Making friends with your data: Improving how statistics are conducted and reported[1]

Daniel B. Wright*

Psychology Department, University of Sussex, UK

Aim. This paper highlights some of the areas where there are problems with the way that statistics are conducted and reported in psychology journals. Recommendations are given for improving these problems.

Sample. The choice of topics is based largely on the questions that authors, reviewers, and editors have asked in recent years. The focus is on null hypothesis significance testing (NHST), choosing a statistical test, and what should be included in results sections.

Results. There are several ways to improve how statistics are reported. These should improve both the authors' and the readers' understanding of the data.

Conclusions. Psychology as a discipline will improve if the way in which statistics are conducted and reported is improved. This will require effort from authors, scrutiny from reviewers, and stubbornness from editors.

In a lecture to the Institution of Civil Engineers, Lord Kelvin said: 'When you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind' (3 May, 1883, as cited in Thompson, 1910/1976). Psychologists have learned how to express their findings in numbers, but often the expression is neither clear nor convincing, leaving the reader bewildered at the author's intent. There are two related problems. The first is whether the author reports the statistical results clearly and in a manner suitable for the typical reader. The second is whether the statistics conducted are appropriate. An American Psychological Association (APA) task force produced a report (Wilkinson *et al.*, 1999) on these issues which is worth consulting. Dozens of other papers have addressed these concerns (see http://www.coe.tamu.edu/~bthompson/ for examples).

---

[1] The title is based on a phrase Robert Rosenthal gave in an interview; details are in the paper's final paragraph.

* Requests for reprints should be addressed to Daniel B. Wright, Psychology Department, University of Sussex, Brighton BN1 9QK, UK (e-mail: DanW@cogs.susx.ac.uk).

As the statistics editor of *BJEP* I am often asked statistical questions. The aim of this paper is to provide suggestions to address these issues and also to impart a flavour of how data analysis should be conducted and reported for *BJEP* and other psychology journals. It would be easy, but not helpful, just to criticise how people report statistics. Instead, I try to provide direction on how to improve the reporting of data. I have organised this paper around three themes: null hypothesis significance testing, choosing a statistical test, and presenting results. There are several additional topics that could have been included. My sampling was based on those questions raised by *BJEP* authors, editors, and reviewers.

Null hypothesis significance testing (NHST) (the p value approach)

The dominant approach for statistical analysis in psychology is to report p values and use these to reject or not reject hypotheses (Hubbard & Ryan, 2000). Most students and researchers wrongly believe that the p value is a probability about the null hypothesis (Oakes, 1986; Wright, 2002b). It is not. It is the probability of observing data as extreme (or more extreme) as observed assuming that the null hypothesis is true. Numerous authors have pointed out the limited value of this approach and that if people understood what p was the probability of, they would use this approach less (e.g., Carver, 1978; Schmidt, 1996). Even those who feel that there is a place for NHST agree that it is overused and often misused (see papers in Harlow, Mulaik, & Steiger, 1997, and Chow, 1998, for discussion).

Improvement: NHST and p values should be used less and it should be made clear that they are of limited value. If NHST is used, the word 'significant' should be avoided if referring to a result with $p < .05$ because it implies an effect is of practical value. Words like 'reliable' and 'robust' are also problematic because they suggest more than what $p < .05$ implies. The word 'detected' is closer to actual meaning, but is still not perfect. Thompson (1996) suggests always prefacing 'significant' with 'statistically'. This makes it clear that $p < .05$ does not mean 'significant' in its standard English usage, but it does not say what it means. Kirk (1999, p. 337) provides a useful definition: 'In the simplest terms, a statistically significant result is one for which chance is an unlikely explanation'. There is no simple phrase that properly conveys this information, but 'detected' or 'statistically significant' are improvements over just 'significant'.

Researchers should consider alternatives to NHST. This may include substantial changes like using Bayesian analysis (see Gill, 2002, for introduction), which brings some of its own problems, or it may simply be reporting more information to help the reader. Researchers should avoid the dichotomous reject/not reject aspect of NHST, should report effect sizes, should calculate confidence intervals, and should conduct power analyses. These topics are discussed below.

Further reading: Cohen (1994); Thompson (1996); Wilkinson *et al.* (1999).

(1) Should $p < .05$ or $p = .03$ be reported?

Before computers, people (who were actually called 'computers') painstakingly produced tables of critical values for selected probability levels, usually 1%, 5%, and 10%, for different test statistics (Salsburg, 2001). Researchers would report $t(47) = 2.25$, $p < .05$. This tells the reader that the actual p value could be anywhere from 0 to 5%. Modern computers produce more precise p values. The question is: Why report the

range 0 to 5% when the t value and the degrees of freedom allow anyone with Excel, SPSS, etc., to calculate that $p = .03$?

Improvement: Report $p = .03$ rather than $p < .05$. This is a move away from strict reliance on the convention of .05 and the dichotomous decision-making aspect of NHST. There is some disagreement about what should be reported for extremely small probabilities. The computer often prints p as .000. Some people argue that printing $p = .00$ or $p = .000$ is fine because readers will realise that the probability is a very small number and has been rounded down from a number greater than zero. Other people, including myself, dislike seeing $p = 0$ because it may suggest to some readers that the probability is absolute zero; that it is an impossibility. We prefer reporting $p < .001$ in these situations.

(2) Which measures of effect size should be used?

Many scientific journals, including APA journals, require reporting the size of any effect. Effect sizes tell the reader how big the effect is, something that the p value does not do. The purpose of reporting effect size is to communicate to the reader the size of the effect and to allow comparisons with other effects. Thus, the measure should be meaningful, understood by others, and comparable with other effect sizes. To be meaningful, it is important to report the units of measurement of the effect size. There are different ways to classify different measures of effect size. Here two distinctions are made. The first is whether the effect is reported in units of the original variables or in standardised units. For example, it is often valuable to report that the mean on a test is 3 correct answers higher in one group than in another, or that an increase of 1 hour of homework per week is associated with getting 2 more correct answers. However, to compare across studies that use different dependent measures it is often valuable to report that one group scores, on average, 1 standard deviation higher than another, or that there is 25% shared variance between homework time and scores on the test. Most of the discussion about effect sizes relates to these standardised measures. A second useful distinction (Kirk, 1996; Richardson, 1996) is between effects for the differences between group means and effects in terms of proportion of variation or association. There are several measures for each of these. Because all can these can be transformed into proportion of variance, some researchers say that standardised measures based on proportion of variation should be used (e.g., Field & Hole, 2003).

Measures of effect size for categorical variables are often treated differently than those for metric variables. There are several different measures. The nature of categorical measures is that the most meaningful units are the percentage of cases in each category or the odds of a case being in a category. Suppose the odds of passing an exam for one group is 3 to 1 (three passes for every fail), and for another group is 2 to 1 (two passes for every fail). The ratio of the odds, $3/2 = 1.5$, is a good choice for measure of association. This is called the odds ratio and is appropriate for 2×2 contingency tables and is used in more complex loglinear models. Another popular measure for more complex contingency tables is Cramer's V. It has the advantage of being based on the well known χ^2 statistic and being related to Pearson's r statistic.

Improvement: Report effect sizes with the appropriate units. Often the effect size should be reported both in the units of the original variable and in standardised units like the correlation.

Further reading: Kirk (1996), Rosenthal, Rosnow, and Rubin (2000).

(3) Confidence intervals are not available from my statistics program?
Confidence intervals give a region which usually includes the true population value of the parameter. They are useful because they both show the estimate of the parameter and indicate how precise this estimate is. Like NHST they can be used to test specific point hypotheses (if the interval does not include a particular point, then NHST would reject that hypothesis), but they can also be used when there is no specific hypotheses and can be used for comparing results across studies.

Sometimes it is difficult to find confidence intervals for everything that you report. Some are reported in the main statistical packages, but for many measures free software can be downloaded to calculate these (see below). It is probably best to use a search engine, like google.com, to find exactly what you are looking for. Others, like within subject confidence intervals, can be calculated using the main statistical packages with a few additional calculations (Loftus & Masson, 1994). Confidence intervals have traditionally just been for means and other parameter estimates, not effect sizes. The computation for confidence intervals of effect sizes is more difficult, but they are now becoming more popular (see Thompson, 2002, for a review).

An alternative is using *bootstrapping* to calculate confidence intervals. Boot-strapping can be used to calculate confidence intervals for anything. It involves taking hundreds or thousands of 're-samples' from the observed data and calculating the statistic (e.g., the mean, the correlation) each time. With modern computers this can be done in seconds. The middle 95% of the estimates provides a confidence interval that is not as reliant on assumptions as confidence intervals calculated in traditional ways. Many statistics packages, like SPSS and SYSTAT, now offer bootstrapping options for some of their procedures.

Improvement: Report confidence intervals routinely. Often this can be done instead of using p values. The interval should be reported as, for example, 8±3 or (5, 11). Some confidence intervals, like those found through some forms of bootstrapping, are not likely to be symmetric, therefore the lower and upper bounds must be explicitly given.

Further reading: Efron and Gong (1983); Thompson (2002); Wright (2002b); http://glass.ed.asu.edu/stats/analysis/ for free software that calculates confidence intervals for many measures.

(4) How many participants should I use; when should I do a power analysis?
There are several considerations that a researcher must take into account when deciding how many participants to use in a study. Practical and financial issues are concerns. Also, it is important to consider what the likelihood of detecting an effect is. This is where power analysis can help.

The power of a test is the probability of rejecting a specific effect size for a specific sample size at a particular α level (ie., the critical level to reject H0). It should be used when calculating the number of participants used in a study and should be described in the Methods section of the paper. The convention is to have power of at least 80%. The standard α level is 5%. It is more difficult to decide the minimum effect size that you are trying to detect. Cohen (1992) lists small, medium, and large effect sizes, and researchers tend to use one of these. However, it is preferable to calculate an effect size based on past research and practical implications. Several statistics packages (e.g., SYSTAT, S-Plus) have modules for calculating power. There is also freeware available for this. G*Power is one of the most popular (website below).

Improvement: Power analysis should be used to calculate the number of participants

to be used in any study. Using power analysis will tend to increase the size of the samples, but will increase the number of effects detected, so it should be cost effective in the long term. Even when practical issues dictate how many participants are used, it is worth doing a power analysis so that the likelihood of detecting effects of different sizes is known. This can help prevent over-interpretation when effects are not detected.

Further Reading: Cohen (1992); http://www.psycho.uni_duesseldorf.de/aap/projects/gpower/ to download free software.

Choosing a statistical test

Many readers would like a list of the appropriate tests to run for every particular situation. This could be like the flow charts that appear in many introductory textbooks directing readers to the 'right' statistical test, providing they can answer questions about, for example, the level of measurement of the data. I am not going to do this here, but rather describe some of the factors that people often consider when choosing a test. The flow chart approach is too simple and can often be misleading. First, it suggests that answers to questions like whether the data are interval or what hypotheses are being investigated are easily answered. Second, it makes the tests appear very different from each other, when arguably researchers should focus on the commonalities of statistical tests (i.e., the generalised linear model). Third, and most important, it makes it sound as if there is one 'right' test. Most situations are not this straightforward.

(1) Are my data interval?

The level of measurement is not an inherent characteristic of a particular variable, but a characteristic that we, as researchers, bestow on it based on our theories of that variable. It is a belief we hold about the variable. We act as if the data are of a certain level, and we therefore need to convince readers that this is an appropriate assumption. Because interval level means the distance between points 2 and 4 on some scale is the *exact* same as between 5 and 7, some have argued that to have precisely this equality is impossible and therefore is never true. However, because level of measurement is something inside the researcher's head, s/he is perfectly welcome to believe this particular equality, and therefore act as if the data *are* interval even if for another researcher this would not be appropriate. Lord (1953) and Wright (1997) provide examples of this.

How does level of measurement relate to the choice of statistical test? Some statistical tests are only *meaningful* for data of a particular level of measurement. For example, calculating a mean implies interval data. We can calculate the mean of any numerical variable, but sometimes this statistic will not convey what the researcher wants.

Improvement: Methodologists differ on how important they think level of measurement is for determining a statistical test, so I will avoid making any strong recommendations. Make sure that you can justify your choice of level and consider alternatives. Many recent rank-based statistics, sometimes called R estimators, have been designed for ordinal data and these are improvements over earlier procedures (Cliff & Keats, 2002; Hettmansperger & McKean, 1998).

Further reading: Cliff and Keats (2002); Lord (1953); Wright (1997).

(2) Does it matter if my data are not normally distributed?

'Experimentalists think that [the Normal distribution] is a mathematical theorem while the mathematicians believe it to be an experimental fact.' (Gabriel Lippman, 1845-1921, as cited on http://math.furman.edu/~mwoodard/mquot.html)

Many statistical procedures assume that the data arise from a Normal distribution and these procedures are popular because of the beliefs expressed by the physics Nobel Laureate Gabriel Lippman. The facts are that Normal distributions are rare in psychology (Micceri, 1989) and that minor deviations from Normality can affect these statistics (Tukey, 1960). Since 1960 numerous procedures (for example M estimators, trimmed statistics, see Hoaglin, Mosteller, & Tukey, 2000) have been developed that not only are more robust, meaning that they are less affected by a small number of outliers, but are also usually more powerful than their Normal-based counterparts (see Wilcox, 2001, for an introduction and Wilcox, 1997, for more details). This means that by using these procedures researchers are less likely to report rogue findings due to a few errant outliers and are more likely to detect actual effects. Thus, the shape of the distribution should greatly affect your choice of statistical test!

These are newer procedures. Many statistical programs do not incorporate them (though S-Plus is excellent), and there are not robust alternatives available for every technique psychologists use. Transformations can often be used to make the distributions more Normally distributed, but these affect the measurement level of the variable.[2] An example transformation is ranking the variables. However, researchers must be cautious in their interpretations. Inferences about transformed variables may not be valid for their untransformed counterparts.

Improvement: If you are worried about a few points, run the statistical tests with and without these points. If you reach the same conclusion then you should feel more comfortable with your results. There are several free robust libraries, including those in Wilcox (2001), which can run on the package S-Plus and the freeware package R (see http://www_rcf.usc.edu/~rwilcox/). It is worth exploring these techniques. Their use is greatly increasing.

Further reading: Wilcox (1998, 2001).

(3) When should I use a median split?

The answer is probably never. To understand why, consider the following situation. You are interested in the relationship between hours of study per week and performance. Figure 1a shows a scatterplot of hypothetical data between these variables and some related statistics. These statistics are what the introductory textbook flowcharts would suggest. The scatterplot shows a positive relationship between hours studied and performance. There is an increase in the performance estimate of about 1 point for each extra hour studied. When doing a median split, all values less than the median, here 10.5 hours, are treated the same and all those above the median are treated the same. This is depicted in Figure 1b, where the number of hours variable has been collapsed into two categories. You might give the value 0 to those below the median and 1 to those above the median. The model being estimated assumes that there is no difference between studying 1 hour and 10 hours, and no difference between

---

[2] *If you think a variable is interval, then in general the transformed variable will not be interval. For example, if you believe that a 1–7 attitude scale is interval, this means the difference between 1 and 2 is the same as between 4 and 5. If your transformation squares the variable, as might be used with negatively skewed data, the difference between 4 and 5 is now three times the difference between 1 and 2 because $5^2 - 4^2 = 25 - 16 = 9$ and $2^2 - 1^2 = 4 - 1 = 3$.*

studying 11 hours and 20 hours. The predicted performance score for someone studying for 2 hours is 63 points, for 10 hours is 63 points, for 11 hours is 71.5 points, and for 18 hours is 71.5 points. There may be examples where the model tested with a median split is appropriate, but these are rare.

There are situations where a continuous variable should be dichotomised. For example, you may record a person's age but only be interested in whether the person is old enough to vote. Similarly, sometimes there is step function, where above a certain threshold people perform differently than below this threshold. However, it is unlikely that it is exactly at the median where this step takes place. The level of measurement implied by median splits, where all values either side of the median are equal, is very peculiar. As stated above, the level of measurement is in the researcher's head so it may be that this particular level is appropriate, but it would be difficult to argue for it. There are procedures for deciding if a continuous variable or set of continuous variables should be collapsed into categories (Waller & Meehl, 1998).
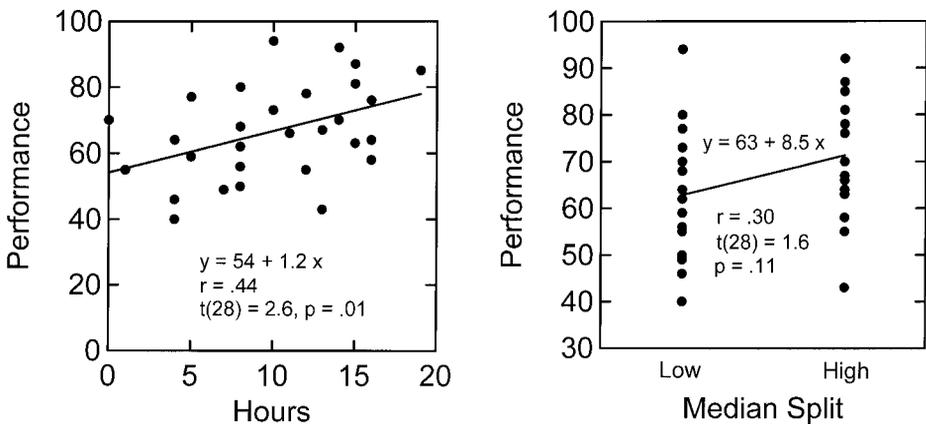

Figure 1. Scatterplots comparing academic performance with hours studied. Figure 1a shows the original hours variable ranging from 0 to 20 hours. In Figure 1b the data are collapsed into those below the median and those above the median. The linear regression with accompanying statistics are included with both. For Figure 1b, the linear regression is identical to a *t* test

Improvement: Avoid using median splits. This often means using a regression procedure rather than an ANOVA. Mathematically these are equivalent (Cohen, 1968). Running the regression will often require computing new variables for interactions and this should be done carefully (Jaccard, Turrisi, & Wan, 1990).

Further reading: Cohen (1983); Waller and Meehl (1998).

(4) Do I really need to learn all these fancy new statistics?

Statistics is a lively discipline with new techniques being published frequently. Some of these techniques, like structural equation modelling, multilevel modelling, item response modelling, etc., have become popular in *BJEP* and related journals. If the research design and questions require one of these 'fancy' techniques, then yes, you will need to use them. This may mean learning a new statistical technique, buying new software (though much freeware exists), or consulting a statistician. Sometimes, however, hypotheses can be addressed with simpler techniques. In these cases, the

simpler techniques should be reported and if appropriate the authors may report that the more advanced techniques led to similar conclusions. In general, the best designs require the simplest statistics.

Improvement: Do not be afraid to learn new statistical procedures. This may create delays for the publication and may cost money, but this is part of the publishing process. Plan which statistical techniques you may need while you are designing your study. Minor changes to the design can make the statistical analyses much simpler. Do not try to impress the reviewers with statistical complexities. If you use advanced statistics, beyond what the typical psychology undergraduate would know, make sure that these are clearly described.

(5) How do I analyse test re-test designs? - Lord's Paradox

The following example illustrates how it is important to match the particular model that you have in mind with the statistical test you report. This is true in all research. Here it is shown for a common example in education, measuring change for two groups.

Consider the following hypothetical study. You have two naturally existing groups of pupils, the stars and circles in Figure 2, and are interested in the effect of a reading programme. You measure reading at time 1 and then one of the groups, say the circles, takes part in the reading programme. You then measure reading at time 2. The question of how to analyse this simple design is at the heart of a paradox raised by Lord in 1967. In some situations it would be reasonable to conduct any of the following: a t test on scores at time 2, a t test on the difference between the scores at time 2 and time 1, an analysis of covariance partialling out scores at time 1, etc. All these relate to whether the reading programme makes a difference, but they make different assumptions and test different models (Hand, 1994). Here, the *stars* are about 40 points *higher* at time 2 (95% CI from 28.3 to 48.1 points, $t(18) = 8.1$, $p < .001$), but they clearly began with higher scores. This is unlikely to occur if you had randomly allocated people into conditions, but often this is not feasible. For example, the groups may be boys and girls, dyslexic and non-dyslexic pupils, etc. Running a t test on the difference, time 2 minus time 1, shows that *circles increased* by about 10 points while the stars did not increase (95% CI of difference between groups 1.1 to 22.7 points, $t(18) = 2.33$, $p = .03$). Running an ANCOVA, partialling out time 1 scores, shows the *stars* have predicted values about 14 points *higher* for any score at time 1 (95% CI from 4.7 to 23.3, $t(17) = 3.18$, $p = .005$). The effect appears in the opposite direction of the t test on the difference despite testing a similar hypothesis. All of these are accurate descriptions of the situation, but are asking different questions (Hand, 1994). Deciding whether the reading programme is causing a difference requires assumptions beyond the data (see Wainer, 1991, for details).

Improvement: The only procedure that is always correct in this situation is a scatterplot comparing the scores at time 2 with those at time 1 for the different groups. In most cases you should analyse the data in several ways. If the approaches give different results, as they do here, think more carefully about the model implied by each.

Further reading: Hand (1994); Lord (1967); Wainer (1991).

(6) Do the data fit my model?

Once you have conducted your analyses and created a model (whether the model is just no difference between groups or a complex equation) you need to evaluate whether
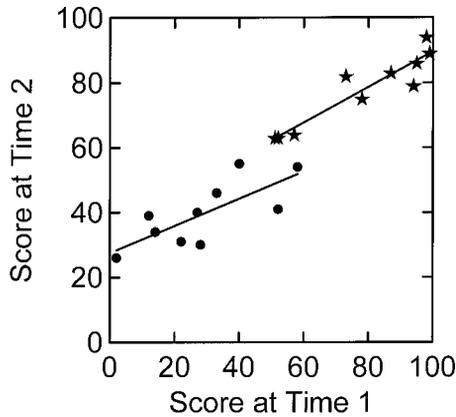
Figure 2. A scatterplot showing hypothetical data for two groups of children. These hypothetical data demonstrate Lord's paradox. A *t* test on the differences (time2 – time1) shows the improvement for the circle group is larger. An ANCOVA shows that once accounting for time 1 scores, the star group does better (i.e., its regression line is above the line for the circles)

the observed data fit the model. Figure 3 shows how a model can be evaluated. Suppose you are trying to determine if a linear regression is an appropriate model. First, you can see whether a non-zero correlation is detected or not (i.e., is $p < .05$). On its own this tells us little about the effect other than that one has been detected. Reporting the size of the correlation and the slope of the regression line in appropriate units are steps in the right direction. Next, you should make sure that the correlation is not due to just a few points, that it is robust. The scatterplot should be examined to make sure that the effect is linear and that there are no oddities in the data. Finally, the effect should be consistent with what is known about science. If your model implies gravity does not exist, then regardless of the first four tests of your data, your model is probably wrong.

Improvement: Go beyond the simple reject/fail to reject dichotomy and beyond the numerical output in evaluating your models. This will also help you to understand your data and models. Statistics is only one way to assess the value of your model.

Figure 3. Five ways for assessing the fit of a regression model. All should be used. Using just the ones on the left is discouraged. (Adapted with permission from Figure 8.6 of Wright (2002a) *First Steps in Statistics*. London: Sage Publications)

Presenting the results section

Even when the appropriate statistics are conducted, often the information is not adequately presented to the reader. The rules of style, both generally (Gibaldi, 1999; Strunk & White, 1979) and specifically to psychology (Sternberg, 2000), apply to results

sections as well as the rest of the paper. A results section should not be a list of the statistical analyses conducted, but a description of the data, highlighting the important aspects and providing the evidential bases for any conclusions. A second year undergraduate should be able to read the results section, on its own, and know what the main findings are.

Further readings: Gibaldi (1999); Sternberg (2000); Strunk and White (1979).

(1) Causal and associative hypotheses

Broadly speaking, there are two types of hypotheses that can be explored: causal and associative. Causal hypotheses imply that changing some aspect of the environment will tend to create some difference. In order to have a causal hypothesis it is necessary to think about manipulating some aspect of the system. Causal hypotheses are most easily investigated using experimental designs. Associative hypotheses describe how variables relate to each other in the absence of manipulation. Sampling is critical for investigating associative hypotheses. It should be made clear which type of hypothesis is being investigated. While some research combines these, investigating the situations where causal effects apply, the more common error is wrongly describing an associative hypothesis as causal. For example, you should not think of gender as having a causal effect (Holland, 1986).

Improvements: Make clear which type of hypothesis is being examined. The type of hypothesis examined is determined by the design and the researchers' theories, and not the statistical procedures (i.e., using structural equation modelling does not mean you are necessarily examining causal hypotheses, Thompson, 2000).

Further reading: Cook and Campbell (1979); Cronbach (1957).

(2) Reliability of scale scores

Psychometric tests are critical for much psychology and education research. There are two common complaints about manuscripts. First, often it is not clear what the psychometric test is. Second, the reliability of the sample data for the test is not reported. There are different kinds of reliability. Most often what is being referred to is the internal consistency of the scores on a particular scale.

Improvement: Say which psychometric tests you are using, give brief descriptions, and cite appropriate sources. Report the reliability for the sample data in the results section.

Further reading: Strube (2000); Thompson and Vacha-Haase (2000).

(3) Tables and graphs

Tukey's (1977) *Exploratory Data Analysis* made the science of pictorially representing data an important part of mainstream statistics (for example, Tufte, 2001). Good tables and graphs can convey information more clearly than text. Bad tables and graphs can confuse and mislead the reader. Bad graphs make the paper look unprofessional. Given the software available, there are no excuses for bad graphs. Seldom are the default settings for software acceptable. It may take several hours to make an acceptable graph. One rule that is often broken is when authors add useless information, like a false third dimension, to a graph (Fischer, 2000). This adds complexity without adding information.

Improvement: Carefully think about which type of graph will best illustrate the information that you wish to communicate. Graphs should communicate the information as simply, as clearly, and as accurately as possible. Take time to prepare good graphs.

Further readings: Wainer (1984); Wainer and Velleman (2001); Wright (2002a, Chapter 2).

(4) Descriptive statistics

Descriptive statistics – the basic means, percentages, standard deviations, etc. – should always be reported. These are often the most influential statistics for communicating the results to your readers. This can be done in tables, figures, or in the text. However, you should not report the same statistics more than once.

Summary

On the final day of my year-long statistics course I tell the blurry eyed students that there are four things that they need to remember from the course:

(1) Data = Model + Error

This equation can be made to look much more complex, but it is this simple equation that underlies all our statistics. For example, with a group *t* test Model is the null hypothesis that there is no difference between groups. The size and distribution of the Error tells us how well the data fit the model.

(2) Tell a story with your data; think about your audience; say it as simply as possible

Why is the results section of a paper *not* the most exciting part of the paper? The introduction and methods should leave readers on the edge of their seats. The results should relieve the suspense, announcing whatever results are so important that they deserve publication. The reason that most results sections are not exciting is because of the way they are written: statistical phrases are not well explained, trivial findings are not differentiated from crucial findings, and numbers are reported without any substantive meaning. The goal of all writing is communicating your ideas.

(3) Current practice is not good

The reason for this paper, the reason for the APA report (Wilkinson *et al.*, 1999), the reason for other similar reports, is that psychologists are not getting the most out of their data and are often misleading their audience. Some of the suggested improvements are:

    If a *p* value is reported, it should include an effect size and with confidence intervals,
    The units of measurement for the effect size should be reported where practical,
    Try to use the word *detected* or *statistically significant* rather than just *significant*,
    Provide rationale for the sample size (for example, with a power analysis),
    Carefully consider which statistical tests you report, and justify your decision,
    Make clear the nature of the hypotheses being explored,

Write clearly, as if for second year undergraduates.

I hope that authors follow this guidance, that reviewers remind authors about them, and that editors insist that results are reported as clearly and as accurately as possible.

(4) 'Make friends with your data' (Rosenthal, from Azar, 1999)

The final piece of advice that I leave students with is the most important. Robert Rosenthal was one of the co-chairs of the APA task force that looked into statistics reporting and he said that researchers should 'make friends with your data'. Researchers often spend months or years painstakingly designing a study and collecting data, and then throw their data into a computer and try to analyse it in minutes. The data deserve better. The quick and reckless approach to data analysis often fails to identify important aspects of the data. You should become friends with them! Conducting data analysis is like drinking a fine wine. It is important to swirl and sniff the wine, to unpack the complex bouquet and to appreciate the experience. Gulping the wine doesn't work.

Acknowledgements

References

Azar, B. (1999). APA statistics task force prepares to release recommendations for public comment. *APA Monitor Online*, *30 (5)*, http://www.apa.org/monitor/may99/task.html.

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

Chow, S. L. (1998). Precis of *Statistical significance: Rationale, validity, and utility* (with comments and reply). *Behavioral and Brain Sciences*, *21*, 169-239.

Cliff, N., & Keats, J. A. (2002). *Ordinal measurement in the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, *70*, 426–443.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7, 249-253.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cook, T. D., & Campbell, D. T. (1979). Causal inference and the language of experimentation. *Quasi-experimentation: Design & analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, *37*, 36-48.

Field, A. P., & Hole, G. (2003). *How to design and report experiments*. London: Sage Publications.

Fischer, M. H. (2000). Do irrelevant depth cues affect the comprehension of bar graphs? *Applied Cognitive Psychology*, *14*, 151-162.

Gibaldi, J. (1999). *MLA handbook for writers of research papers* (5th ed.). New York: Modern Language Association of America.

Gill, J. (2002). *Bayesian methods for the social and behavioral sciences*. New York: Chapman & Hall/CRC.

Hand, D. J. (1994). Deconstructing statistical questions. *Journal of the Royal Statistical Society: A*, *157*, 317–356.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds) (1997). *What if there were no significance tests?* Hove: Lawrence Erlbaum Associates.

Hettmansperger, T. P., & McKean, J. W. (1998). *Robust nonparametric statistical methods. Kendall's library of statistics 5*. London: Arnold.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.) (2000). *Understanding robust and exploratory data analysis*. New York: John Wiley & Sons.

Holland, P. W. (1986). Statistics and causal inference. *Journal of American Statistical Association*, *81*, 945–960.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology and its future prospects. *Educational and Psychological Measurement*, *60*, 661–681.

Jaccard, J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression*. London: Sage Publications.

Kirk, R. E. (1996). Practical Significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746–759.

Kirk, R. E. (1999). *Statistics: An introduction*. London: Harcourt Brace.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476–490.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, *8*, 750–751.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *72*, 304–305.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester: John Wiley & Sons.

Richardson, J. T. E. (1996). Measures of effect size. *Behavior Research Methods Instruments & Computers*, *28*, 12–22.

Rosenthal, R., Rosnow, R., & Rubin, D. B. (2000). *Contrast and effect sizes in behavioral research: A correlational approach*. Cambridge: Cambridge University Press.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: Henry Holt & Company.

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

Sternberg, R. J. (Ed.) (2000). *Guide to publishing in psychology journals*. Cambridge: Cambridge University Press.

Strube, M. J. (2000). Reliability and generalizability theory. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 23–66). Washington, DC: American Psychological Association.

Strunk, W. Jr., & White, E. B. (1979). *The elements of style* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26–30.

Thompson, B. (2000). Ten commandments of structural equation modelling. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261–283). Wash. DC: American Psychological Association.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 24–31.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics *is* datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174–195.

Thompson, S. P. (1976, originally published 1910). *The life of Lord Kelvin. Vol. 2* (2nd ed.). New York: Chelsea Publishing Company.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow & H. Mann (Eds.), *Contributions to probability and statistics* (pp. 448–485). Stanford, CA: Stanford University Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wainer, H. (1984). How to display data badly. *American Statistician*, *38*, 137–147.

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin*, *109*, 147–151.

Wainer, H., & Velleman, P. F. (2001). Statistical graphics: Mapping the pathways of science. *Annual Review of Psychology, 52*, 305–335.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua*. Thousand Oaks, CA.: Sage Publications.

Wilcox, R.R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*, 300–314.

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York: Springer-Verlag.

Wilkinson, L & the Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Wright, D. B. (1997). Football standings and measurement levels. *The Statistician: Journal of the Royal Statistical Society Series D*, *46*, 105–110.

Wright, D. B. (2002a). *First steps in statistics*. London: Sage Publications.

Wright, D. B. (2002b). *Confidence intervals, probability, and dead cats*. Manuscript submitted for publication. (Available from the author.)