Psychonomic Bulletin & Review
2001, 8 (4), 847-850

Interpretation of significance levels by psychological researchers: The .05 cliff effect may be overstated

JACQUES POITEVINEAU
Centre National de la Recherche Scientifique, Paris, France

and

BRUNO LECOUTRE
Centre National de la Recherche Scientifique and Université de Rouen, Rouen, France

Comments about previous studies indicate that the interpretation of significance levels by psychological researchers is unequivocally dictated by a binary decision-making framework. In particular, confidence in a p level would drop abruptly just beyond the fateful .05 level (*cliff effect*). A replication of Rosenthal and Gaito's (1963) experiment on the degree of confidence in p levels shows that these claims should be moderated. Detailed analysis of individual curves reveals that the attitude of researchers toward p values is far from being as homogeneous as might be expected. However, most psychological researchers in our study rated graduated confidence judgments, as either exponential or linear. Only a minority of *all-or-none* respondents exhibited an abrupt drop in confidence.

In 1963, Rosenthal and Gaito carried out one of the first experimental studies on the use of null hypothesis significance tests. They asked psychological researchers and graduate students "to rate their degree of belief or confidence in a variety of p levels" on a 6-point scale (from *extreme confidence or belief*, 5, to *complete absence of confidence or belief*, 0). Surely, this should be understood as the degree of belief in some research hypothesis, given the sample size and a p value associated with the appropriate statistical test. Fourteen p values ranging from .001 to .90, associated with two different sample sizes ($n = 10$ and $n = 100$), were considered. The degree of confidence of both the researchers and the students was found to be approximately a decreasing exponential function of the p level, and for any p level the larger sample size always gave rise to more confidence. The authors concluded that both Type I and Type II errors were used as criteria of belief by the subjects, Type I errors explaining the decrease in confidence when p levels increase, and Type II errors explaining the influence of sample size. These results were confirmed in subsequent studies (Beauchamp & May, 1964; Minturn, Lansky, & Dember, 1972, quoted by Nelson, Rosenthal, & Rosnow, 1986). Furthermore, Rosenthal and Gaito (1963, 1964) argued in favor of the existence of a *.05 cliff effect*—that is, an abrupt drop in confidence in a p level just beyond the .05 point—and came to the conclusion that it was an effect of the emphasis on the .05 level. This result was invoked by Oakes (1986, p. 83) in support of his *significance hypothesis*, according to which the outcome of the significance test is interpreted in terms of a dichotomy: An effect either *exists* when it is significant or *does not exist* when it is nonsignificant. A similar interpretation was made by Nelson et al. (1986), who concluded that "this early study also provided preliminary evidence that research decisions to believe or not to believe (accept not accept) the null hypothesis are made in a binary manner based simply on whether p does or does not reach the .05 level" (p. 1299). Nevertheless, it should be stressed that the cliff effect reported by Rosenthal and Gaito (1963) was of relatively moderate magnitude and was of any consequence only for the data from student subjects. Moreover, in a series of other experiments, Lecoutre (1983, 2000) observed a *gap* between the spontaneous comments made by researchers about the data and the dichotomous reject–accept decisions that resulted from significance tests. She also noted that the interpretation of test results could vary considerably from one individual to another. This led us to replicate the Rosenthal and Gaito study and to investigate individual responses. Our aim was to identify distinct categories of subjects, possibly corresponding to different conceptions of statistical inference, referring in particular to Neyman–Pearson, Fisher, and Bayes. With this purpose in mind, the instructions were made more explicit than those in the original experiment. It was specified that confidence was related to *the effect of the experimental treatment*, and the subjects were asked to state their degree of confidence in the hypothesis that the "experimental treatment really has an effect."

Our special thanks go to Victoria Bishop for her editorial help. Correspondence concerning this article should be addressed to J. Poitevineau, LCPE, INaLF, CNRS., 44 rue de l'Amiral Mouchez, 75014 Paris, France (e-mail: jacques.poitevineau@ivry.cnrs.fr).

METHOD

Subjects

The subjects were 18 psychological researchers from various universities in France, all with a Ph.D. and all with practical experience processing experimental data.

Procedure

The subjects carried out the task individually. In order to clarify the meaning of the sample size, it was specified that the test was Student's *t* test for paired groups. Twelve *p* values (.001, .01, .03, .05, .07, .10, .15, .20, .30, .50, .70, .90) crossed with two sample sizes ($n = 10$ and $n = 100$, as in the original experiment) were presented at random on separate pages of a notebook. The subjects were asked to mark a point on a nongraduated 1-dm (i.e., 10-cm) line corresponding to their degree of confidence, from *zero confidence* (0, left extremity of the line) to *complete confidence* (1, right extremity).

RESULTS

The subjects perceived the task as easy and quickly completed it (about 5 min on average, and never more than 10 min). When asked to comment on the task, not one researcher professed to having knowledge of previous studies or referred to the existence of a cliff effect. The average curves, plotted in Figure 1A, were rather similar to those in Rosenthal and Gaito's (1963) article, with a degree of confidence that was always greater for $n = 100$ than for $n = 10$. A .05 cliff effect was apparent for the two sample sizes. Nevertheless, these average curves were fairly well fitted by an exponential function ($r^2 = .925$ for $n = 10$ and $r^2 = .949$ for $n = 100$). The standard errors for means varied from .011 to .082. The degree of confidence given for *p* values greater than or equal to .10 was greater than that in the original experiment (where it was 0 for $p = .70$ and $p = .90$). However, this probably reflected the effects of using a continuous scale. Although our experiment was conducted about 35 years after the original one and in another country, the (average) results appear to be similar. This is hardly surprising if we take into account the fact that the teaching and the practice of statistical inference methods in psychology are largely standardized and, in addition, have changed little since the original experiment was performed (Schmidt, 1996).

Although previous reports paid little or no attention to individual differences, we examined the individual data and found them to be qualitatively heterogeneous. Three distinct categories of functions could be easily identified: (1) a decreasing exponential curve, (2) a negative linear curve, and (3) an all-or-none curve representing a very high degree of confidence when $p \leq .05$ and quasi-zero confidence otherwise. Therefore, three models, each with two parameters (exponential, $y = \exp(a + bp)$; linear, $y = a + bp$; all-or-none, $y = a$ if $p \leq .05$, $y = b$ otherwise), were fitted to each of the two curves ($n = 10$ and $n = 100$) for each subject. Then the curves were classified according to the model yielding the greatest r^2. In all cases, the two curves belonging to each subject fell into the same category. Consequently, the 18 subjects were classified as follows (see Figures 1B, 1C, and 1D):

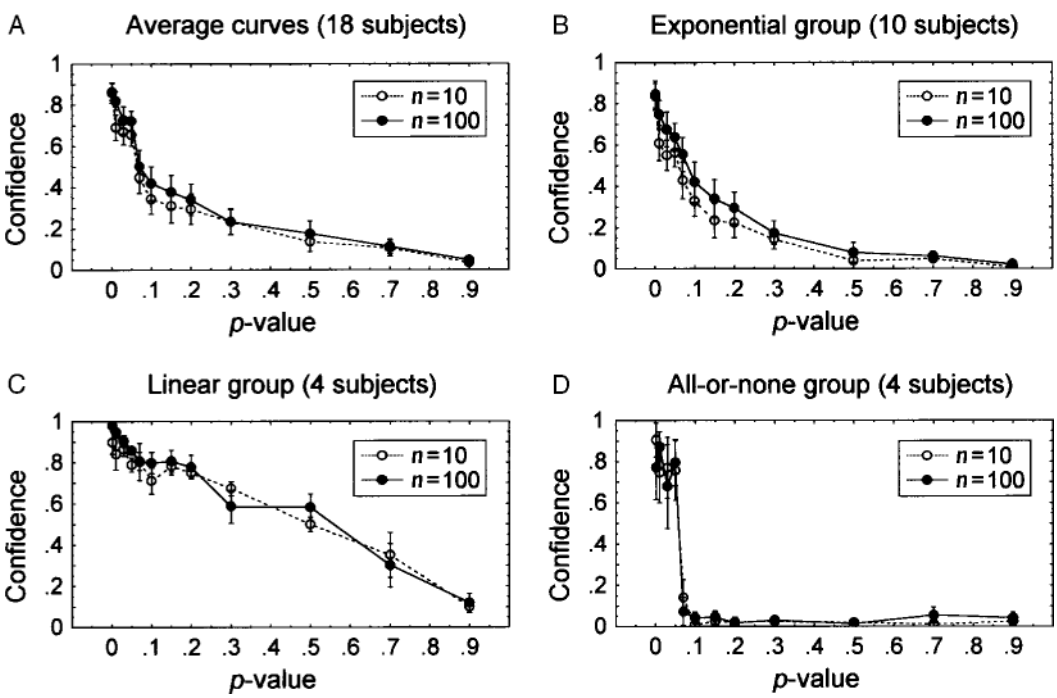

Figure 1. Confidence in the effect of an experimental treatment as a function of the *p* value and sample size *n*. Average ratings (with *SEM* error bars) are displayed for the whole group of subjects and for each of the three identified subgroups.

There were 10 subjects with decreasing exponential curves (average adjusted curve, $r^2 = .961$ for $n = 10$ and $r^2 = .990$ for $n = 100$), 4 subjects with negative linear curves ($r^2 = .970$ for $n = 10$ and $r^2 = .966$ for $n = 100$), and 4 subjects with all-or-none curves ($r^2 = .981$ for $n = 10$ and $r^2 = .985$ for $n = 100$).

According to Nelson et al. (1986), a test for the .05 cliff effect "controlling for the ordinary decline in confidence as p increases" can be based on the cubic contrast assigned to the four consecutive p values .03, .05, .07, and .10. The reason for this procedure is that it is equivalent to testing whether the patterns of the corresponding parent means and the predicted means based on a second-degree polynomial equation are the same (*no cliff effect*). Consequently, a "natural" measure of the strength of the cliff effect is given by an index of departure. Given the second-degree polynomial equation that best fits the data for the considered p values, the (raw) cliff effect size can be estimated by the quadratic mean of the residuals between the observed and the predicted means. This index is equal to the (absolute) numerical value of the usual cubic contrast, up to a constant of proportionality. Taking into account the unequal spacing of the four p values and introducing the appropriate constant of proportionality, the cliff effect size is given here by the coefficients of the cubic contrast −.1310, +.3668, −.3057, and +.0699. The sign of this contrast being arbitrary (but relevant), it was chosen so that a positive value represented a cliff effect. For instance, in the exponential group, the observed .05 cliff effect size was 0.016 dm, quite small.

The observed cliff effect sizes and their 95% confidence intervals were respectively: +.016 [−.010,+.043] for the exponential group, −.007 [−.049,+.035] for the linear group, +.159 [+.117,+.201] for the all-or-none group,[1] and +.043 [+.023,+.063] for the average curve (all subjects). The corresponding standardized cliff effect sizes, as measured by the r statistic, and their 95% confidence intervals were the following: +.321 [−.184,+.649] (exponential), −.094 [−.515,+.383] (linear), +.902 [+.763,+.948] (all-or-none), and +.766 [+.462,+.879] (whole set). It was obvious that the minority all-or-none group was largely responsible for the substantial average cliff effect.

The standardized .05 cliff effect size obtained by Nelson et al. (1986) was $r = .34$, far smaller than that for the set of subjects studied here (.77) and almost equal to that for the present exponential group (.32). If the variability of their data was, at most, equal to the variability of our data, the raw effect found by Nelson et al. was, at most, of the same magnitude as the one found for the present exponential group and was, therefore, moderate. The Rosenthal and Gaito (1963) functions appear to be more similar to those of the present exponential group than to those for the present group average. Furthermore, on the basis of Rosenthal and Gaito's (1963) average curves, we estimated the average scores associated with .03, .05, .075, and .10 p values. In this case, the appropriate coefficients for the cubic contrast were −.1393, +.3510, −.3120, and +.1003. The raw estimate of the .05 cliff effect found by Rosenthal and Gaito (1963; expressed in units of the present 0–1 scale) was .015 for the researchers and .018 for the students. Thus, their average cliff effect was also small. Comparison with other studies was impossible because of the lack of sufficient information. Therefore, the hypothesis that the samples from previous studies consisted of a great number of all-or-none subjects is not likely to be correct.

The role of sample size was only noticeable in the exponential group (and consequently in the averaged data). The observed raw effect sizes (the difference between the mean for $n = 100$ and the mean for $n = 10$) and their 95% confidence intervals were +.070 [+.012,+.128] (exponential), +.033 [−.058,+.124] (linear), −.002 [−.093,+.090] (all-or-none), and +.046 [+.003,+.089] (whole set). The raw effect sizes and their 95% confidence intervals for interaction between the cliff effect and the sample size were −.021 [−.059,+.017] (exponential), +.028 [−.033,+.088] (linear), +.049 [−.011,+.109] (all-or-none), and +.005 [−.023,+.034] (whole set).

CONCLUSION

The identification of three clearly distinct categories of subjects shows that researchers' confidence in p levels does not reflect monolithic statistical knowledge. Although we would not claim that researchers' judgments are each based on a particular statistical model, some connections may be made. Only the stepwise curves of the third (all-or-none) minority group clearly suggest a decision-making criterion of the Neyman–Pearson (1933) variety. However, an analysis consistent with Neyman–Pearson would also lead to considering the probability of a Type II error. Consequently, the effect of the sample size would depend on whether the p value was below or above the α level (Royall, 1986). Since this is not the case here, it is not clear whether these subjects are really concerned with Type II errors. On the contrary, all the other subjects expressed *graduated* confidence judgments about p values, which were either exponential or linear. The results for the second (linear) group appear compatible with the common misinterpretation of a p value as the complement of the probability that the alternative hypothesis is true, which Carver (1978) called the *valid research hypothesis fantasy.* Moreover, from another point of view, the functions for these subjects correspond to the standard Bayesian or the Fisher fiducial reinterpretation of one-sided p values (e.g., Casella & Berger, 1987; Rouanet, 1998). The exponential curves of the first group are, in fact, similar to the type of curve often obtained in psychophysics experiments, as if these subjects considered the p values to be a physical measure of evidence.

Lastly, a major finding in our experiment was that cliff characteristics were only noticeable for a minority of all-or-none respondents. They were of limited magnitude for the other subjects, including the exponential group of subjects who responded most like the subjects in pre-

vious studies. Therefore, one should be cautious about accepting previous claims about the existence of *an abrupt drop* in a *p* level just beyond the fateful .05 level. This is particularly true if one thinks that the representativeness of the samples used in the various experiments was not sufficient to ensure the general validity of the conclusion. The results of the present study should be contrasted with the common publication practice to treat a test outcome as a dichotomous decision. Such a decision probably reflects an essentially circumstantial attitude ("it's the norm"), "mechanical behavior" (Gigerenzer, 1991), or "automatic routine" (Falk & Greenbaum, 1995). Furthermore, it supports one of the long-standing criticisms of null hypothesis significance tests, expressed by Rozeboom (1960), for instance: "But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested" (p. 420).

REFERENCES

Beauchamp, K. L., & May, R. B. (1964). Replication report: Interpretation of levels of significance by psychological researchers. *Psychological Reports*, **14**, 272.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, **48**, 378-399.

Casella, G., & Berger, L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with discussion). *Journal of the American Statistical Association*, **82**, 106-111.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, **5**, 75-98.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "Heuristics and biases." In N. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology 2* (pp. 83-115). New York: Wiley.

Lecoutre, M.-P. (1983). La démarche du chercheur en psychologie dans des situations d'analyse statistique de données expérimentales. *Journal de Psychologie Normale et Pathologique*, **3**, 275-295.

Lecoutre, M.-P. (2000). And . . . What about the researcher's point of view. In H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre, & B. Le Roux, *New ways in statistical methodology: From significance tests to Bayesian inference* (2nd ed., pp. 65-95). Bern: Peter Lang.

Minturn, E. B., Lansky, L. M., & Dember, W. N. (1972). *The interpretation of levels of significance by psychologists: A replication and extension*. Paper presented at the meeting of the Eastern Psychological Association, Boston.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, **41**, 1299-1301.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London: Series A*, **231**, 289-337.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* New York: Wiley.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, **55**, 33-38.

Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports*, **15**, 570.

Rouanet, H. (1998). Significance testing in a Bayesian framework: Assessing direction of effects. *Behavioral & Brain Sciences*, **21**, 217-218.

Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. *American Statistician*, **40**, 313-315.

Rozeboom, W. W. (1960). The fallacy of the hypothetico-deductive significance test. *Psychological Bulletin*, **57**, 416-428.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, **1**, 115-129.

NOTE

1. The cliff effect for this last group would have been estimated more efficiently by the difference between the two estimates of the parameters *a* and *b* of the all-or-none curve, which results in quite a large effect (+.750).

(Manuscript received December 2, 1999;
revision accepted for publication March 27, 2001.)