

From the Editor

ON *p*-VALUES

Periodically, social scientists debate the strengths and weaknesses of hypothesis testing (for which researchers pose the question, e.g., “Are my group means the same or different?”) compared with effects estimation (motivated by the question, “How large is the difference between my group means?”). As is often the case, the extreme positions are clear but they approach ideology, and a moderate stance seems the more constructive prescription.

The testing of null hypotheses affords researchers many advantages (Abelson 1997; Cortina and Dunlap 1997; Frick 1996; Greenwald et al. 1996; Hagen 1997; Harris 1997; Mulaik, Raju, and Harshman 1997). Of primary importance, the test of a null hypothesis is conducted in the context of a simple decision rule and provides a dichotomous outcome (Greenwald et al. 1996, 177). While critics would argue that hypothesis tests provide less information compared to alternative techniques, supporters argue that the binary decisions nevertheless enable scholarly progress and theory testing, which “requires nothing more than a binary decision about the relation between two variables” (Chow 1988, 105; Wainer 1999).

ISSUES WITH NULL HYPOTHESIS TESTING

When hypothesis testing is criticized, it derives not so much from anything inherent to the technique or philosophy itself as from users’ fallacious interpretations of the corresponding results. There are several major concerns.

The first concern is the misinterpretation of a finding when a computed test statistic does not exceed a critical value. We quickly learn to avoid the improper semantics of “accepting” the null hypothesis, in part because there are “many ways (including incompetence of the researcher), other than the null hypothesis being true, for obtaining a null result” (Greenwald 1975, 2). Indeed, researchers frequently (and appropriately) point to insufficient sample size and lack of power as the likely culprit for not obtaining expected results. Given the logic and statistical machinery of hypothesis testing, Greenwald (1975, 2) continues: “a null result is only a basis for uncertainty. Conclusions about relationships among variables should be based only on rejections of null hypotheses.” Recall from your basic statistics training: “The logic of null hypothesis tests is as follows: Assume the null hypothesis is true, and then examine the likelihood of obtaining the sample results [or a result more extreme] based on this assumption; if this likelihood is lower than a specified criterion (the alpha level), reject the null hypothesis. . . . Such tests cannot be used to confirm the null hypothesis” (Dar, Serlin, and Omer 1994, 76). Thus insignificance is “inconclusive” and the philosophy among most social scientists is that “non-significant results should not take up” journal space (Bakan 1966, 427–28). (While we are savvy enough to not typically err and speak of “accepting the null,” nevertheless consider how frequently we test an effect such as “order of stimulus

I am grateful for feedback from Eric Bradlow, Alexander Chernev, John Deighton, Adam Duhachek, Jennifer Escalas, Kent Grayson, Donald Lehmann, Sidney Levy, Mary Frances Luce, Robert Meyer, James Oakley, Joseph Priester, Americus Reed, and Howard Wainer.

presentation” to demonstrate that order did not matter or, for example, that there were no preexisting group differences, presenting an “*n.s.*” in support of that stance.¹)

The second criticism of hypothesis testing is due to the occasional misinterpretation of *p*-values. For example, the *p*-value does not indicate in any simple manner “how likely it is that the results will replicate” (Dixon 1998, 391). Methodological researchers are critical of those who interpret small *p*-values as “highly significant” and borderline ones as “marginally significant” (Dar, Serlin, and Omer 1994, 76). Strictly speaking, it’s true that a finding is merely “significant” or “not significant” (recall the binary characterization), so the adjectives of “highly” or “marginal” are inappropriate. In addition, yes, .05 is “arbitrary,” but it is our accepted standard, and therefore it levels the playing field. (It is of some note that Fisher intended .05 to be relatively “liberal” compared to more stringent values; Wainer and Robinson 2003, 23.) A manuscript with a tiny portion of findings at $p = .06$ is tolerable, but when a paper is peppered with results that do not fall below the relatively liberal cut-off of $\alpha = .05$, researchers should be wary of interpreting significance in the results. Better to bite the bullet, collect more data, and demonstrate that, with more power, the study’s focal premises had indeed been tested definitively.

Yet when researchers deny that “*p* measures the . . . strength of the results” (Dar, Serlin, and Omer 1994, 76), that is not quite accurate either. Of course it is true that *p*-values reflect strengths of results—*ceteris paribus* (primarily, for equal sample sizes), a *p*-value of .005 indicates a larger effect in the sample than $p = .05$. Recall a *p*-value is defined precisely as “the probability of obtaining a test statistic as large or larger than that which was obtained in the data, given the null holds.” Thus, $p = .005$ implies the null is less plausible, the likelihood of obtaining a data result that extreme is even less than for the $p = .05$ result. (Greenwald et al. [1996, 178] offer a whimsical version of this likelihood interpretation—that a *p*-value “is an approximate measure of how surprised we should be by a result.”) Yet if the sample sizes entering into the test statistics vary, as can be the case even in the same sample as a result of proportionally greater missing data for one of the tests, then to say $p = .05$ is a moderate result compared to the more conservative $p = .005$ would be incorrect. While the decision criterion is indeed binary—that is, “Is my *p*-value less than .05?”—it seems like valuable information to convey that $p = .001$ or that, while one failed to reject the null, nevertheless $p = .06$ (Dixon 1998, 391).

Methodologists are increasingly recommending that researchers report precise *p*-values, for example, $p = .04$ rather than $p < .05$ (Greenwald et al. 1996, 181). To use $\alpha = .05$ “is an anachronism. It was settled on when *p*-values were hard to compute and so some specific values needed to be provided in tables. Now calculating exact *p*-values is easy [i.e., the computer does it] and so the investigator can report [$p = .04$] and leave it to the reader to [determine its significance]” (Wainer and Robinson 2003, 26).

Finally, Greenwald et al. (1996, 175–81) also point to the utility of *p*-values as a translating equalizer function. Specifically, the *p*-value allows the comparison across different statistics, for example, z , t , F , r , β , χ^2 , which are otherwise indices that operate on different scales.

No technique answers all research inquiries, and, even if one agrees that hypothesis testing has its limitations, the question for the researcher becomes, “What is the proposed solution?”

¹This issue is challenging. It is tempting to suggest that if the test of order is conducted on roughly the same sample size as the tests of the more focal hypotheses, and the focal hypotheses yield significance, then the lack of significance on the order test is meaningful, i.e., for comparable power, some results were obtained, others not. Unfortunately, one must also factor in the subtle effect of the possibly differentially sensitive measures, i.e., more reliable measures contribute to power by effectively reducing error variances (Meehl 1978, 822), so it is conceivable that the measures of the focal hypotheses were more sensitive and the order measure less reliable, thereby yielding “inconclusive” results. What is one to do? Make more cautious interpretations of *n.s.*, to be sure, at a minimum changing “Order was not significant” to “Order is not sufficient to conclusively explain the results.” Another form of this issue occurs when researchers seek to eliminate rival hypotheses by demonstrating significance using their constructs and *n.s.* on the alternatives. A superior solution would require greater theoretical complexity, e.g., obtaining significant findings within one’s focal constructs in the anticipated direction and significant findings on the alternatives in the opposite direction.

There are four classes of proposed alternatives to hypothesis testing. However, as Cohen (1990, 1001) warns, "Don't look for a magic alternative. . . . It doesn't exist."

ALTERNATIVES TO NULL HYPOTHESIS TESTING

Confidence Intervals

Of all the proposals, confidence intervals are the easiest to compute and the solution whose communication and interpretation lend the greatest facility (Krantz 1999). Beyond "Yes or no, are the population means the same or not?" confidence intervals address a broader question, "What are [good estimates for] the population means?" (Loftus 1991, 103). We know that in examining a confidence interval, we are implicitly assessing an infinite number of hypotheses; a confidence interval "treats all the alternative hypotheses with glacial impartiality" (Rozeboom 1960, 427). We also know that confidence intervals provide precision information as well, reflecting sample size in an informative manner: "The more power you have, the smaller are your confidence intervals (i.e., the better your knowledge of where population means are)" (Loftus 1991, 103). When scholars are reluctant to report confidence intervals, it may be because they provide too much information, for example, a confidence interval makes it painfully obvious when a parameter estimate is disappointingly small or the confidence interval itself embarrassingly wide (Reichardt and Gollob 1997).²

Bayesian Estimation

The second class of proposed alternative to hypothesis testing is Bayesian estimation. The logic of the classic null hypothesis test is to assess $P(d|H)$; that is, "What is the probability of obtaining my data (specifically a test statistic this big or bigger) given the (null) hypothesis is true?" Yet researchers seem to find more appealing, and interpret their results more akin to $P(H|d)$ (Cohen 1994, 997); that is, "Given my data look as they do, what is the likelihood that my hypothesis holds?" The latter is a conditional probability that is part of Bayesian estimation. Thus, the argument goes, if the researcher's intuition more naturally lends itself to Bayesian logic, why not fully embrace it (Bakan 1966, 436; Dixon 1998, 392; Nickerson 2000, 241; Rozeboom 1960, 420)?

Bayes's theorem, $[P(H_0|d)]/[P(H_1|d)] = \{[P(H_0)]/[P(H_1)]\} \times \{[P(d|H_0)]/[P(d|H_1)]\}$, isn't terribly complicated when the components are explained. The term $P(H_0)/P(H_1)$ is the prior odds ratio, a means of representing the relative a priori belief in two hypotheses or models before data are collected. The term $P(d|H_0)/P(d|H_1)$ captures the relative odds of the data assuming either of the hypotheses holds (the "classic" statistics orientation), and $P(H_0|d)/P(H_1|d)$ is the posterior odds, an update in the relative confidence in the two hypotheses given what one has found in the data.

Note, however, that the Bayesian approach brings two new problems that preclude it from being an ideal solution. First, researchers do not always have good estimates for all these new components. Compelling uses of Bayes often refer to applications with population incidences and relatively well-known base rates (e.g., the proportion of schizophrenics in some population), as well as the diagnosticity of the instrument yielding the data (e.g., the accuracy of a test for detecting and properly identifying the schizophrenic; Hagen 1997, 15). If one's own research paradigm does not easily yield counterparts to the Bayesian elements, then the added components of Bayes merely introduces more subjectivity into the interpretation of results rather than clarification. Second, if researchers are making interpretive mis-

²Recall that a null hypothesis would be rejected for values not contained in the confidence interval. Instead of reporting, e.g., "the groups were significantly different ($F_{1,29} = 13.693, p = .0009; \bar{M}_1 = 4.75, \bar{M}_2 = 4.00$), a researcher would report, "the 95th percent confidence interval around $\bar{M}_1 - \bar{M}_2 = [0.642, 0.858]$ indicates the groups are significantly different." Even given the clear advantages of confidence intervals (over the other three alternatives to hypothesis testing described shortly), note that this computation and reporting would be somewhat cumbersome as the researcher progresses to interaction terms, thus undermining its potential status as a perfect alternative solution.

takes with relatively simple tools that have been taught and used for 100 years, it is possible that the implementation of Bayes might induce its own problems (Harlow, Mulaik, and Steiger 1997, 13).

Effect Size

A third alternative (or supplement) to hypothesis testing is the estimation of the size of the effects in the data (Bakan 1966, 436). This suggestion is usually motivated by one of two concerns. First, in reporting the complement of hypotheses tested and effect sizes, the researcher can make clear the distinction between statistical “significance” and scientific “importance.” (Here too, we might not frequently commit this sin of confusing significance with importance, yet we may come close to doing so when we occasionally overstate “implications” in discussion sections.) The distinction between significance and importance also stems from researchers who query, “Why are we testing a null, which we (almost certainly already) know to be false?” (Loftus 1991; Meehl 1978, 817ff). That is, most groups’ means are not identical, most constructs are at least somewhat correlated, and so on; thus the likelihood of finding significant mean differences or significant correlation patterns is primarily a function of sample size (whereas effects indices are essentially independent of sample size). If we know the group means are significantly different but the difference is small, or a correlation is statistically nonzero but also small, that supplemental magnitude estimation is informative.

A second motivation for computing effect sizes is to have indices to enable the accumulation of knowledge in science (Hedges 1987, 443) as data in meta-analyses (Hunter 1997; Schmidt 1992, 1996). To do so, there exist a plethora of effect size indices, such as: r^2 , η^2 , ω^2 , R^2 , ϕ^2 , Cohen’s d , Hedge’s g , and so on, therein lying one problem—an ambiguity as to which index should be used in which circumstances. Further, these indices are descriptive, not inferential, thus the interpretation of the size of the estimates produced is entirely subjective, with “rules of thumb” that are simplistic (i.e., contingent upon no experimental design parameters) being offered as to what comprises a “large,” “medium,” or “small” effect size (Cohen et al. 2003, 179; Rosnow and Rosenthal 1996, 332). Finally, critics would argue that large effects could be obtained in experiments by simply using “heavy-handed” manipulations. Thus, while effect size indices are appealing in spirit, they are problematic in execution.

Miscellany

A fourth class of alternatives is a potpourri collection of indices that researchers have proposed to address various concerns, and each is accompanied by its own issues. Posavac (2002) offers a means of taking a p -value and computing an estimate of the likelihood of replicating the findings. Rosenthal and Rubin (1994) and Rosnow and Rosenthal (1996) suggest computing the “counternull,” an alternative hypothesis that, when used as the null, generates the same p -value as had the null hypothesis of “zero difference.” Dixon (1998, 393) points to p -values as interpretable as the linear reflections of the aforementioned (classic) likelihood ratio and suggests researchers obtain support for their hypothesis, compared with the alternative, on the order of 10:1 (a well-argued, though ultimately also arbitrary and subjective rule of thumb).

SUMMARY

In close, no technique is perfect. Hypothesis tests may carry the disadvantage of being simple, but they simultaneously carry the advantage of . . . being simple.

The American Psychological Association’s Board of Scientific Affairs commissioned a white paper, “Task Force on Statistical Inference” (available at apa.org), which influenced policies

in the fifth edition of its publication manual. We endorse these stylistic suggestions for reporting quantitative (e.g., experimental or survey) data:

When reporting inferential statistics (e.g., *t* tests, *F* tests, and chi-square), include information about the obtained . . . value of the test statistic, the degrees of freedom, the probability of obtaining a value as extreme as or more extreme than the one obtained [i.e., the *p*-value]. . . . Be sure to include sufficient descriptive statistics (e.g., per-cell sample size, means, correlations, standard deviations). . . . The reporting of confidence intervals (for estimates of parameters, for functions of parameters such as differences in means, and for effect sizes) can be an extremely effective way of reporting results . . . because confidence intervals combine information on location and precision and can often be directly used to infer significance levels (2001, 22).

PRESCRIPTION? KEEP DOING WHAT YOU'RE DOING

In truth, current practice in reporting quantitative research in the *Journal of Consumer Research* largely resembles these guidelines. If the presentation of confidence intervals proper is actually rare, nevertheless hypothesis tests are nearly always supplemented with the full array of means, cell sizes and standard deviations, and so on, thus enabling the knowledge accumulators to cobble together estimates of whichever effect size indices they prefer, hence satisfying both the hypothesis testing and effects estimation perspectives on scientific inquiry.

Dawn Iacobucci
Editor
March 2005

REFERENCES

- Abelson, Robert P. (1997), "On the Surprising Longevity of Flogged Horses: Why There Is a Case for the Significance Test," *Psychological Science*, 8 (1), 12–15.
- American Psychological Association (2001), *Publication Manual*, 5th ed., Washington, DC: American Psychological Association.
- Bakan David (1966), "The Test of Significance in Psychological Research," *Psychological Bulletin*, 66, 423–37.
- Chow, Siu L. (1988), "Significance Tests or Effect Size?" *Psychological Bulletin*, 103, 105–10.
- Cohen, Jacob (1990), "Things I Have Learned (So Far)," *American Psychologist*, 45, 997–1003.
- (1994), "The Earth is Round ($p < .05$)," *American Psychologist*, 45, 1304–12.
- Cohen, Jacob, Patricia Cohen, Stephen G. West, and Leona S. Aiken (2003), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed., Mahwah, NJ: Lawrence Erlbaum.
- Cortina, Jose M. and William P. Dunlap (1997), "On the Logic and Purpose of Significance Testing," *Psychological Methods*, 2 (2), 161–72.
- Dar, Rueven, Ronald C. Serlin, and Haim Omer (1994), "Misuse of Statistical Tests in Three Decades of Psychotherapy Research," *Journal of Consulting and Clinical Psychology*, 62, 75–82.
- Dixon, Peter (1998), "Why Scientists Value *p* Values," *Psychonomic Bulletin & Review*, 5, 390–96.
- Frick, Robert W. (1996) "The Appropriate Use of Null Hypothesis Testing," *Psychological Methods*, 1 (4), 379–90.
- Greenwald, Anthony (1975), "Consequences of Prejudice against the Null Hypothesis," *Psychological Bulletin*, 82, 1–20.
- Greenwald, Anthony G., Richard Gonzalez, Richard J. Harris, and Donald Guthrie (1996), "Effect Sizes and *p* Values: What Should Be Reported and What Should Be Replicated?" *Psychophysiology*, 33, 175–83.
- Hagen, Richard L. (1997), "In Praise of the Null Hypothesis Statistical Test," *American Psychologist*, 52, 15–24.
- Harlow, Lisa L., Stanley A. Mulaik, and James H. Steiger, eds. (1997), *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum.
- Harris, Richard J. (1997), "Significance Tests Have Their Place," *Psychological Science*, 8 (1), 8–11.
- Hedges, Lawrence V. (1987), "How Hard is Hard Science, How Soft Is Soft Science? The Empirical Cumulativeness of Research," *American Psychologist*, 42, 443–55.
- Hunter, John E. (1997), "Needed: A Ban on the Significance Test," *Psychological Science*, 8 (1), 3–7.
- Krantz, David. H. (1999), "The Null Hypothesis Testing Controversy in Psychology," *Journal of the American Statistical Association*, 44, 1372–81.

- Loftus, Geoffrey R. (1991), "On the Tyranny of Hypothesis Testing in the Social Sciences," *Contemporary Psychology*, 36, 102–5.
- Meehl, Paul E. (1978), "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology," *Journal of Consulting and Clinical Psychology*, 46, 806–34.
- Mulaik, Stanley A., Nambury S. Raju, and Richard A. Harshman (1997), "There Is a Time and a Place for Significance Testing," in *What If There Were No Significance Tests?* ed. Lisa L. Harlow, Stanley A. Mulaik, and James H. Steiger, Mahwah, NJ: Lawrence Erlbaum, 65–115.
- Nickerson, Raymond S. (2000), "Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy," *Psychological Methods*, 5, 241–301.
- Posavac, Emil J. (2002), "Using p Values to Estimate the Probability of a Statistically Significant Replication," *Understanding Statistics*, 1 (2), 101–12.
- Reichardt, Charles S. and Harry F. Gollob (1997) "When Confidence Intervals Should Be Used Instead of Statistical Significance Tests, and Vice Versa," in *What If There Were No Significance Tests?* ed. Lisa L. Harlow, Stanley A. Mulaik, and James H. Steiger, Mahwah, NJ: Lawrence Erlbaum, 259–84.
- Rosenthal, Robert and Donald B. Rubin (1994), "The Counternull Value of an Effect Size: A New Statistic," *Psychological Science*, 5, 329–34.
- Rosnow, Ralph L. and Robert Rosenthal (1996), "Computing Contrasts, Effect Sizes, and Counternulls on Other People's Published Data: General Procedures for Research Consumers," *Psychological Methods*, 1 (4), 331–40.
- Rozeboom, William W. (1960), "The Fallacy of the Null Hypothesis Significance Test," *Psychological Bulletin*, 57, 416–28.
- Schmidt, Frank L. (1992), "What Do Data Really Mean? Research Findings, Meta-Analysis, and Cumulative Knowledge in Psychology," *American Psychologist*, 47, 1173–81.
- (1996), "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers," *Psychological Methods*, 1 (2), 115–29.
- Wainer, Howard (1999), "One Cheer for Null Hypothesis Significance Testing," *Psychological Methods*, 4 (2), 212–13.
- Wainer, Howard and Daniel H. Robinson (2003), "Shaping Up the Practice of Null Hypothesis Significance Testing," *Educational Researcher*, 32, 22–30.