

Some Suggestions for Improving Research Methods in the Social and Behavioral Sciences:
Putting the MAGIC Back into Your Research

Jane Buck

Delaware State University

In Statistics as Principled Argument, Abelson (1995) stated that five elements, which can be summarized by the acronym, MAGIC, control the persuasiveness of research results. They are magnitude, or the degree of empirical support for the claim made by the data; (2) articulation, or the level of detail provided by the results; (3) generality, or the degree to which the results apply to a broadly defined population; (4) interestingness, which involves both the importance of the issue being investigated and the data's ability to change existing beliefs; and (5) credibility, which arises out of sound theory and method. Although these five elements are intertwined and almost equally consequential, my focus here will be on magnitude, the element that typically creates the most misunderstanding in the interpretation and implementation of research results, especially in studies involving gender, race, or ethnicity.

Is it true that women are more sensitive than men to nonverbal cues? Are women more phobic than men about mathematics? Do the race, gender, age, and ethnicity of the researcher affect participants' responses? Questions such as these have received a great deal of attention over the past few decades, often with conflicting results and a paucity of information concerning the degree to which the results confirm or negate the research hypothesis of interest. A major reason for this state of affairs is that most research in the natural, social, and behavioral sciences relies on rejection of the null hypothesis as the basis for claiming statistical significance.

Rejection of the null hypothesis is a necessary but not sufficient condition for establishing the magnitude of results. The only claim that can be made for statistically significant results based solely on rejection of the null hypothesis is that the results are unlikely to have been produced by sampling error or chance factors. For example, if the null hypothesis states that men and women do not differ in their sensitivity to nonverbal cues, stating that the results reveal a difference that is sufficiently large to reject the null hypothesis at the .05 level of significance tells us merely that there is only a 5% probability that the difference is actually zero. Another way of stating this is that there is only a 5% probability that we have rejected a true null hypothesis--a Type I error. We know nothing about the magnitude of the difference.

This lack of specificity has led some statisticians to criticize, or even to call for the abolition of, the test of significance (Bakan, 1966; Hunter, 1997; Shea, 1996). Others, while recognizing the limitations of the technique, have defended its usefulness (Abelson, 1997; Harris, 1997). Still others have suggested that the problem lies not in the procedure but in the widespread misunderstanding of the term; many non-statisticians believe that statistical significance implies importance or practical usefulness (Scarr, 1997). Even defenders of the significance test tend to agree that the addition of measures of magnitude would substantially improve the usefulness of research results.

Time does not permit a discussion of the many issues involved in sampling, research design, and the reliability and validity of measurement, all of which have an impact on the magnitude of research results. I shall limit this presentation to a brief summary of some of the more promising suggestions for addressing the issue of magnitude. They include confidence limits, meta-analysis, power analysis, the reporting of null results, and comparing variances.

Confidence Limits

Virtually every elementary statistics textbook includes the calculation of confidence limits along with the calculation of tests of significance, and most statistical software packages will provide both measures with a few key strokes. Some find that confidence intervals not only provide more information than do tests of significance, but are easier to understand (Hunter, 1997). As early as the mid-1980s, the American Journal of Public Health all but banned the use of significance tests in favor of reporting confidence limits (Shrout, 1997).

Why the campaign to replace or at least augment tests of significance with confidence limits? Remember that tests of significance yield very limited information. The statement that the difference between two means is statistically significant at the .05 level means only that there is a 5% probability that the difference is non-chance or non-zero. The corresponding 95% confidence limits, on the other hand, provide a range of values between which one can be 95% confident of having included the true difference. If results are reported as significant at the .01 level, the corresponding confidence limits are 99% limits.

Let's look at a hypothetical example to illustrate the point. In an extensive study involving 1000 men and 1000 women, a psychologist finds that the mean score for women on a test of verbal fluency is 102, and the mean score for men is 100, with a standard deviation of 16 for each group. The t test for the difference between two means reveals that the two-point difference is significant at the .01 level. It is all too easy to exaggerate the importance of such a result because of the very low probability that the difference is zero or caused by chance factors. Although reporting the confidence limits does not eliminate the problem, it substantially reduces it. The 99% confidence limits for these data are .157 and 3.843, less than one-fifth of a point on the low end and almost four points on the high end. In other words, we can be 99% confident that we have included the true mean difference in the interval between .157 and 3.843.

Although we can be reasonably certain that the mean difference is not zero, it would be a serious error to make too much of a two-point difference in this case for at least two reasons. In the first place, the true difference might be very close to zero, as indicated by the lower limit of .157; in the second place, even if the true difference were 3.843, the upper limit, that would be a difference of just under one-fourth standard deviation. The best bet, statistically, is that the true mean difference is two points, or one-eighth standard deviation. These results are hardly exciting from the standpoint of practical significance. This example illustrates that small differences are easily detected and yield statistically significant results when using large samples. Reporting confidence limits instead of or in addition to the test of significance gives a clearer picture of the magnitude of the results.

Meta-analysis

Meta-analysis is a much-touted and much-criticized relatively recent technique that is very useful in estimating the magnitude of results, especially when a number of studies give conflicting results. The method involves combining the results of all the well-designed studies of a given phenomenon. Estimates of the average magnitude of the results, or effect size, and of the degree to which the results vary across studies are obtained.

By analyzing differences in research protocols among studies, one might discover possible explanations for conflicting results (Abelson, 1995). For example, if all or most of the studies yielding positive results employed participants over the age of 40, and all or most of the studies yielding negative results employed participants under the age of 40, it is reasonable to speculate that age is a variable that should be controlled in further research on the phenomenon.

By combining the results of a number of studies, meta-analysis reduces the risk of relying

too heavily on individual studies, some of which almost certainly contain Type II errors as the result of low power. Additionally, the technique has the potential to contribute to theory by revealing regularities that are often obscured by more traditional methods (Schmidt, 1992)

Power Analysis

Power is the probability of rejecting a false null hypothesis, in other words, of obtaining statistically significant results. All other things being equal, power is directly related to sample size. Obtaining an estimate of power prior to gathering data on a sample of a given size can help the investigator avoid the pitfall of failing to reject a false null hypothesis. If estimated power is only .20 with a sample of 25, it would be foolhardy to proceed with the study without obtaining a larger sample. Too many investigators who fail to estimate power find themselves in the position of having to retain the null hypothesis when it is false, thus committing a Type II error. Hunter (1997) claims that, because of the small samples typically used in social and behavioral research, the average Type II error rate might be as high as 60%.

Even if one has failed to reject the null hypothesis, a post hoc estimate of power can provide extremely useful information about the probability that the decision was correct. A failure to reject the null hypothesis is much more likely to be the correct decision when power is high than when it is low. If power was .90, the probability of committing a Type II error was only .10 compared with .80 if power was only .20.

There are situations, however, in which high power is not necessarily desirable. If the magnitude of the results or effect size is so small as to be practically or theoretically trivial, it is better not to reject the null hypothesis or at least to interpret such a result with extreme caution. Scarr (1997) gives a telling hypothetical example. On the basis of a minuscule, but statistically significant, correlation between eating oatmeal and developing brain cancer, we might foolishly avoid eating oatmeal.

Reporting Null Results

There is a bias in the social and behavioral sciences against publishing null results (Hubbard & Armstrong, 1997; Rosenthal, R., 1979). Most journals are reluctant to publish studies that fail to reach the arbitrary and conventional .05 level of significance. Without knowledge of the number of studies that have been conducted and filed away because of their failure to reach significance, it is impossible to estimate the prevalence of Type II errors.

The publication of null results would have a number of salutary effects on the research enterprise. It is informative and useful to know that a given result is not significant even in the presence of high power. This is suggestive, albeit not conclusive, of a very small effect size. Researchers would be well advised not to pursue further research in such a case. (Hubbard & Armstrong, 1997). On the other hand, failure to reject the null hypothesis when power is low suggests a high probability of a Type II error, and one should be encouraged to pursue further studies using improved research methods.

Comparing Variances

Although most elementary statistics textbooks advise the comparison of variances when employing statistical techniques for comparing means, they tend not to address the issue of heterogeneity of variance as the result of experimental treatment (Buck, 1990). If two or more equivalent groups receive different experimental treatments, it is quite possible that the treatments will have effects such that there is no significant difference between means, but a statistically significant difference between variances.

A study of attitudes towards penalties for various criminal offenses found no difference

between the means of two groups. A comparison of variances, however, revealed that the experimental group, which had been convicted of driving under the influence of alcohol, proposed a more variable list of penalties for speeding than did the control group. The control group, on the other hand, proposed a much more variable list of penalties for shoplifting and embezzlement (Landauer, Harris, & Pocock, 1982). Had the investigators not analyzed the data for differences in variances, they would have come to the erroneous conclusion, on the basis of comparing only the means, that there was no difference between the two groups.

Without suggesting that we should abandon tests of significance, I would urge that we add other approaches and techniques to our research armamentarium. Some of the more promising are confidence limits, meta-analysis, power analysis, the reporting of null results, and comparing variances.

References

- Abelson, R. P. (1995). Statistics as Principled Argument. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. Psychological Science, *8*, 12-15.
- Bakan, D. (1966). The test of significance in psychological research. Psychological Bulletin, *66*, 423-437.
- Buck, J. L. (1990). On testing for variance effects. *Teaching of Psychology*, *17*, 255-256.
- Hubbard, R. & Armstrong, J. S. (1997). Publication bias against null results. Psychological Reports, *80*, 337-338.
- Hunter, J. E. (1997). Needed: A ban on the significance test. Psychological Science, *8*, 3-7.
- Landauer, A. A., Harris, L. J., & Pocock, D. A. (1982). Inter-subject variances as a measure of differences between groups. International Review of Applied Psychology, *31*, 417-423.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin, *86*, 638-664.
- Scarr, S. (1997). Rules of evidence: A larger context for the statistical debate. Psychological Science, *8*, 16-17.
- Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. American Psychologist, *47*, 1173-1181.
- Shea, C. (1996). Psychologists debate accuracy of significance test. The Chronicle of Higher Education, *49*, (August 16), A12 & A17.
- Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. Psychological Science, *8*, 1-2.