A Shortcut to Rejection: How Not to Write the Results Section of a Paper

David L Streiner, PhD[1]

This article discusses common errors in writing up the results of papers. It covers the following: 1) giving details about irrelevant topics, such as what program was used to enter the data, while ignoring important ones (for example, which options were chosen for various statistical tests); 2) reporting *P* levels of 0.0000 and negative values for *t* tests; 3) giving the *P* levels but not the actual values of the statistical tests; 4) not including confidence intervals and measures of the magnitude of an effect; 5) testing irrelevant hypotheses, such as whether reliability or validity coefficients are significantly different from zero; 6) attributing reliability and validity to tests rather than to the circumstances under which they are given; and 7) reporting levels of accuracy that cannot be supported by the data. Suggestions are made regarding the proper way to report findings.

(Can J Psychiatry 2007;52:385–389)

Information on funding and support and author affiliations appears at the end of the article.

Highlights

- Despite what the computer says, *P* levels should never be reported as 0.0000 (as opposed to < 0.0001) and should not appear without the associated test statistic; further, the number of decimal places should reflect the amount of accuracy that the sample size can support.
- Estimates of parameters (for example, correlations and ORs) should always be accompanied by CIs, and the results of tests of hypotheses should be accompanied by effect sizes.
- Reliability and validity are not fixed attributes of a test and should not be reported as such; they are dependent on the sample and the situation. The magnitude of the correlation or effect size is important in reporting results of reliability and validity studies; the *P* levels are irrelevant.

Key Words: *results, writing style, statistics*

Benjamin Franklin told us that the only sure things in life were death and taxes. What that proves is that he lived in the days prior to peer-reviewed publications; otherwise, he would have written, “But in this world, nothing can be certain, except death, taxes, and rejection for badly written Results sections of papers.” That is not to say that other sections of research articles are immune to criticism. After all, articles can be rejected for what is written, or not written, in every section of a paper. For example, the Introduction can present a one-sided or biased view of a controversial area; it can cite articles that are out of date; and worst of all, it can omit important ones (that is, those written by the reviewers). The Methods section is also a rich area for reviewers to mine for criticism. Entire books have been written about designing various types of studies, and this has several implications. First, it illustrates just how many things must be borne in mind at every stage of a project and that therefore can (and likely will) go wrong. Second, it means that every reviewer has become a maven and feels competent, if not obligated, to comment about the design and execution of the study. The Discussion section is where the authors interpret their results and give their implications (concluding, needless to say, that “further research is needed”). This gives the reviewers ample opportunity to lambaste the authors on their misinterpretation of what they found and ridicule the conclusion that the world will never be the same after these results gain the recognition they so richly deserve.

However, it is in the Results section that authors can really excel in demonstrating their lack of understanding of research and statistics. This article is not a tutorial on how to correctly analyze data—no single article can do this, and there are many books available that do a good job. Rather, it focuses on some of the more common mistakes researchers make in presenting their findings, which serve as red flags signalling to reviewers that the authors are out of their depth. Sadly, these examples are not made from whole cloth but are based on my experience reviewing papers for more than 40 journals over a span of some 35 years. It is not a comprehensive list of things that can go wrong but is, rather, highly idiosyncratic and based solely on the criterion of what drives me up the wall. It is a tutorial about what to do to increase the chances that your article will be rejected. Sometimes, my pet peeves contradict editorial policy. In these cases, I will tell you what I think is wrong but also what you should do to keep benighted editors happy (a category that of course does not include the editor of this journal).

Give Details About Data Entry, Not Analysis

Although this is a problem seen more often in grant applications than in articles, it's not unusual to run across paragraphs similar to this:

The data were entered into Program X, Version 2.03.01 (precise reference given) and analyzed with Program Y, Version 14.9b (another precise reference). Univariate and multivariate tests were used to test for differences between the groups.

The preceding 2 sentences give too much information about trivia and not enough regarding important issues. Some journals require you to state the statistical package and version you used, but for the life of me, I don't have the foggiest idea why. It's important if you're running highly sophisticated analyses, such as structural equation modelling, item response theory, or cluster analysis, because unfortunately, different programs can give different results. However, for the vast majority of statistical tests, the program is irrelevant: they all calculate the mean or do an ANOVA the same way and come up with identical answers. Even more meaningless is what program was used to enter the data. A 4.2 is a 4.2, whether that number was entered into a spreadsheet, into a dedicated data entry program, directly into the statistical program, or even with a word processor. Telling me which program you used is about as relevant as saying whether you used a pencil or a pen to record what the subject said.

Abbreviations used in this article

ANOVA	analysis of variance
CI	confidence interval
OR	odds ratio
SD	standard deviation

As a reviewer, what I want to know are the details of the tests that you used. This isn't too much of an issue for the simple univariate tests because you don't have too many options to choose from. However, if you do have options, I want to know which ones you chose and why. For example, if you do a factor analysis, you have a choice of many different extraction methods (and the defaults in many programs are the wrong ones), a larger number of rotation methods, and different ways of determining how many factors to retain. Similarly, if you ran a stepwise regression (although about 98.34% are done for the wrong reason),[1] I want to know what criteria were used for entering and deleting variables. My job as a reviewer is to determine whether you did things correctly or whether you screwed up somewhere; to do so, I have to know what you did. I'm not a very trusting soul (this being the prime requisite for a reviewer, trumping even expertise); if you don't indicate what you've done, I'll assume either that you're unaware of the implications of the various options or that you're trying to hide something. At best, the article will be sent back with a request for clarification and rewriting; at worst, it will be rejected out of hand. The bottom line is that you should report the relevant details of the analyses, not the irrelevant ones.

Report *P* Levels of Zero

P levels of statistical tests often determine whether an article will be submitted to a journal[2] and, if submitted, whether or not it is accepted.[3] Consequently, articles are replete with them, often reported poorly. If you truly want to demonstrate to the reviewer that you do not understand statistics, the best way is to report the *P* level as $P = 0.0000$. You may want to take some comfort from the fact that this is how many of the most widely used computer programs indicate highly significant results. However, it merely shows that, contrary to being "giant brains"—the name applied to the old mainframes—computers in fact manifest many symptoms of organic brain syndrome. They are concrete, literal, and sadly deficient in logical reasoning (although one wonders if this description should be applied to the computers or to the people who program them). The only things that have a probability of zero are those that violate one of the basic laws of nature: travel that is faster than the speed of light, perpetual-motion machines, or politicians who keep campaign promises. For everything else in the world, and especially study results, probabilities may be extremely small, but they are never zero. Therefore, do not report *P* levels as 0.0000; they should be given as less than some value. Very often, they're written as $P < 0.0001$, but later, we'll see why it's usually more accurate to say $P < 0.01$.

Report Naked *P* Levels

Almost as bad as *P* levels of zero (especially to 4 decimal places) are *P* levels that appear alone, unsullied by any association with the results of a statistical test. If you've run a *t* test or an ANOVA, you must report the value of the test and the degrees of freedom, in addition to the *P* level. For multiple regressions, the minimum necessary information is to report the standardized and unstandardized coefficients, their standard error, the *t* test for each variable, and the overall multiple correlation. Simply saying that the regression was significant (even when reporting that $P = 0.0000$) is meaningless, especially when the sample size is large, because even very small multiple correlations can be statistically significant.

Other statistical tests are reported in different ways, and you should know what they are. If you're unfamiliar with reporting the results of multivariate statistics, treat yourself to a copy of Tabachnick and Fidell's excellent book.[4] Each chapter ends with an illustration of how to do just this.

Don't Give CIs

Until about 5 or 10 years ago, journals were quite content if you reported the value of some parameter (for example, a mean, a proportion, or an OR) and the result of any tests of significance. However, there was growing discontent with this practice, especially among statisticians, for 2 reasons: first, the dichotomous decision of significant (if *P* were equal to or less than 0.05) or not significant (if *P* were greater than that value); and, second, the recognition that neither the results of the test nor the significance level addressed the issue of the magnitude of the effect. Regarding the first point, many researchers clung (and unfortunately, in the case of drug regulatory agencies, still cling) to the naive belief that a phenomenon doesn't exist if *P* is 0.051 but suddenly appears if *P* is 0.049—ignoring the facts that probabilities exist along a continuum and that the criterion of 0.05 is a totally arbitrary one chosen by Sir Ronald Fisher. Nevertheless, as Rosnow and Rosenthal[5] said, "Surely God loves the .06 nearly as much as the .05." The *P* level reflects the probability that, if the null hypothesis is true, these results could have arisen by chance. If *P* is less than 0.05, they still could have arisen by chance, and conversely, if *P* is greater than 0.05, the findings could be real ones. The 0.05 criterion is an agreed-upon convention and not a reflection of the way the world operates: results don't pop into and out of reality like virtual particles as the *P* level changes. What changes is not the results themselves but our confidence in the results.

The second point is that neither the value of a statistical test nor the associated *P* level says anything about the clinical importance of a finding. If you were to see that the results of an ANOVA were $F_{2,128} = 3.25$, $P = 0.04$, you wouldn't know how much variance the grouping variable accounted for or whether the effect was a large or a small one. Similarly, a significant OR of 2.1 could generate a lot of interest in the phenomenon if it were a good estimate of the population value, but it would elicit yawns if the estimate were only a rough approximation.

After much discussion, the American Psychological Association published guidelines for reporting the results of statistical tests.[6] There were 2 very strong recommendations: all point estimates of a parameter should be accompanied by 95%CIs, and whenever possible, statistical tests should be accompanied by an effect size. For example, a multiple correlation would report the value of R^2, an ANOVA would give the value of η^2 (eta-squared) or ω^2 (omega-squared) in addition to *F*, and differences between means would be accompanied by the standardized mean difference (to find out how to calculate these, see[7–10]). All psychology journals and many medical journals, including this one, have adopted these guidelines.

Report Negative Values of *t*

As long as we're on the topic of reporting results, avoid another travesty perpetrated by mindless computers: reporting negative values for the results of *t* tests. The minus sign appears in the output because the program subtracts the mean of group 2 from that of group 1. If group 2's mean is larger, then the result is negative. However, what is called group 1 and what is called group 2 is totally arbitrary. We can label the treatment group "1" and the comparison group "2," or vice versa, and nothing will change except the sign of the *t* test, so the sign tells us absolutely nothing. It's even more meaningless when it's reported in a table or in the text without any indication of which group is which—so, lose the sign.

Commit Type III Errors

We are all familiar with type I and type II errors: the first erroneously concludes that there is a statistically significant effect when in fact there isn't one, and the second is the converse of this—falsely concluding that there is no effect when one is actually there. There is also what we[7] have called a type III error—getting the right answer to a question that no one is asking. Perhaps the most egregious (but also the most common) example of this is testing whether a reliability or a validity coefficient is significantly different from zero. There are 2 reasons why this is asking the wrong question. First, unless you have done something terribly wrong in the execution or analysis of the study (for example, correlating the indvidual's social insurance number with his or her telephone number), I guarantee that the correlation will be greater than zero: never forget Meehl's sixth law,[11] that everything is correlated with everything else (usually around 0.30). Thus a significant correlation by itself is no guarantee of a Nobel Prize. More important, however, that's not what the reader needs to know. The important issue is the magnitude of the correlation, not its statistical significance. A correlation of 0.50 will be statistically significant if there are at least 16 subjects in the study, but I hope nobody would trust a scale if the test–retest reliability were that low.

People often test for a correlation that is significantly different from zero under the false impression that the word "null" in null hypothesis means "nothing." As Cohen[12] has pointed out, however, this is actually the "nil" hypothesis; the null hypothesis is the hypothesis to be nullified. Now, in many cases, the 2 are the same: we want to test whether some parameter, such as a difference between means, is bigger than zero—but it needn't be so, and in the case of reliability and validity coefficients, it shouldn't be so. If I had my druthers, people wouldn't even bother to report the *P* levels because that's a type III error. However, I doubt I'll win my battle over this one, so the most sensible course is not to test whether the correlation is greater than zero but whether it's larger than some unacceptable value, say 0.60. Better yet, I would simply calculate a CI around the parameter and check to see that the lower bound exceeds this minimal value. This would satisfy reviewers and editors as well as the requirement that parameters be accompanied by CIs.

Say That the Test Is Reliable and Valid

While we're on the topic of reporting reliability and validity coefficients, a very nice way of demonstrating your ignorance of scales is to talk about the reliability and validity of a test, as if it were an immutable property that, once demonstrated, resides with the test forever. Actually, it's not just ignorance you're showing but also the fact that you're more than a quarter of a century out of date. Until the 1970s, it was in fact common to talk about a test's reliability and validity, to say that validity is the determination of what a scale is measuring. However, that changed dramatically following a series of articles by Cronbach[13] and Messick,[14] in which the focus of validity testing shifted from the test to the interaction between the test and the specific group of people completing it. As Nunnally said, "Strictly speaking, one validates not a measurement instrument but rather some use to which the instrument is put."[15, p 133] For example, a scale of positive symptoms of schizophrenia may show good validity when used with English-speaking, non-Native individuals. However, it is invalid when used with those who have been raised in some traditional First Nations cultures where it is quite appropriate to hear from and speak to deceased relatives.[16]

Similarly, a test can show good reliability when it is used with groups that manifest a wide range of the behaviour being tapped. However, because of restriction in the range of scores, the same test will have much poorer reliability when used with more homogeneous groups of people.[17]

Table 1 An example of a table reporting baseline differences between groups

Variable	Group A	Group B
Number of women/men	9/11	10/10
Age (SD)	35.45 (6.00)	37.25 (5.52)
Education (SD)	13.25 (4.11)	12.50 (4.04)

The bottom line is that reliability and validity are not inherent properties of a test. You cannot say a test is reliable or valid: you have to demonstrate that these psychometric characteristics obtain for the group in which you are using it.

Be Inaccurate With Too Much Accuracy

In the previous section, I said that reporting a highly significant result as $P < 0.01$ is more accurate than reporting it as $P < 0.0001$. This seems paradoxical because more decimal places usually reflect a higher degree of precision. After all, 0.3333 is a closer approximation to 1/3 than is 0.3. However, the issue is whether you have enough subjects and have gathered the right information to justify a large number of digits to the right of the decimal point. Take a look at Table 1, which is much like the first table encountered in most articles, comparing 2 groups in terms of their baseline characteristics. Age and education are both reported to 2 decimal places. Is that degree of accuracy warranted? For age, that second decimal place represents 1/100th of a year, or just under 4 days. If you ask an individual how old he or she is (and assuming the individual rounds to his or her nearest birthday and doesn't lie), the average response is accurate only to within plus or minus 3 months—that's over 90 days' worth of error! Stating that you know the average age of the study participants to within half a week is bordering on the delusional.

The situation is even worse insofar as education is concerned. Assuming that the average school year is 200 days long, the difference between 13.25 years of education and 13.26 years is 2 days in the classroom. Would you be willing to stand up in court and defend that degree of precision, when all you asked for is the highest grade completed? I thought not.

This problem is exacerbated when you realize that the sample size in this example is only 20 subjects per group. That means that 1 year of schooling for each individual changes that last digit by 0.05. Draw a new sample that is identical to the first, except that one individual of the 20 finished an additional year of school, and the mean increases to 13.30. Small sample sizes result in large variations from one sample to the next. With 20 subjects and an SD of 4.11, the 95%CI around the mean is plus or minus 1.8 years (that is, $1.96 \times 4.11 / \sqrt{20}$). In this case, even the first decimal digit is probably accuracy overkill. Thus the number of decimal places you report in a table should reflect sampling error.

Conclusions

I have tried to show how you can write Results sections in such a way as to almost guarantee that your article will be rejected. Unfortunately, it isn’t a comprehensive list: it is also possible, for example, to draw graphs that are misleading or distort what is—or more often, what is not—going on[7,18]; however, it’s enough to get started.

How you report your results reflects how well you understand statistics and, by implication, whether you are aware of the possible limitations of your results. If you commit any of these, or other, errors, you will be signalling to the reviewer that something is amiss—that you simply pressed the compute button without being fully aware of what you were doing or what the results may mean. In *The Doctor’s Dilemma*, Shaw had the crusty physician say, “I tell you, Cholly, chloroform has done a lot of mischief. It’s enabled any fool to be a surgeon.”[19] In the same way, desktop computers and the ready availability of statistical packages have enabled anyone to be a statistician. Nevertheless, just as chloroform doesn’t take the place of training in surgery, so computers don’t obviate the need for knowledge of statistics. If you don’t have it, find someone who does. Don’t be afraid to ask—most statisticians are (relatively) tame and friendly.

References

1. Streiner DL. Regression in the service of the superego: the do’s and don’ts of stepwise regression. Can J Psychiatry. 1994;39(4):191–196.
2. Sutton AJ, Song F, Gilbody SM, et al. Modelling publication bias in meta-analysis: a review. Stat Methods Med Res. 2000;9(5):421– 445.
3. Olson CM, Rennie D, Cook D, et al. Publication bias in editorial decision making. JAMA. 2002;287(21):2825–2828.
4. Tabachnick BG, Fidell LS. Using multivariate statistics. 4th ed. Boston (MA): Allyn and Bacon; 2001.
5. Rosnow RL, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. Am Psychol. 1989;44(10):1276 –1284.
6. Wilkinson L, Task Force on Statistical Inference, APA Board of Scientific Affairs. Statistical methods in psychology journals: guidelines and explanations. Am Psychol. 1999;54(8):594 – 604.
7. Norman GR, Streiner DL. Biostatistics: the bare essentials. 2nd ed. Toronto (ON): BC Decker; 2000.
8. Fern EF, Monroe KB. Effect size estimates: issues and problems in interpretation. J Consum Res. 1996;23(2):89–105.
9. Fleiss JL. Estimating the magnitude of experimental effects. Psychol Bull. 1969;72(4):273–276.
10. Hojat M, Xu G. A visitor’s guide to effect sizes: statistical significance versus practical (clinical) importance of research findings. Advances in Health Sciences Education .2004;9(3):241–249.
11. Meehl PE. Why summaries of research on psychological theories are often uninterpretable. Psychol Rep. 1990;66(1):195–244.
12. Cohen J. The earth is round (p < .05). Am Psychol. 1994;49(12):997–1003.
13. Cronbach LJ. Test validation. In: Thorndike RL, editor. Educational measurement. Washington (DC): American Council on Education; 1971.
14. Messick S. Test validity and the ethics of assessment. Am Psychol. 1980;35(11):1012–1027.
15. Nunnally J. Introduction to psychological measurement. New York (NY): McGraw-Hill; 1970.
16. Hoffmann T, Dana RH, Bolton B. Measured acculturation and MMPI-168 performance of Native American adults. J Cross Cult Psychol. 1985;16(2):243–256.
17. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. Oxford (GB): Oxford University Press; 2003.
18. Streiner DL. Speaking graphically: an introduction to some newer graphing techniques. Can J Psychiatry. 1997;42(4):388–394.
19. Shaw B. The doctor’s dilemma. Lenox (MA): Hard Press; 2006. Act 1.

Manuscript received July 2006, revised, and accepted November 2006.
This is the 27th article in the series on Research Methods in Psychiatry. For previous articles please see Can J Psychiatry 1990;35:616–620, 1991;36:357–362, 1993;38:9–13, 1993;38:140–148, 1994;39:135–140, 1994;39:191–196, 1995;40:60–66, 1995;40:439–444, 1996;41:137–143, 1996;41:491–497, 1996;41:498–502, 1997;42:388–394, 1998;43:173–179, 1998;43:411–415, 1998;43:737–741, 1998;43:837–842, 1999;44:175–179, 2000;45:833–836, 2001;46:72–76, 2002;47:68–75, 2002;47:262–266, 2002;47:552–556, 2003;48:756–761, 2005;50:115–122, 2006;51:185–191, 2007;52:121–128.

[1] Director, Kunin–Lunenfeld Applied Research Unit, Baycrest Centre, Toronto, Ontario; Professor, Department of Psychiatry, University of Toronto, Toronto, Ontario.
Address for correspondence: Dr DL Streiner, Kunin–Lunenfeld Applied Research Unit, Baycrest Centre, 3560 Bathurst Street, Toronto, ON M6A 2E1; dstreiner @ klaru-baycrest.on.ca

Résumé : Un raccourci au refus : comment ne pas écrire la section « résultats » d’un article

Cet article présente les erreurs fréquentes commises en écrivant les résultats des articles. Il porte sur ce qui suit : 1) donner des détails sur des sujets non pertinents, comme le programme utilisé pour saisir les données, en ignorant les sujets importants (par exemple, quelles options ont été choisies pour les divers tests statistiques); 2) inscrire des niveaux *P* de 0,0000 et des valeurs négatives pour les tests *t*; 3) donner les niveaux *P* mais pas les valeurs réelles des tests statistiques; 4) ne pas inclure les intervalles de confiance ni les mesures de l’ampleur d’un effet; 5) vérifier des hypothèses non pertinentes, comme estimer si les coefficients de fiabilité ou de validité sont significativement différents de zéro; 6) attribuer la fiabilité et la validité à des tests plutôt qu’aux circonstances dans lesquelles ils sont donnés; et 7) déclarer des niveaux d’exactitude que ne peuvent appuyer les données. Des suggestions sont faites concernant la manière adéquate de rapporter les résultats.