page.

Procedure. Participants came to the experimenter's office at an appointed time and were run individually, sitting across a desk facing the experimenter. They first signed an informed consent form and read the instructions. Then, to allow participants to become familiar with the stimuli prior to performing the task, they were shown 10 scatterplots and informed that there were 5 representing a higher correlation and 5 representing a lower correlation (though individual plots were not identified as such). They were instructed to examine but not respond to the 10 plots, and they could look at each of the 10 more than once if they wished. Next, participants were told that they would see a sequence of 100 scatterplots, one at a time, each showing the higher or lower of the two possible correlations, and that they should mark "Y" (for *yes*) or "N" (for *no*) in the upper corner of the page to indicate whether the plot did or did not represent the higher correlation. The scatterplot was turned over after a judgment had been made, and the participant was not permitted to look back at previously judged plots. The order of the 100 scatterplots was randomized individually for each participant. After the 100th scatterplot, the participant was asked to "estimate the value of the higher coefficient of correlation represented on the scatterplots." The participant was then debriefed, paid $10, and thanked.

Results

Table 2 presents the point estimates, discriminability index (d'), a measure of bias (c) from Macmillan and Creelman (1991), the hit rate (HR), and false alarm rate (FAR) for each participant, and the value of χ^2 calculated from the 2 × 2 table from which the HR and FAR were calculated, for each level of the correlation, *r*. Figure 6 shows the relationship of primary interest, that is, the relation between d' and *r*. The second-order polynomial regression of d' on *r* was such that $R = .74$ ($df = 18$, $p < .001$). A striking feature of the results is the marked individual differences in d', even to the extent that the data of 3 of the participants failed to reach even a marginal level of statistical significance by the χ^2 test of independence ($p > .30$). Given the evidence that these participants' responses were essentially randomly related to the stimuli, their data were dropped from the analysis. When the data

Table 2
The Point Estimate (r_{est}), Discriminability Index (d'), Bias Index (c), Hit Rate (HR), False Alarm Rate (FAR), and χ^2 As a Function of Noise and Signal Correlations (*N*, *S*) for Each Participant in Experiment 2

N, *S*	r_{est}	d'	c	HR	FAR	χ^2
.30, .34	.60	0.50	.45	.42	.24	3.66
.30, .34	.65	0.25	−.13	.60	.50	1.01
.30, .34	.70	0.41	.15	.52	.36	2.60
.30, .34	.70	0.10	.05	.50	.46	0.02
.30, .34	.65	0.57	.18	.68	.46	4.94
.50, .54	.56	0.57	−.24	.70	.48	5.00
.50, .54	.70	0.35	−.03	.58	.44	1.96
.50, .54	.70	0.66	−.03	.64	.38	6.76
.50, .54	.60	0.67	.08	.60	.34	6.78
.50, .54	.50	0.94	.06	.66	.30	12.98
.70, .74	.50	0.56	.08	.58	.36	4.86
.70, .74	.65	0.66	−.24	.72	.46	6.99
.70, .74	.60	1.23	−.03	.74	.28	21.17
.70, .74	.75	1.42	.06	.74	.22	27.08
.70, .74	.70	−0.05	−.08	.52	.54	0.04
.90, .94	.75	2.20	−.18	.90	.18	52.17
.90, .94	.45	1.00	−.14	.74	.36	14.59
.90, .94	.90	2.47	−.32	.94	.18	58.60
.90, .94	.70	1.85	−.15	.86	.22	41.22
.90, .94	.90	0.99	.03	.68	.30	14.46
.90, .94	.58	1.23	.03	.72	.26	21.68

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

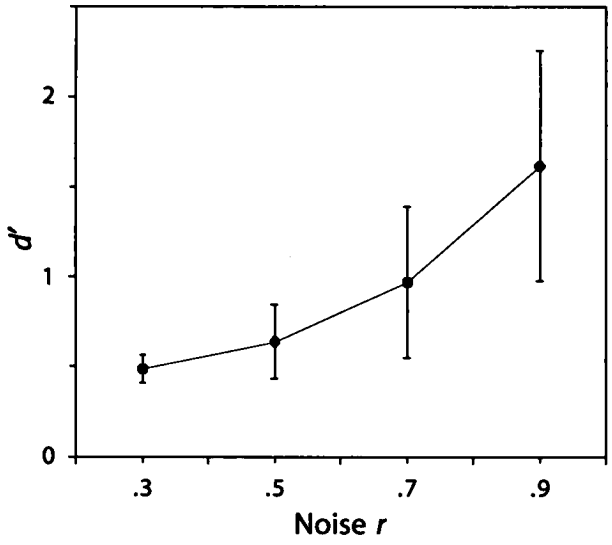

Figure 6. The relation between the mean d' and the Pearson r in Experiment 2 for participants with χ^2 values such that $p < .30$. Error bars are standard deviations.

were reanalyzed, the second-order polynomial regression of d' on r was such that $R = .77$ ($df = 15, p < .002$).

Discussion

The results of Experiment 2 are largely consistent with those of Experiment 1 in that both showed that discriminability increased significantly as the strength of association increased. In spite of the fact that the methods are, as noted, very different, they converge on a common conclusion: The discriminability, or its converse, confusability, of scatterplots are significantly influenced by the level of correlation represented thereon. The key index, d', its components FAR and HR, and bias as indexed by c are all sensitive to variations in level of r, when Δr is held constant.

EXPERIMENT 3

Recall the summary of the literature by Lewandowsky and Spence (1989), who stated that "the consistent finding was that people underestimate correlations over a wide range," and that r^2 is as good a basis for describing people's estimates as any other (pp. 220–221). Given that r^2, as a measure of relatedness, is considered by many to be at least as appropriate as r, Experiment 3 was conducted with the S and N scatterplots spaced in equal units of r^2 rather than in equal units of r.

Method

Participants. Twenty graduate students in psychology, none of whom had served in either Experiment 1 or 2 served as participants. Half of the students were from the same pool of advanced graduate students as in the experiments above, which exhausted the pool of willing participants. The remaining half were in their first semester of graduate school and were enrolled in a course in graduate level statistics and a course in methodology. The mean and standard deviation of the number of statistics and methodology courses previously completed by these ten graduate students were 2.3 and 1.55, respectively, with a range from 1 to 6.

Materials, Design, and Procedure. These were the same as in Experiment 2, except for the spacing of the S and N scatterplots. The Δr value of .04 in Experiment 2 was such that d' values could be obtained for all participants. A corresponding value for Experiment 3 was selected by using the midrange values of S and N of $r = .60$ and .64, respectively, as a standard. The difference between the squared correlations of $r = .60$ and .64 is .05, which was selected as the constant difference, Δr^2, between the S and N scatterplots for this experiment. Table 3 shows the paired S and N values of r for the four levels of correlation. The procedure for Experiment 3 was identical to that of Experiment 2.

Results

The results are presented in Table 3, and the function representing the relation between d' and r is shown in Figure 7. As is evident from Figure 7, the results of this experiment are qualitatively highly similar to those of Experiment 2. Given the data of all 20 participants, the second order polynomial regression yielded a correlation of .51 ($df = 17, p < .075$), but a striking feature of the results is the marked individual differences in d', even to the extent that the data of 6 of the participants failed to reach even a marginal level of statistical significance when the 2×2 tables representing their responses were subjected to a χ^2 test of independence ($p > .30$). When these participants were dropped from the analysis, the second-order polynomial regression of d' on r was such that $R = .77$, the same as in Experiment 2 ($df = 11, p < .007$). The alpha level of .30 was selected to ensure that we were dropping only participants who, for whatever reason, did not provide meaningful data. Had we set the alpha for dropping participants at .40 or .50, one more participant in each of Experiments 2 and 3 would have remained in the data set, and the pattern of significance would have been unchanged. Note

Table 3
The Point Estimate (r_{est}), Discriminability Index (d'), Bias Index (c), Hit Rate (HR), False Alarm Rate (FAR), and χ^2 As a Function of Noise and Signal Correlations (N, S) for Each Participant in Experiment 3

N, S	r_{est}	d'	c	HR	FAR	χ^2
.30, .374	.60	0.36	−.18	.64	.50	2.00
.30, .374	.47	0.05	.06	.48	.46	0.04
.30, .374	.45	0.85	−.22	.74	.42	10.51
.30, .374	.10	0.78	.14	.60	.30	9.09
.30, .374	.40	0.63	.13	.58	.32	6.83
.50, .548	.40	0.82	.00	.66	.34	10.24
.50, .548	.50	0.48	.34	.46	.28	3.48
.50, .548	.75	0.57	−.18	.68	.46	4.94
.50, .548	.34	0.56	−.08	.64	.42	4.86
.50, .548	.50	0.20	−.02	.54	.46	0.64
.70, .735	.70	1.33	−.08	.88	.44	21.57
.70, .735	.92	0.15	.08	.50	.44	0.36
.70, .735	.85	0.15	.02	.52	.46	0.36
.70, .735	.80	−0.11	−.46	.66	.70	0.18
.70, .735	.70	1.11	.03	.70	.28	1.11
.90, .927	.88	−0.05	.18	.42	.44	0.04
.90, .927	.90	1.01	.20	.62	.24	14.73
.90, .927	.80	2.35	.00	.88	.12	57.76
.90, .927	.70	1.09	.29	.60	.20	16.67
.90, .927	.80	2.28	−.14	.90	.16	54.96

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

that participants were dropped at every level of r. Again, treatment of the point estimate data is deferred to a section prior the introduction to Experiment 4.

Discussion

The d' data show clearly that discriminability increases sharply as the magnitude of the Pearson r increases. Hence, as with Experiments 1 and 2, Experiment 3 implies that high correlations are further apart in psychological space than are low correlations. The data of all three experiments confirm Pollack's (1960) original conclusion that discriminability varies directly with the magnitude of r. The d' results of Experiment 2 reflect an increase in the discriminability of r, confirming his speculation in that early study. The relations between the magnitudes of r and the SDs of ranks in Experiment 1, and between the magnitudes of r and d' in Experiment 2, entail the conclusion that the relation between the standard metric of correlation (i.e., the Pearson r) and the subjective representation of the relationship represented is decidedly nonlinear.

The error bars are clearly and unexpectedly higher in the cases with higher objective correlations. In this experiment, the confusability with low objective correlations appears to place an upper limit on the d' values. However, there is no practically similar lower limit associated with objectively high correlations. Any of a number of factors, including inattention, lack of motivation, lack of understanding of the task, and so forth, may operate to create low values of d' even when the S and N scatterplots are quite discriminable. It was certainly unexpected that several participants who were making discriminations between high actual correlations would produce data that did not even come close to traditional levels of statistical significance.

As stressed above, all prior research on scatterplot perception, except for the early work by Pollack (1960), has used some form of point estimate as the response measure. The psychophysical models based on that point estimate research (Boynton, 2000; Meyer et al., 1997) are both consistent with the results of the present study. The research reported herein was clearly not intended as a test of those models: Not only did the response scales differ, but characteristics of the scatterplots other than error variance that would provide tests of the models were held constant.

Point Estimation Data

When we included the point estimate responses in what were designed primarily as the discrimination studies reported as Experiments 1, 2, and 3, we fully expected to replicate the underestimation and positive acceleration commonly found in the literature. Examination of the individual r_{est} values in Tables 1, 2, and 3 reveals that our point estimate data are largely inconsistent with the conclusions in the literature cited. In no case did the function relating r_{est} to r look remotely like the expected positive linear acceleration, nor was there systematic underestimation, except when $r = .90$. In the $m = 9$ data of Experiment 1, when $r \leq .70$, there were six overestimates and four underestimates. In the $m = 100$, $r \leq .70$ data, there were nine overestimates and four underestimates. In Experiment 2, 9 of 10 participants overestimated r when $r < .74$, as did half of the participants in Experiment 3. These unexpected results led us to reexamine the literature with an eye toward possible explanations, and to see whether the failure to replicate the typical findings would occur with a new sample of participants who had not been involved in making discriminative responses. In order to maximize the response rate in a study in which no compensation would be offered, only a single scatterplot was used.

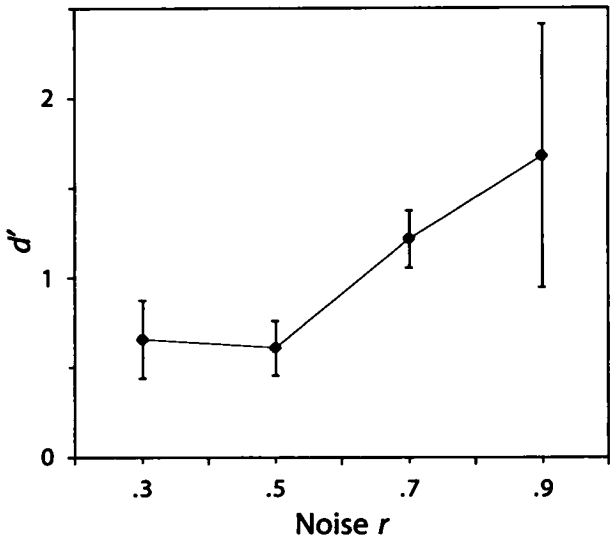

Figure 7. The relation between the mean d' and the Pearson r in Experiment 3 for participants with χ^2 values such that $p < .30$. Error bars are standard deviations.

EXPERIMENT 4

The literature reviewed in the introduction to this article notes that r_{est} based on scatterplots typically underestimates r, and that the function relating r_{est} to r is positively accelerated. Such underestimation may be highly consequential. One implication was noted by Bobko and Karren (1979) when they wrote "A theorist who obtains the statement '$r = .4$' from a computer may, in fact, be envisioning a display with validity in the range .5 to .6 . . ." (1979, p. 322). In a similar vein, Meyer and Shinar (1992) speculated that

> the consistent underestimation of correlation values indicates that when subjects receive a statistical estimate for the level of correlation between two variables (e.g., r^2), they imagine a point cloud that is denser (more highly correlated) than the one leading to the estimate. If, for example, the validity of a psychometric exam is reported as $r = .60$, subjects probably imagine a point cloud that corresponds to a much higher correlation level. . . . Thus the use of a correlation coefficient might cause overconfidence in the validity of the test and, accordingly, inappropriate decisions. (p. 346)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

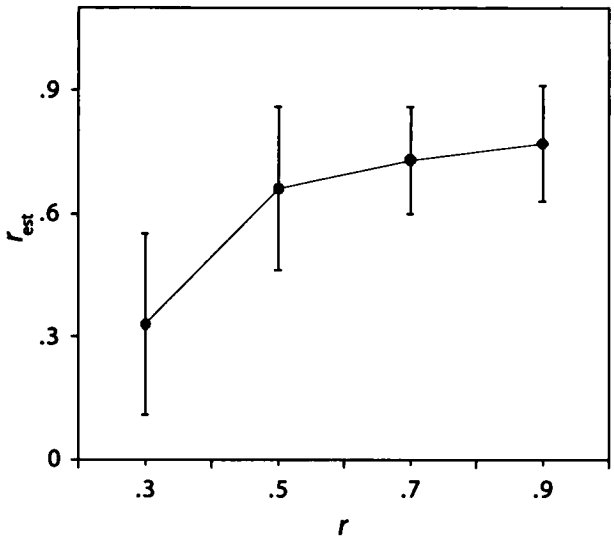

Figure 8. Values of r_{est} as a function of r in Experiment 4. Error bars are standard deviations.

Given that the results of Experiments 1, 2, and 3 failed to find such potentially consequential, systematic underestimation, except at high levels of r, Experiment 4 was undertaken.

Method

Participants. Faculty members and graduate students in the departments of psychology and sociology were solicited to serve as participants by placing packets in campus mailboxes. The numbers of responses and response rates were 15 (54%) for faculty and 20 (33%) for graduate students in psychology, and 8 (30%) for faculty and 15 (43%) for graduate students in sociology.

Materials and Procedure. A three-page packet with a Human Subjects Review Board consent form, a page of instructions, and a single scatterplot with an r value of either .30, .50, .70, or .90 was placed in the mailbox of each faculty member and graduate student solicited. Each participant was provided a unique scatterplot. All scatterplots had 100 data points, generated by the same algorithm and printed on standard 8½ × 11 paper, as in the above experiments.

The instructions informed the participant that the scatterplot was not a plot of behavioral data, but was a plot of X, Y pairs created on a computer such that the mean and variance of X equaled the mean and variance of Y. The participants were asked to write "F" for faculty or "G" for graduate student and the value of the correlation coefficient directly on the scatterplot page. Psychology participants were instructed to return it to the first author's mailbox, and to assure anonymity, to return the consent form to the second author's mailbox. Sociology participants were given virtually identical instructions, but returned the forms by placing them separately in two large envelopes in the sociology departmental office.

Results

First, the four subgroups were combined, and the data are shown in Figure 8. Performing a polynomial regression on the pooled data resulted in an overall $R = .71$. The best-fitting quadratic equation was $Y' = 2.95X - 1.87X^2 - 0.38$, with the linear and quadratic components both being significant at $p < .01$.

None of the faculty had been participants in Experiments 1, 2, or 3, but it is highly likely that many of the graduate students in psychology had been. To deal with the possibility of confounding by prior experience with the research, the point estimates for faculty were analyzed separately. Performing a polynomial regression on the data for faculty resulted in an $R = .84$. The best-fitting quadratic equation was $Y' = 4.19X - 2.92X^2 - 0.73$, with the linear and quadratic components both being significant at $p < .01$.

Discussion

These data are inconsistent with the generalizations that people underestimate correlations and that the function relating r_{est} to r is positively accelerated. The r_{est} values were a significantly *negatively* accelerated function of r. Participants were, on the average, relatively accurate at estimating $r = .30$, but *overestimated* midrange correlations and underestimated high ones. For actual r values of .30, .50, .70, and .90, the overall median r_{est} values were .33, .66, .73, and .77, respectively. These results are highly consistent with the point estimate data of the first three experiments. The results of all four investigations reported herein are inconsistent with those reported in the literature.

The consistency of the results reported herein, the consistency of the results reported in the literature, and the inconsistency between them led to a reexamination of the literature with an eye toward possible explanations. We propose one tentative possibility. Methodological variations abound in the studies cited, but there is one feature common to all the investigations reviewed, which is not characteristic of our experiments. All of the studies cited presented participants with multiple scatterplots, ranging from 13 (Strahan & Hansen, 1978) to 351 (Boynton, 2000, Experiment 2). Of all the multiple-scatterplot studies cited, several of which report more than one experiment, only Meyer et al.'s (1997) Experiment 1 reported results that are inconsistent with the accepted view that r_{est} underestimates and is a positively accelerated function of r. Their results are like those of the present study in one salient way, that being the finding of a negatively accelerated relation between r_{est} and r. The one salient methodological difference between the present study and those in the prior literature is that participants in the present study judged a single scatterplot whereas those in prior studies judged many scatterplots. One possible explanation for the difference between the present results and those typically reported is that judgments of correlation on scatterplots are influenced by immediately preceding judgments. There are other possible explanations. The scatterplot stimuli have been presented in many ways and a variety of response scales have been used, and the resulting function form may be influenced not only by any one of these but by some combinations of them, as well.

GENERAL DISCUSSION

The ideas in the opening paragraph of this article bear repetition: Point estimates, however assessed, and discriminative measures are operationally different from one another. The results presented above show that these two classes of measures are also quite different behaviorally, given that the present investigations led to fundamentally different conclusions concerning the relation between

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

judgments and scatterplots, depending on the nature of the response measure.

Why might behavioral differences be expected, depending on whether point estimates or discriminative judgments are employed? In some ways the two kinds of judgments are very different, the first having an important cognitive component in addition to the obvious perceptual component. Asking people to make point estimates, on the other hand, appears to rely on the extent to which people have good prototypes for scatterplots of different r values. This argument recalls Meyer and Shinar's (1992) speculation that there may be a bias in the mapping of numerical values of r and r^2 to the level of correlation exhibited in a scatterplot.

Experiments 1, 2, and 3 clearly confirm Pollack's (1960) conclusion that pairs of scatterplots reflecting similar, but different, high correlations are less confusable with one another than are pairs of scatterplots reflecting similar, but different, low correlations. Thus, assessments of discriminability by both classical and contemporary methods show clearly that participants are more sensitive to differences at higher levels of relatedness, whether the differences between adjacent scatterplots are measured in Δr (Experiments 1 and 2) or Δr^2 (Experiment 3).

Experiment 4 revealed a very different function relating r_{est} to r, at least for the conditions of this particular investigation. An obvious possible explanation for the negatively accelerated function may be the boundedness of the response scale on to which the perceptual representation of relatedness must be mapped. The errors associated with a scatterplot with a .90 correlation can range much farther below .90 than they can range above it. The converse is true for scatterplots with a correlation of .30. Inspection of the error distributions shows that this is, in fact, the case. Thus, the negative acceleration may be attributable to a form of regression effect.

The reason for this discrepancy from previous literature that also used point estimates is unclear, but the difference may be due to one or more of the specific features of scatterplots that have been shown to affect results. This speculation highlights the fact that we have found this negative acceleration using an experimental strategy of holding all those variables constant that have been shown to influence inferences from scatterplots.

Numerous influences on the perception of correlation from scatterplots have been documented. Lane et al. (1985) manipulated slope, error variance, the variance of X, and presentation format, and found that scatterplots with the same value of r led to different values of r_{est}. Eade (1967; cited in Lane et al., 1985) found that increasing the size of the scales, and thus manipulating the density of the point cloud, affects r_{est} (Cleveland, Diaconis, & McGill, 1982). Effects of outliers, differences in slope, heteroscedasticity, restriction of range, and missing data have been shown by Bobko and Karren (1979). Lauer and Post (1989) found effects of the standard deviations of X and of Y, the number of data points and even which computer was used to present the scatterplots. An obvious candidate for which influences have also been demonstrated is expertise (Bobko & Karren, 1979; Cleveland et al., 1982; Collyer et al., 1990; Lane et al., 1985; Meyer & Shinar, 1992; Strahan & Hansen, 1978; Wainer & Thissen, 1979), but the results are mixed. Note that the generalizability of the conclusion regarding the superior discriminability of scatterplots that reflect high correlations is restricted to the comparisons in which scatterplot features other than error variance are themselves comparable. It might well be possible to design scatterplots, taking advantage of one or more of the previously demonstrated influences on scatterplot perception, that would violate our conclusion.

The commonly accepted view of the relation between r_{est} and r appears to reflect an overgeneralization. The claim here is not that the earlier experiments have been wrong, but that, in addition to all of the influences just cited, context may be yet another influence on the relation between r_{est} and r. Wainer and Thissen's results, discussed briefly above, implicate a context effect, as does the distinction between theory-based and data-based inferences (Jennings, Amabile, & Ross, 1982).

There are two practical implications, predicated on the whole literature on the perception of correlation from scatterplots as well as on the results presented in this paper. The first is that it would be desirable for further standardization of publication conventions in the presentation of scatterplots (see also Meyer & Shinar, 1992). Ideally, recommendations for standardization ought to be based on empirical research that shows close correspondences between perceptual estimates of relatedness and statistical indices thereof. It seems clear, however, that the myriad of factors influencing such judgments, some of which have been explored herein, suggest that pursuit of such research would be prohibitively costly in terms of time and effort. Clarity of scientific communication would, we believe, be enhanced if editors moved toward further standardization of what are now essentially arbitrary, current practices based on varying conventions.

The second is yet one more call for investigators to look at their data. It has long been preached that investigators ought to examine scatterplots because true relationships might be missed or underestimated because default r values are often based on the assumptions of linearity and homoscedasticity. The research discussed above suggests that even linear relationships are more likely to be misinterpreted in the absence of scatterplots. In short, scatterplots and r values are complementary sources of intuitions about relationships, and neither should be interpreted alone.

AUTHOR NOTE

The authors thank David M. Boynton and Joachim Meyer for substantial comments and suggestions on drafts of the manuscript. This research was supported in part by a grant from the National Science Foundation. Correspondence concerning this article should be addressed to M. E. Doherty, Department of Psychology, Bowling Green State University, Bowling Green, OH 43403 (e-mail: mdoher2@bgnet.bgsu.edu).

REFERENCES

BOBKO, P., & KARREN, R. (1979). The perception of Pearson product moment correlations from bivariate scatterplots. *Personnel Psychology*, **32**, 313-325.

BOYNTON, D. M. (2000). The psychophysics of informal covariation as-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

sessment: Perceiving randomness against a background of dispersion. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 867-876.

CLEVELAND, W. S., DIACONIS, P., & MCGILL, R. (1982). Variables on scatterplots look more highly correlated when the scales are increased. *Science*, **216**, 1138-1141.

COLLYER, C. E., STANLEY, K. A., & BOWATER, C. (1990). Psychology of the scientist: LXIII. Perceiving scattergrams: Is visual line fitting related to estimation of the correlation coefficient? *Perceptual & Motor Skills*, **71**, 371-378.

GARNER, W. R, HAKE, H. W., & ERIKSEN, C. W. (1956). Operationism and the concept of perception. *Psychological Review*, **63**, 149-159.

GUILFORD, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.

JENNINGS, D., AMABILE, T. M., & ROSS, L. (1982). Informal covariation assessment: Data-based vs. theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 211-230). New York: Cambridge University Press.

LANE, D. M., ANDERSON, C. A., & KELLAM, K. L. (1985). Judging the relatedness of variables: The psychophysics of covariation detection. *Journal of Experimental Psychology: Human Perception & Performance*, **11**, 640-649.

LAUER, T. W., & POST, G. V. (1989). Density in scatterplots and the estimation of correlation. *Behaviour & Information Technology*, **8**, 235-244.

LEWANDOWSKY, S., & SPENCE, I. (1989). The perception of statistical graphs. *Sociological Methods & Research*, **18**, 200-242.

MACMILLAN, N. A., & CREELMAN, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.

MEYER, J., & SHINAR, D. (1992). Estimating correlations from scatterplots. *Human Factors*, **34**, 335-349.

MEYER, J., TAIEB, M., & FLASCHER, I. (1997). Correlation estimates as perceptual judgments. *Journal of Experimental Psychology: Applied*, **3**, 3-20.

POLLACK, I. (1960). Identification of visual correlational scatterplots. *Journal of Experimental Psychology*, **59**, 351-360.

STRAHAN, R. F., & HANSEN, C. J. (1978). Underestimating correlation from scatterplots. *Applied Psychological Measurement*, **2**, 543-550.

WAINER, H., & THISSEN, D. (1979). On the robustness of a class of naive estimators. *Applied Psychological Measurement*, **3**, 543-551.

WICKENS, T. D. (2002). *Elementary signal detection theory*. Oxford: Oxford University Press.

NOTE

1. In light of concerns about scaling qualities, all analyses in Experiment 1 were also done with medians and Q, the semi-interquartile range. No conclusions would change, hence those analyses are not reported here.

(Manuscript received February 10, 2006;
revision accepted for publication April 20, 2007.)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.