THE DOG THAT DIDN'T BARK… INTERPRETING NON SIGNIFICANCE

Diana Eugenie Kornbrot, Rachel M. Msetfi
University of Hertfordshire, UK
d.e.kornbrot@herts.ac.uk, r.msetfi@herts.ac.uk

Abstract

Hypothesis testing is a crucial component of science. This leads to guidelines (often ignored) in most discipline including psychology. Unfortunately, most focus on significant effects. Non-significant effects are sidelined, in spite of their importance to scientific progress. This study reports a survey of practicing scientist on how they would report and interpret explicit scenarios with non-significant effects. There was no consensus on interpretation in terms of predicting future results. Respondents agreed about how to report the significance of a hypothesis test. Most chose not to report any descriptive statistics, or the sample size, or anything about power, or sufficient information to enable replication or meta-. These results shed light on statistical thinking and so should enable more useable guidelines. For non-significant effects, the importance of a priori power is emphasised.

"It is impossible to prove a negative, i.e. non-existence of some phenomenon". Furthermore, "falsifiability", the ability to prove a positive, is claimed to be the very hallmark of science. This perception of the nature of science may be one reason that the search for statistically significant results appears to dominate empirical research guidelines. Examples of such guidelines include: (APA, 2001; Nickerson, 2000; Roberts & Pashler, 2000; Steingrimsson & Luce, 2005; Wilkinson, 1999)for psychology; (Duran et al., 2006) for education and (Campbell, Elbourne, & Altman, 2004; CONSORT, 2001; QUORUM, 2000; STROBE, 2005)for medicine. Most guidelines focus on significant results only, and recommend reporting effect sizes. Only Nickerson(, 2000) explicitly recommends giving a priori power for non-significant results.

It is our contention that identifying an absence of difference is one of the most important tools of science, both theoretically and practically. In the physical sciences, conservation laws (e.g. energy, speed of light ion a vacuum) are cornerstones of theory. For psychologist, it is important to establish an absence of difference in the following situations: 1. when theory predicts equality constraints (Birnbaum, 2004; Steingrimsson & Luce, 2005); 2. when establishing lack of evidence for harm of beneficial procedure, e.g. the MMR vaccine or side effects of drugs; 3. when establishing that an effect is highly improbably, e.g. psycho kinesis; and 4. when establishing that controls such as counterbalancing have been successful. Power is crucial for all these applications.

The current study reports an internet survey on the views of practicing scientists on how to report and interpret non-significant results. The aim is a better understanding researchers' perceptions in order to underpin guidelines on non-significance. There is continuing evidence that many guidelines are ignored (Cumming et al., 2007).

Method

Respondents

There were 230 respondents who replied to an appeal to 1 of 17 email lists (5 psychological methods, 4 statistical packages, 2 statistics 2 hci/surveys; 3 education and 1 history.

Materials

There were 8 versions of the survey (http://web.mac.com/kornbrot/iWeb/statspublic.html). Each version comprised two scenarios comparing the proportion of people in 2 groups with high blood pressure. The (non-significant) chi-square value and p(null) were also provided. Two factors were counterbalanced in the scenarios. GROUPs could be either gender (men, women) or video (travel or health). VALUEs could be either high (group1=28%, group 2 = 38%) or low (group 1=16%, group 2 = 24%). People were also asked for confidence judgements for each scenario: in either MAGNITUDE of confidence or AGREE with "I am confident" statement. Values for the 1st scenario are shown in Table 1. The 2nd scenario had opposite values on each factor.

Table 1. Values for each factor for 1st Scenario in Survey

	V1	V2	V3	V4	V5	V6	V7	V8
Group	gender	gender	gender	gender	video	video	video	video
Value	low	low	high	high	low	low	high	high
Confidence	agree	magnitude	agree	magnitude	agree	magnitude	agree	magnitude

Variables

Respondents were first asked how they would REPORT the results to lay and professional audiences. The reports were then coded to produce the following response variables: descriptive statistics, inferential statistics, sample size, sufficiency, confidence limits & effect size, odds ratio.

Respondents were also asked a multiple-choice question on how they would INTERPRET the results in terms of what they would predict for a replication of the same study. The choices available were: 1. numbers equal to the original separate group values; 2. a number equal to the combined original value; 3. 50%; 4. no prediction possible; and 5. other.

Observational explanatory factors were obtained from participant responses as to their: DISCIPLINE (coded to 1. biological science, 2. social science, 3. psychology, neuroscience, cognitive science or education, 4. statistics, maths, physics, computer science, 5. not given); ROLE (coded to 1. teaching or teaching + research, 2. research only, 3. not given); theoretical STANCE (coded to 1. includes Bayesian approaches, 2. frequentist, 3. unfamiliar with Bayes approach); and most common PROCEDURE (1. t-test, 2. ANOVA, 3. regression, 4.other). GROUP and VALUE, derived from the survey version described under materials, were experimentally manipulated factors.

Results and Discussion

The median time was 16.3 minutes,, with 15 taking more than an hour (breaks are possible). Some respondents did not progress past the 1st scenario and hence did not provide information of role and discipline (43 gave no bio data). Table 2 shows numbers by role and discipline.

Table 2. Numbers of respondents by role and discipline

	social science	biological science	psychology, education, cog sci, neuropsy	statistics, maths, phys sci
teaching +	37	10	23	34
research only	35	10	10	13

Likelihood ratio chi-squared tests on contingency tables have been used throughout, with a 95% confidence level. The measure of effect size is the contingency coefficient, w = √(χ^2/N), with w^2 giving a measure of variance accounted for. Power to detect medium effects (w = .3) was reasonable, from .79 to .98; but power to detect small effects (w = .1) was small, from .15 to .25. The effect sizes that could be detected with power of 0.8 varied from w = .21 to w = .40, i.e. 4% to 9% variance accounted for. There was no effect of GROUP in the scenario or the VALUE of the proportion with blood pressure, or of the order in which people saw the scenarios. However, as noted power was 'modest' in this exploration.

REPORT Variables: Coded Free Form Reports to Professional & Lay Audiences

There were 15 respondents who considered the design so awful that they did not give an analyzable response to any 'how would you report study'. No purpose was given for the scenarios and most appeared to *assume* that the purpose was to generalize to the adult population at large. With hindsight, it would have been better to provide a realistic purpose. For example, a concerned local group is considering setting up a healthy living clinic in a shopping mall to reduce the prevalence of high blood pressure in the community by providing advice, exercise facilities and health promoting videos. Their first concern is the potential demand. They also want to pilot a video and estimate separate demands for men and women, so as to provide sufficient facilities (e.g. changing rooms for women and men). Dichotomising the continuous blood pressure variable was also criticized.

- *Descriptive statistics*

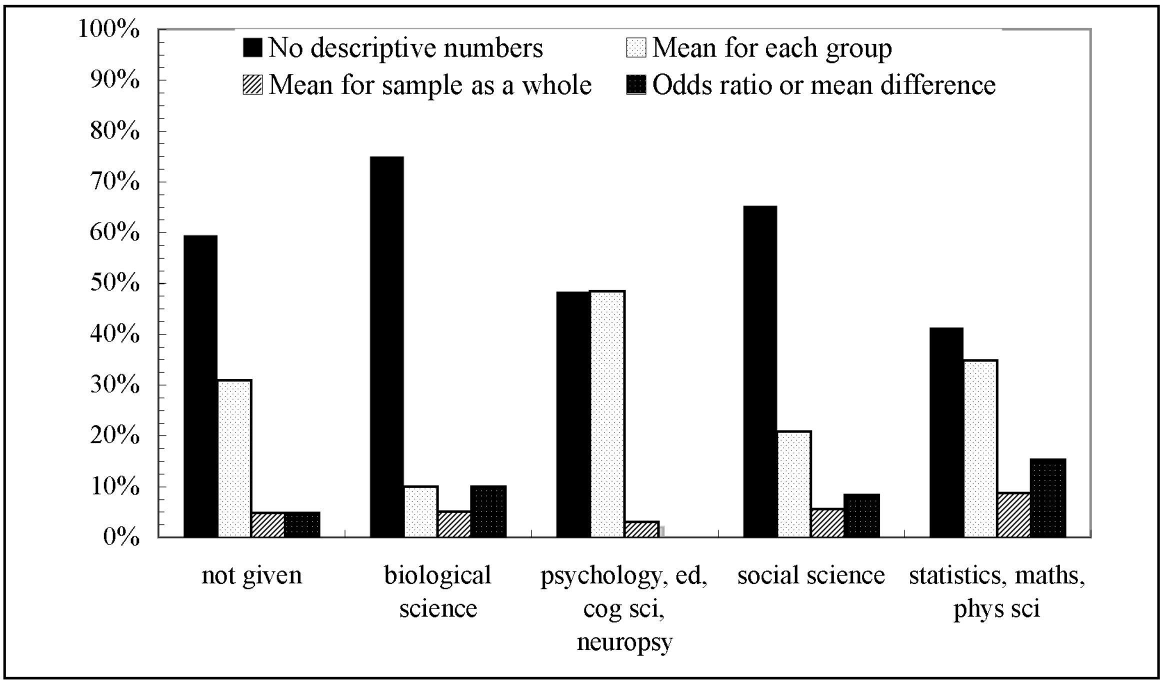

Figure 1. Percentage choices of descriptive statistics reporting as a function of discipline.

Figure 1 shows the proportion of people choosing each way of reporting the descriptive statistics. The most salient feature is that the majority of the 213 respondents chose not to give any descriptive statistics at all (57%). The next most popular choice was to give the mean proportion for each group separately (29%). There was a significant effect of discipline, LR $\chi^2(12)$ = 23.5, p = .024, effect size w = .33. Post hoc analyses show that statisticians (59%) and psychological scientists (51%) were more likely to report some form of descriptive

For effect size/ confidence interval, $\chi^2(6) = 14.3$, p = .024. For odds ratio, , $\chi^2(2) = 6.4$, p = .041. Overall, 15% of people made some mention of effect size or confidence intervals. Just 7 people suggested odds ratios (3.3%), interesting but not directly relevant to non-significance reporting. It appears that having a role in teaching (either with or without research) was associated with a greater probability of suggesting both effects sizes/confidence intervals and odds ratios.

INTERPRETATION: prediction of outcome of replication

The predictions for the 1st and 2nd scenarios were analyzed separately. Anomalous predictions (50% or other were dropped from the analyses). This left 199 predictions for scenario 1 and 165 for scenario 2. The scenario 1 choices were: 38% separate groups, 30% combined groups, 32% not predictable. The scenario 2 choices were: 34% separate groups, 31% combined groups, 35% not predictable. Thus, there is *no* consensus as to the interpretation of the implications of non-significant results for future studies.

The only statistically significantly effect was the most common procedure used by the respondents, see Fig. 3. For scenario 1, N =170 , $\chi^2(8) = 14.3$, p = .067, w = .27. For scenario 2, N =170 , $\chi^2(8) = 20.0$, p = .010, w = .30.

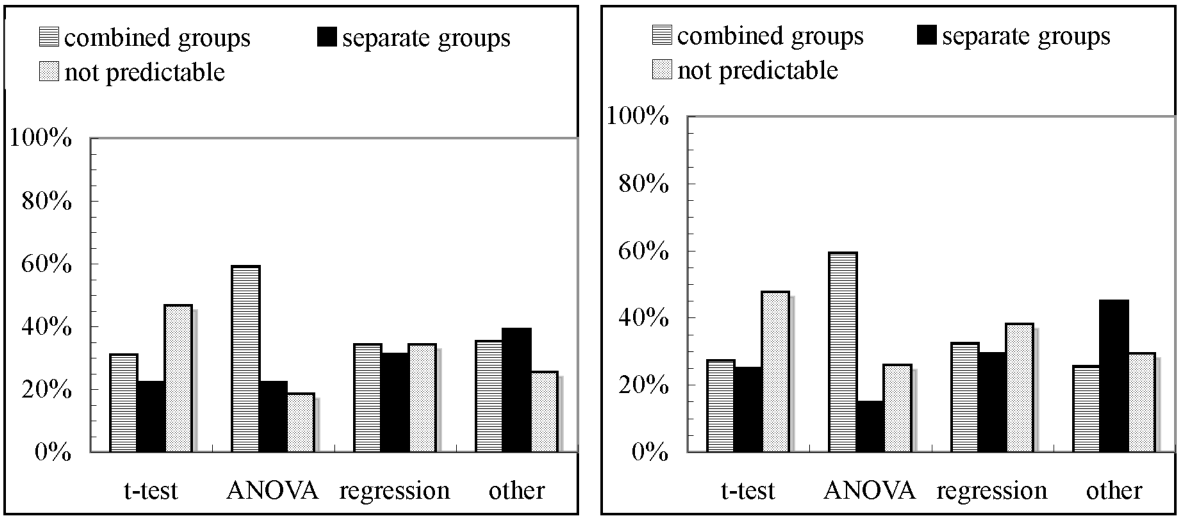

Figure 3. The effect of most frequent analysis on predictions of replications. Scenario 1, left panel; scenario 2, right panel.

The respondents who most frequently used a t-test chose ‘not predictable’ most often (48%) Those who most frequently used ANOVA chose ‘combined groups’ most often (59%). The rest were about equally distributed among the three alternatives.

Summary

Informative reporting of non-significant results has a key role to play in theory and practice in all areas of science, including psychophysics. Guidelines need to *explicitly* address the reporting of non-significant results. Our recommendation, following Nickerson 2000, would be to *always* give a priori power for non-significant effects. This is the flip side of *always* giving effect size for significant effects. Of course, if there the reporting is sufficient by inclusion of sample size, descriptive statistics (proportions or mean and s.d.) and inferential test-statistics, then the effect sizes and a priori power can be calculated. However, we believe statistics than biologists (25%) or social scientists (35%). There were no statistically significant effects of any other factors on the reporting of descriptive statistics.

- *Inferential statistics*

There was far greater consistency in reporting inferential statistics. Only 2.8% gave no inferential statistics at all. The majority (85%) reported to professional audiences that the difference between groups was not statistically, significant so there was insufficient evidence to support a group difference. However 10.8% who did not want to confuse the public and reported to lay audiences baldly that ‘there is no effect”, i.e. non-significant = no effect. There were 2.8% who reported “no effect” even to professional audiences. There were 7% who thought the scenario designs were so awful that they would not report at all. There were no significant differences according to any explanatory factor. The hypothesis testing culture is strongly embedded, for better or worse.

- *Sample size, sufficiency of information, power*

There were no significant effects of any factor on these variables. With 213 participants, power to detect a medium effect, w = .3, was .90 for most common procedure, .95 for theoretical stance, .96 for discipline and .98 for role. Effect sizes that could be detected with power of 0.8 ranged from w= .24 (4.5% variance) to w = .27 (7.1% variance).

Nearly three quarters (73%) chose NOT to provide sample size in their report to professional audiences. This might be because sample size is usually given in the method section of a report, so respondents felt it unnecessary to repeat the information in the results.

Only 14% gave sufficient information to support meta analysis, or equivalently, for future researchers to check whether a replication had produced significantly different results. In this case, sample size and χ^2 would have been sufficient, since it is obvious that df = 1 and p(null) can easily be determined from a spreadsheet.

About a quarter (24%) warned that although tests gave no significant effect, power was low so an effect was not precluded.

- *Effect size and confidence levels, odds ratios*

Other reporting suggestions included effect sizes and/or confidence limits, and presenting odds ratios. The proportion making these suggestions, depended on role see Fig. 2.

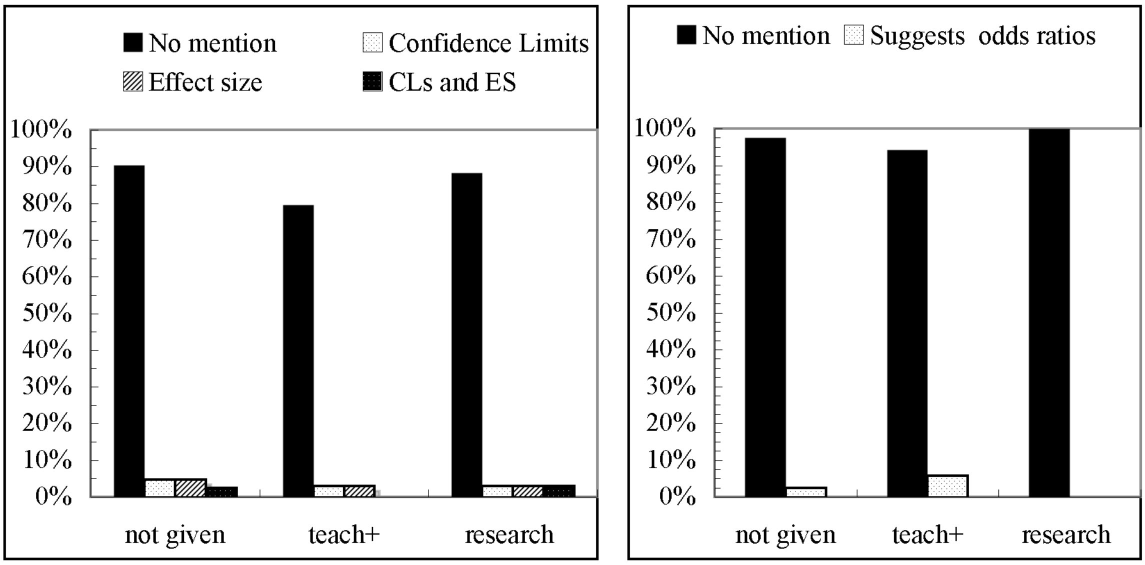

Figure 2. Reports of effect sizes/confidence interval (left panel) and odds ratios (right panel)

it is a researcher's responsibility to make these key parts of any empirical study immediately salient and available to the reader.

Respondents to this survey are obviously committed to sound statistical practice. Nevertheless, the focus was clearly more on inference than description. Almost everyone reported the result of the hypothesis test. However, only a minority reported descriptive statistics of any kind. Furthermore, only a minority reported sufficient information for replication. Clearly, in a real scientific Ms. these issues would be addressed. What this study shows is that descriptive statistics and sufficient information are not *salient* to many scientists. This is perhaps why guidelines fall on such stony ground. On a more optimistic note, perhaps knowing what is salient will help improve future guidelines, and practice.

References

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al. (2001). The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration. *Annals of Internal Medicine, 134*, 663-694.

APA. (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, D.C.: American Psychological Association.

Birnbaum, M. H. (2004). Tests of rank-dependent utility and cumulative prospect theory in gambles represented by natural frequencies: Effects of format, event framing, and branch splitting. *Organizational Behavior and Human Decision Processes, 95*(1), 40-65.

Campbell, M. K., Elbourne, D. R., & Altman, D. G. (2004). CONSORT statement: extension to cluster randomised trials. *BMJ, 328*(7441), 702-708.

CONSORT. (2001). CONSORT Statement on randomized controled trials (RCT. Retrieved 1 Aug, 2007, from http://www.consort-statement.org/Downloads/download.htm

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., et al. (2007). Statistical Reform in Psychology: Is Anything Changing? *Psychological Science, 18*(3), 230-232.

Duran, R. P., Eisenhart, M. A., Erickson, F. D., Grant, C. A., Green, J. L., Hedges, L. V., et al. (2006). Standards for Reporting on Empirical Social Science Research in AERA Publication. *American Educational Research Association*, from http://www.aera.net/uploadedFiles/Opportunities/StandardsforReportingEmpiricalSocialScience_PDF.pdf

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

QUORUM. (2000, May, 2000). THE QUOROM STATEMENT on Systematoc Reviews in Medicine. Retrieved 1 Aug, 2007, from http://www.consort-statement.org/QUOROM.pdf

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*(2), 358-367.

Steingrimsson, R., & Luce, R. D. (2005). Evaluating a model of global psychophysical judgments. I: Behavioral properties of summations and productions. *Journal of Mathematical Psychology, 49*(4), 290-307.

STROBE. (2005, 28 Jun 2007). STROBE Statement: STrengthening the Reporting of OBservational studies in Epidemiology. Retrieved 1 Aug, 2007, fromhttp://www.strobe-statement.org/PDF/STROBE-Checklist-Version2.pdf

Wilkinson, L. (1999). Statistical methods in psychology journals - Guidelines and explanations. *American Psychologist, 54*(8), 594-604.