ELSEVIER

Archives of Clinical Neuropsychology
20 (2005) 1053–1059

Archives of CLINICAL NEUROPSYCHOLOGY

Misuse of statistical tests in *Archives of Clinical Neuropsychology* publications

Philip Schatz *, Kristin A. Jay, Jason McComb, Jason R. McLaughlin

Saint Joseph's University, Department of Psychology, 222 Post Hall, Philadelphia, PA 19131, USA

Accepted 10 June 2005

Abstract

This article reviews the (mis)use of statistical tests in neuropsychology research studies published in the *Archives of Clinical Neuropsychology* in the years 1990–1992 and 1996–2000, and 2001–2004, prior to, commensurate with the internet-based and paper-based release, and following the release of the American Psychological Association's Task Force on Statistical Inference. The authors focused on four statistical errors: inappropriate use of null hypothesis tests, inappropriate use of *P*-values, neglect of effect size, and inflation of Type I error rates. Despite the recommendations of the Task Force on Statistical Inference published in 1999, the present study recorded instances of these statistical errors both pre- and post-APA's report, with only the reporting of effect size increasing after the release of the report. Neuropsychologists involved in empirical research should be better aware of the limitations and boundaries of hypothesis testing as well as the theoretical aspects of research methodology.

Keywords: Neuropsychology; Statistics; Effect size; Null hypothesis testing; Multivariate

The psychology and neuropsychology literature are replete with discussions of proper approaches to statistical testing, designing and implementing research paradigms, and the utility of various statistical operations. Specific guidelines for statistical reporting have appeared in both the medical (Bailar & Mosteller, 1998) and psychology (Wilkinson, 1999) literature, which is often read and cited by researchers in the field of neuropsychology. However, while researchers have identified knowledge of statistics related to neuropsychological assessment measures (e.g., sensitivity, specificity, base rates) (LaBarge, McCaffrey, & Brown, 2003), and

* Corresponding author. Tel.: +1 610 660 1804; fax: +1 610 660 1819.
E-mail address: pschatz@sju.edu (P. Schatz).

0887-6177/$ – see front matter
doi:10.1016/j.acn.2005.06.006

even guidelines for calculating specific statistics (Zakzanis, 2001), to date there has been no systematic study of adherence among neuropsychologists to these published guidelines.

Over the past several decades of psychological research and testing, null hypothesis testing (NHT) has become a popular methodology used in empirical research (Dar, Serlin, & Omer, 1994; Leavitt, 2001), despite considerable criticism (Denis, 2003; Onwuegbuzie, 2001). The null hypothesis states that there is no effect of the independent variable or the experimental manipulation on observed results, or that any effect is the result of chance. Frequently, tests of significance are performed at the beginning of a study to establish that the two groups are equivalent at the start, thus confirming the null hypothesis (Dar et al., 1994). NHT, or inappropriate use of NHT, refers to this common practice of using statistical tests to confirm the null, that there is no difference between groups.

Another problem found in the statistical analyses is an inappropriate use of the *P*-value (Cohen, 1990; Denis, 2003) which is often used to determine whether the predetermined criterion of Type I error was exceeded (Dar et al., 1994). However, researchers frequently fail to mention a predetermined *P*-value, and instead make vague references to whether the achieved *P*-value meets or exceeds the requirements for significance (e.g., near-significant, approaching significance, not significant but representing a trend). Researchers also use *P*-values to inappropriately determine the strength of their results; by using inappropriate terminology, researchers make insignificant results seems significant and vice versa (Dar et al., 1994; Denis, 2003). Additionally, when reporting *P*-values, the values should be written as an exact value ($P = .03$), as opposed to the frequently viewed *P*-value stated as an inequality ($P < .05$).

Inappropriate uses of NHT and *P*-values result in an increased likelihood of Type I error, or when the researcher incorrectly rejects the null hypothesis. Such error usually occurs because of incorrect statistical procedures and inappropriate emphasis on *P*-values. In most cases, researchers treat each statistical test individually, instead of examining the results as a whole, demonstrating a lack of control for Type I error. Performing multiple ANOVAs or following a MANOVA with univariate ANOVAs without adjusting the alpha level accordingly commonly results in Type I error (Dar et al., 1994). The Bonferonni correction, commonly referred to in methodology and statistical texts and articles, is a simple-to-use control for inflated Type I error (e.g., Cohen, 1990; Stevens, 2002). This simple procedure involves decreasing your alpha level to account for the number of statistical analyses conducted on that independent variable, and is often referred to as a means of reducing "family-wise error." For example, the researcher analyzing the effects of an intervention on five separate dependent measures (but using a sample size too small to meet the assumptions of a MANOVA) would "correct" the *P*-value to .01 to maintain an acceptable likelihood of Type I error. Whereas five separate analyses performed with a .05 alpha level would result in a 25% likelihood of Type I error ($5 \times .05 = .25$), "correcting" the alpha level to .01 maintains a "familywise" error rate of 5% likelihood of Type I error ($5 \times .01 = .05$).

Inclusion of information regarding effect size has been identified as another approach for diminishing the problems of inappropriate use of *P*-values and increased risk of Type I error. Effect size refers to the strength or magnitude of the relationship, or the degree of departure from the null hypothesis (Rosenthal & Rosnow, 1991). The *Publication Manual of the American Psychological Association* (APA) suggests that researcher should not only provide the reader with information about statistical significance, but also with enough information

to understand the magnitude of the observed effect or relationship (APA, 2001, p. 26). Too often, researchers rely on *P*-values to demonstrate the strength of the effect of the experimental manipulation on their measures, using terms such as "borderline significant," "approaching statistical significance," "near significant," or "highly significant" (Cohen, 1965, 1990; Dar et al., 1994). In what is now an oft-referenced quote, Rosnow and Rosenthal (1989) state: "surely, God loves the .06 nearly as much as the .05" (p. 1277). Despite widespread documentation of the importance of effect sizes over *P*-values in determining the strength of effect (Cohen, 1990; Rosnow and Rosenthal, 1989), effect sizes are not routinely documented in published literature.

In 1996, the American Psychological Association established a Task Force on Statistical Inference (APA, 1996), the purpose of which was to clarify many of the issues surrounding statistical procedures used in psychological research, as well as to provide alternatives in order to increase the validity of psychological research. They identified topics relevant to this article, suggesting improvements to enhance psychological research. One of these topics focused on how to improve data usage and discourage "misrepresentation of quantitative results," emphasizing inappropriate use of NHT and *P*-values, recommending that researchers include more extensive data (such as standard deviations, sample sizes, and box-and-whisker plots), and that effect sizes and confidence intervals be included in results sections (in order to provide more information about the results than a simple *P*-value can provide). Another topic focused on keeping statistical testing to a minimal level. While they acknowledged that complex testing is sometimes necessary, they also found that keeping the analysis simple yielded more accurate results and fewer errors; using the simplest tests applicable to the data also makes the research simpler to understand when read by others.

The purpose of the current study was to evaluate the adherence to the recommendations put forth by the APA Task Force, by documenting the presence of the following four statistical errors (before, commensurate with the internet-based release, and after the American Psychological Association's Statistical Inference Report) in neuropsychology research: (1) neglect of effect size, (2) increased risk of Type I error, (3) inappropriate use of *P*-values, and (4) inappropriate use of null hypothesis testing. We hypothesized that each of the four statistical errors would occur less frequently in the *Archives of Clinical Neuropsychology* publications after the APA's Statistical Inference Report, than in publications released prior to the report.

1. Method

1.1. Materials

Empirical research papers from six volumes of *Archives of Clinical Neuropsychology* (vols. 5–7, 11–19) were reviewed. Book reviews, poster abstracts, grand rounds articles, case studies, and non-empirical or other theoretical studies were not included in the analysis, based on the inherent lack of statistical analyses. Only articles from *Archives of Clinical Neuropsychology* were reviewed, as this study was a pilot study to identify if trends existed prior to expanding the research to several journals in the field of neuropsychology. This journal in particular was chosen because the first author had promoted membership in the National Academy of

Neuropsychology to his student co-authors, and the *Archives of Clinical Neuropsychology* is the official NAN journal.

1.2. Procedures

Empirical research articles from the *Archives of Clinical Neuropsychology* were reviewed before (1990–1992; $n = 62$), commensurate with and shortly after (1996–2000; $n = 174$), and after (2001–2004; $n = 170$) the release of APA's Task Force on Statistical Inference Report. The rationale for these time period delineations was that although the Task Force met in 1996 and 1997, the results were posted on-line but were not published in the *American Psychologist* until August 1999. Thus, by the time the 1999 report was published, journal articles appearing in print in 1999 and 2000 would have been in progress and likely submitted or accepted without the benefit of the Task Force's report.

Four-hundred and six articles were reviewed for neglect of effect size, increased risk of Type I error, inappropriate use of *P*-values, and inappropriate use of null hypothesis testing. Articles were reviewed by the first author and one of the co-authors. Any discrepancies were reviewed, checked for accuracy, and discussed between the reviewers until both reviewers agreed on the decision.

Increased risk of Type I error was operationally defined as the researcher performing multiple statistical tests (i.e., 2 or more) in the absence of corrections to the *P*-value or compensatory use of multivariate analyses. Inappropriate use of *P*-values occurred where researchers documented estimated, or rounded, rather than exact *P*-values. The inappropriate use of null hypothesis testing occurred where statistical analyses were used to establish baseline, between-groups equivalence. Decreased likelihood of making these errors was tested using 2×3 chi-square analyses: year group (1990–1992, 1996–2000, 2001–2004) by error occurrence (yes, no). Bonferonni correction for familywise error rate set the alpha level to .0125.

2. Results

2.1. Neglect of effect size

Of the 12 volumes of *Archives of Clinical Neuropsychology* evaluated, approximately three-quarters of all researchers neglected to mention or address effect size, as only 93 of the 406 articles (22.9%) documented effect size. The release of the Statistical Inference Report appears to have a significant positive effect on likelihood of documentation of effect size [$\chi^2(2, N = 405) = 13.6$, $p = .001$; $\phi = .18$], with 4 of 62 articles (4.3%) prior to the release of the Task Force report, 39 of 174 (22.4%) commensurate with the release of the report, and 50 of 170 articles (29.4%) after the release of the report including this statistic.

2.2. Increased chance of Type I error

Increased risk of Type I error was found in 275 of 406 articles (67.7%). Analysis revealed no significant effect of the Statistical Inference Report on decreased likelihood of risk of Type

I error [$\chi^2(1, N = 406) = .39, p = .87; \phi = .03$], with 40 of 62 articles (64.5%) prior to the release of the Task Force report, 118 of 174 (67.8%) commensurate with the release of the report, and 117 of 170 articles (68.8%) after the release of the report revealing increased risk of Type I error.

2.3. Inappropriate use of P-values

Inappropriate use of *P*-values was found in 293 of 406 articles (72.2%). Analysis revealed no effect of the Statistical Inference Report on decreased likelihood of improper use of *P*-values [$\chi^2(2, N = 406) = 3.14, p = .21; \phi = .02$], with 48 of 62 articles (77.4%) prior to the release of the Task Force report, 130 of 174 (74.7%) commensurate with the release of the report, and 115 of 170 articles (67.6%) after the release of the report revealing inappropriate use of *P*-values.

2.4. Inappropriate use of null hypothesis testing

Inappropriate use of null hypothesis testing occurred in 194 of 405 (47.8%) articles. Analysis revealed no effect of the Task Force report on decreased likelihood of Inappropriate use of null hypothesis testing [$\chi^2(2, N = 406) = 6.62, p = .037; \phi = .18$]; 21 of 62 articles (33.9%) prior to the release of the Task Force report, 92 of 174 (52.9%) commensurate with the release of the report, and 81 of 170 articles (47.8%) after the release of the report revealing inappropriate NHT.

3. Discussion

Our results indicate that despite the documentation in the *Publication Manual of the American Psychological Association* and the recommendations of the 1999 Task Force report, there continue to exist errors when reporting statistical data in neuropsychological research. While such inaccuracies were anticipated in the time period prior to the report's release, a significant improvement was expected in the years following the report. There was a significant increase in the reporting of effect size, and this appears to have become more commonplace throughout the literature. However, there were no other significant effects (with small effect sizes noted), and none of the three other criteria appear to have improved as a result of the Statistical Inference Report. These results are not to imply that neuropsychology researchers are in some way "missing the boat" with respect to statistical analyses and documentation. Rather, as pointed out by Dar et al. (1994), null hypothesis tests have long since been misinterpreted, significance levels have been over-inflated, and Type I error rates have been over inflated in a vast majority of studies.

Such statistical inaccuracies may hinder the progress that can be made by neuropsychological research, so adherence to the recommendations of the Task Force is not simply trivial or without purpose. Over-reliance on null hypothesis testing, especially combined with inflation of Type I error rates, can create invalid between-groups differences, and these differences are often controlled for using the variable as a covariate. Such errors may confound the validity of neuropsychology research, by decreasing the generalizability of results, or at times putting forth results which are not actually substantiated by the data (see Miller & Chapman,

2001). Including information regarding effect size may add validity to many research studies, especially when there is a smaller sample size and effect sizes point to a much stronger outcome than *P*-values alone. Often, when employing a smaller sample, the effect size is a more accurate measure of the result of the experimental manipulation (Cohen, 1992), which should sit well with both graduate students and researchers who have traditionally relied on the *P*-value as the lone determinant of the merit or significance of a study. As well, in the absence of effect size data, documentation of exact *P*-values facilitates post-hoc calculation of effect sizes (see Rosenthal and Rosnow, 1991) as required for meta analyses. We recognize that these recommendations may be battling a long history of trends in psychological research and publications. Researchers have long-since rounded *P*-values to general categories (such as <.05 or <.01), but providing more detailed information will allow future researchers to extract more definitive meaning from retrospective research and literature reviews. In those situations where a future researcher is attempting to identify collective effect sizes across similar articles, specific *P*-values will allow for more precise retrospective identification of effect sizes that cannot be achieved with rounded *P*-values.

In light of these findings, we offer the following recommendations:

1. Researchers and journal reviewers should read the hallmark literature in this area (Rosnow & Rosenthal, 1989), familiarize themselves with APA's recommendations (APA, 1996, 2001), and consult recommended textbooks where appropriate (Rosenthal and Rosnow, 1991; Stevens, 2002).
2. Establish an a priori alpha level in order to minimize Type I and family-wise error.
3. When appropriate, opt for multivariate analyses over multiple univariate analyses.
4. Document pre-established alpha levels, as well as group means, standard deviations, exact *P*-values, and effect sizes.

In considering the limitations of this study, it is possible that authors of articles published from 1996 to 2000 may not have been exposed to the internet-based version of the Task Force's report. As well, authors of articles published from 2001 to 2004 may also not have been aware of the APA's report, thus decreasing the opportunity for the report to have an effect. However, given the sparse adherence to APA's recommendations it appears this research may provide necessary exposure within the field of neuropsychology. It is not clear why researchers have either not been made familiar with the recommendations, or failed to adhere to the Task Force's recommendations. It is possible that while graduate Ph.D. programs do emphasize research methodology and statistics, there are few, if any, continuing education programs at neuropsychology conventions focusing on these topics. Another explanation may be that journal editors and reviewers have incorporated these recommendations in the guidelines traditionally provided to authors wishing to submit articles for publication. A future survey of journal editors and their perceptions of these issues and their importance would perhaps provide further illumination.

To date, the authors have looked at only one journal, and it is possible that other journals have different rates of adherence to APA's recommendations. A future study will compare the *Archives of Clinical Neuropsychology* with other journals in the field of neuropsychology. If only generalizable to the current consumers and producers of research in *Archives of Clinical*

Neuropsychology, we feel this is an accurate "snapshot," and the above-listed recommendations should have a significant effect on improving the adherence to APA's guidelines.

References

American Psychological Association. (1996). *Task Force on Statistical Inference Report*. Washington, DC: American Psychological Association.

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

Bailar, J. C., III, & Mosteller, F. (1998). Guidelines for statistical reporting in articles for medical journals: Amplifications and explanations. *Annals of Internal Medicine*, *108*, 266–273.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). McGraw Hill: New York.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*, 98–101.

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology*, *62*(1), 75–82.

Denis, D. J. (2003). Alternatives to null hypothesis significance testing. *Theory & Science, 4*(1). [online]. Retrieved March 1, 2004 from http://theoryandscience.icaap.org/content/vol4.1/02_denis.html.

Labarge, A. S., McCaffrey, R. J., & Brown, T. A. (2003). Neuropsychologists' abilities to determine the predictive value of diagnostic tests. *Archives of Clinical Neuropsychology*, *18*, 165–175.

Leavitt, F. (2001). *Evaluating scientific research: Separating fact from fiction* (pp. 210–226). Upper Saddle River, NJ: Prentice Hall.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*(1), 40–48.

Onwuegbuzie, A. J. (2001). Common methodological, analytical, and interpretational errors in published educational studies: An analysis of the 1998 volume of the British Journal of Educational Psychology. *Educational Research Quarterly*, *26*(1), 11–22.

Rosenthal, R. L., & Rosnow, R. (1991). *Essentials of behavioral research: Methods and data analysis* (pp. 491–512). Boston, MA: McGraw Hill Inc.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Lawrence Earlbaum Associates.

Wilkinson, L. (1999). Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Zakzanis, K. K. (2001). Statistics to tell the truth, the whole truth, and nothing but the truth: Formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Archives of Clinical Neuropsychology*, *16*, 653–667.