2016

Reporting Practices and Use of Quantitative Methods in Canadian Journal Articles in Psychology

Alyssa Counsell

Lisa L. Harlow
University of Rhode Island, lharlow@uri.edu

Citation/Publisher Attribution
Counsell, A., & Harlow, L. L. (2016, October 6). Reporting practices and use of quantitative methods in Canadian journal articles in psychology. *Canadian Psychology/Psychologie Canadienne*. Advance online publication.
Available at: http://dx.doi.org/10.1037/cap0000074

Running head: STATISTICS IN RECENT CANADIAN JOURNALS

The citation is:

Counsell, A., & Harlow, L. L. (2016, October 6). Reporting practices and use of quantitative methods in Canadian journal articles in psychology. *Canadian Psychology/Psychologie Canadienne*. Advance online publication. http://dx.doi.org/10.1037/cap0000074

Reporting Practices and Use of Quantitative Methods in Canadian Journal Articles in Psychology

Alyssa Counsell[1*] and Lisa. L. Harlow[2]

[1]York University, Toronto, ON, Canada

[2]University of Rhode Island, Kingston, RI, USA

Thanks are extended to the National Institutes of Health grant G20RR030883 for L. L. Harlow

Correspondence concerning this article should be addressed to Alyssa Counsell, Department of Psychology, York University, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3.
E-mail: counsell@yorku.ca

Abstract

With recent focus on the state of research in psychology, it is important to assess the nature of the methods and analyses used and reported. To study this, we coded information about the statistical content reported in articles in the four major Canadian psychology journals published in 2013. We first classified whether the articles were quantitative, qualitative, or theoretical in nature. Our main focus was on articles that used quantitative methods; whereby we investigated the prevalence of different statistical procedures, as well as further details of reporting practices. Few articles in any of the journals used qualitative approaches, as 92.9% of empirical articles included a quantitative study. Analysis of variance (ANOVA), *t*-tests, and multiple regression were the statistical analyses most often reported in the investigated articles. The majority of articles used hypothesis testing, and while most of these tests were accompanied by an effect size, this rarely included a confidence interval. Many of the quantitative studies provided minimal details about their statistical analyses and less than a third of the studies presented on data complications such as missing data and statistical assumptions. Further discussion highlights strengths and areas for improvement for reporting quantitative results. The paper concludes with recommendations for how researchers and reviewers can improve comprehension in statistical reporting.

Key words: Canadian psychology, quantitative methods, statistics, review, reporting practices

Reporting Practices and Use of Quantitative Methods in Canadian Journal Articles in Psychology

Psychology is a diverse discipline that continues to grow. Given its size and diverse nature, some have argued that the field of psychology has become fragmented (e.g., Goertzen, 2008), and many are questioning the nature of our studies and the strength of the findings (e.g., Open Science Collaboration, 2015). Attempting to define common features amongst the branches of psychology and the practice of research can present a challenge. However, one of the commonalities for much of the research in psychology is the use of common research methods and analyses, particularly for quantitative studies. The current paper examines the types of articles published in in recent Canadian journals in psychology to look for commonalities in methodological approaches for answering research questions. This will provide the reader with information about the general nature of the research (e.g., the prevalence of theory/review, qualitative, or quantitative designs). Subsequently, we will investigate the specific information presented in the results sections of empirical articles that use quantitative methods, since this is the most common analytic strategy for research in psychology. Here, we will further discuss commonalities of reporting practices and compare this information to recommended practice.

The current study focuses predominantly on the use of quantitative methods in psychology although the authors note that there is a demand for more integration with qualitative research methods (Gergen, Josselson, & Freeman, 2015). Thus, we include general information about the prevalence of qualitative data methods in comparison to the prevalence of quantitative methods in recent substantive articles in Canadian psychology journals. For those with more interest in the topic of qualitative research, we refer you to a special issue of *Canadian Psychology* devoted to qualitative methods (e.g., see O'Neill, 2002; Rennie, Watson, & Monteiro, 2002).

Whereas quantitative methods are widely used in psychology, it is not done without controversy and debate. For example, a number of articles and books are devoted to discussing reporting practices and the use of null hypothesis significance testing (NHST: e.g., Chow, 1996; Cumming, 2012; Harlow, Mulaik, & Steiger, 2016; Kline, 2013). The NHST debate led to the publication of recommended guidelines from the American Psychological Association (APA) through the creation of a task force on statistical inference (Wilkinson & the APA Task Force on Statistical Inference 1999). Such guidelines around statistical reporting are described in the current edition of the publication manual (APA, 2010). Given the wide number of individuals arguing against NHST or at very least for better supplementing of NHST information, we believe that the following paper contributes by providing information about the extent to which empirical articles in Canadian journals have incorporated these issues into their reporting practices.

Reporting Quantitative Information in the Results Section

Data complications. No researcher collects perfect data without complications. Issues such as missing data (due to attrition, nonresponses, etc.) or violations of statistical assumptions (e.g., nonnormally distributed variables) should be considered when reporting the results of statistical tests. More details on these complications are described in turn.

Missing Data. A wealth of literature on missing data strategies exists (Allison, 2002; Baraldi, & Enders, 2013; Enders, 2010; Little & Rubin, 2002). Several strategies for dealing with missing data have been suggested in the past (e.g., listwise deletion, pairwise deletion, mean substitution, etc.). However, these methods have been called into question and newer methods have been advocated, including multiple imputation and full information maximum likelihood (Graham, 2009). Multiple imputation is a multistage process whereby the missing data points are

replaced by a score predicted from a regression line calculated by including other relevant variables. This procedure is repeated *m* times so that a researcher will have *m* datasets. The analysis of interest is then conducted on each of the *m* datasets, and the *m* analyses are pooled so that the researcher has one set of estimates, test statistics, and standard errors. In full information maximum likelihood, one does not replace missing data points, but instead produces model estimates using all of the available information. With this method, one must run a model that is able to utilize a maximum likelihood method for estimation (e.g., multilevel models, structural equation models, etc.)

Unfortunately, the most common methods used by applied researchers tend to be those that rely on software defaults (e.g., listwise deletion in SPSS) rather than recommendations by methodologists (Bodner, 2006; Wood, White, & Thompson, 2004). The APA task force on statistical inference stated that excluding cases with missing data is "among the worst methods available for practical applications" (Wilkinson et al., 1999, p. 598). Aside from issues with using simple methods for dealing with missing data, both Kline (2013) and Reinhart (2015) discuss how many articles do not explicitly state how the research dealt with missing data problems at all, which makes it hard for the reader to know how to validate the author's decisions and results.

Statistical Assumptions. Parametric statistical tests must satisfy a number of statistical assumptions in order for valid interpretation of the results. Unfortunately few articles include any information about statistical assumptions and the information included may not be comprehensive (Kline, 2013). A lack of information about statistical assumptions could stem from assumptions not being tested, mistaken information about the robustness of statistical tests, (e.g., Bradley, 1978; Glass, Peckham, & Sanders, 1972), being unaware of the importance of

attending to statistical assumptions, or that the data has met the statistical assumptions but the researchers have simply not reported it. The decision for a researcher to use a parametric test should depend on which of these scenarios occur; but this is virtually impossible for readers to validate with the current amount of information reported by applied researchers. Choosing a parametric test and failing to report information about assumptions because they have been met has different implications than a scenario in which a researcher chose a parametric test but statistical assumptions were violated. Unfortunately, research by Hoekstra, Kiers, and Johnson (2012) suggests that researchers rarely test their assumptions, and this appears to stem from having limited knowledge about the robustness of parametric tests and how and why they should test statistical assumptions.

Taken together all of this information is important information to include in a results section, but as Kline (2013) notes, we appear to have a reporting crisis in psychology.

Significance tests. The majority of empirical articles in psychology use NHST (Rodgers, 2010) despite considerable opposition to an exclusive focus on dichotomous significance tests (e.g., Cohen, 1994; Cumming, 2012; Kline, 2013; Rozeboom, 1997; Schmidt & Hunter, 1997; Wilkinson et al., 1999). Amidst these opposing perspectives, a number of researchers endorse the use of significance tests in some circumstances, particularly if accompanied by relevant effect sizes and confidence intervals (CIs) (e.g., Abelson, 1997; Denis, 2003; Hagen, 1997; Harlow, 2010; Harris, 1997; Mulaik, Raju, & Harshman, 1997). Detailed information on reporting the results from inferential significance tests is presented in the APA publication manual:

> "For inferential statistical tests (e.g., t, f and chi square tests) include the obtained magnitude or value of the test statistic, the degrees of freedom, the probability of obtaining a value as extreme or more extreme than the one obtained (the exact p value),

> and the size and direction of the effect. When point estimates (e.g., sample means or regression coefficients) are provided, always include an associated measure of variability (precision), with an indication of the specific measure used (e.g., the standard error)" (APA, 2010, p. 33).

The publication manual recommends reporting results from hypothesis tests, although it includes other recommendations about including information about measures of variability and CIs. Given that most journals follow the APA publication manual for reporting practices, it is unsurprising that applied researchers continue to rely on reporting NHST results.

Effect sizes and confidence intervals. Several individuals opposed to NHST consider effect sizes the viable alternative. Cumming (2008; 2012) and Thompson (2007) have strongly argued against using NHST and advocate reporting effect sizes and their associated CIs without any significance tests. The reasoning is that effect sizes provide information about the magnitude or importance of an effect, which is really what researchers want, rather than whether a nil hypothesis has been rejected. Effect sizes are particularly useful when accompanied by information about variability around the effect (e.g., CI); and if the CI contains the null values (e.g., a mean difference of 0), a hypothesis test would not be statistically significant. The APA publication manual simply recommends the use of effect sizes for the "reader to appreciate the magnitude or importance of a study's findings," stating further that, "the inclusion of confidence intervals (for estimates of parameters, for functions of parameters such as differences in means, and for effect sizes) can be an extremely effective way of reporting results" (APA, 2010, p. 33). Some may argue that these recommendations are not strong enough, and Wilkinson and the Task Force explicitly stated that effect sizes should always be presented, while a measure of variability such as a CI should be included on any effect size reported. It is becoming more

common for journals to require effect sizes but CIs are not typically required, so the extent to which individuals are reporting effect sizes and CIs together is not clear.

Visual representations of data. Figures allow researchers to present a large amount of data in an efficient manner so that readers may examine the results in a more comprehensive manner than solely providing results of significance tests. The APA publication manual devotes a large section to tables and figures in order to discuss the many purposes for visual displays of data such as exploration, communication, calculation, storage or decoration. A number of books and articles on graphical expressions of data can also be found (e.g., Cleveland, 1993; Friendly, 2000; Friendly & Meyer, in press, Margolis & Pauwels, 2011, Tukey, 1977). Unfortunately, informative figures and graphics are not always presented, despite the recommendations that researchers include high quality figures with indicators of variability (e.g., Wilkinson et al., 1999).

Improving Psychological Science through Reporting Practices

All of the issues described thus far contribute to the larger problem in psychology of lack of transparency and issues with replication. Articles and special issues of journals are devoted to this topic (e.g., *Perspectives on Psychological Science*), along with a number of general psychology discussions and theoretical articles discussing research practice (e.g., Anderson & Maxwell, 2016; Funder et al., 2014; Kline, 2013; Nosek, Spies, & Motyl, 2012; Wilkinson et al., 1999). While we believe that these articles are of paramount importance, it remains worthwhile to examine the impact of such papers on articles in practice. We believe that it is important not only to discuss areas for improvement when applied researchers do not follow some of these recommended guidelines, but also discuss what researchers are doing well. As such, the

following paper contributes to the literature by providing concrete information about what and how Canadian journal articles in psychology are reporting.

Purpose of the Current Study

The purpose of the current paper is to investigate the nature of the methods and analyses used and how researchers report results from their analyses. We seek to classify the number of theoretical, and empirical papers (qualitative or quantitative) in psychology articles and further explore the statistical procedures used in recent psychology articles in Canadian journals. Specifically, we aim to examine the nature of the methods used (e.g., quantitative, qualitative) and elaborate on what specific types of information researchers report for their analyses, with the main focus on quantitative analyses. The paper will discuss both strengths and limitations of the current reporting practices of the articles in Canadian journals and conclude with recommendations for reporting quantitative results. We have four specific research aims but the main focus will be on the second and third goals:

1. Classify the methodological focus in the articles examined (i.e., whether they used quantitative or qualitative methods, or presented a theory or review);
2. Examine the frequency of specific statistical procedures (e.g., correlation, *t*-test, etc.);
3. Examine what types of information are reported in conjunction with a quantitative analysis (e.g., figures, effect sizes, confidence intervals, etc.); and
4. Assess whether articles were using statistical procedures appropriately for their research design.

Method

The current study examined all issues of the four major Canadian psychology journals published in 2013. The journals included: *Canadian Psychology* (*CP*), the *Canadian Journal of*

Experimental Psychology (*CJEP*), the *Canadian Journal of Behavioural Science* (*CJBS*), and the *Canadian Journal of School Psychology* (*CJSP*). The first three journals are both Canadian Psychological Association (CPA) and American Psychological Association (APA) journals. After excluding articles that were not appropriate for the study (e.g., editorials, book reviews, and commentaries), our first task was to classify the articles based on whether they were a theory or review paper in psychology or an empirical article with qualitative or quantitative analyses. For articles that included quantitative analyses, we examined specific types of quantitative information, based on our second and third research goals, and sought to assess the appropriateness of the quantitative methods used and reported for our fourth goal. As an example for assessing appropriateness, if a researcher had data with a dependency structure (e.g., individuals nested within couples and both were included in the study), it would be inappropriate to use a traditional linear regression model instead of one that takes the dependency into account (e.g., a multilevel model).

To collect information regarding whether specific types of information were included from a quantitative analysis, the first author created a spread sheet where each column represented a dummy variable about whether that type of information was included or not (e.g., included effect size?). Each row represented a unique analysis within a particular empirical study. The second author reviewed the coding and made suggestions when needed.

Results

Prevalence of Journal Article Type

To address the first goal, we examined the nature of the articles in all of the issues published during 2013 of the four Canadian journals. In total there were 126 articles, however, 25 were excluded from the study as they included editorials, introductions to a special issue,

book or test reviews, and commentaries (on other papers or conference activities). This left 101 relevant articles for the study which included 27 from *CJEP*, 32 from *CJBS*, 25 from *CP*, and 17 from *CJSP*. Of the 101 articles, 65 included quantitative methods, 5 included qualitative methods, and 31 were review or theory papers. For articles that included an empirical study, 92.9% included at least one quantitative analysis, whereas 7.1% included a qualitative analysis. As expected, the majority of the review and theory papers came from *CP* (71%). Most of the articles were written in English ($N = 89$, 88%), 11 were written in French (11%) and one article (1%) was included twice, both in English and in French.

Prevalence of Statistical Analyses and Inferential Procedures

Although there were 65 studies with quantitative information, most articles had several unique statistical analyses, such that there was a total of 153 analyses included in the current study. An analysis was considered unique if it was used to answer a question of substantive interest, was not used as a manipulation check or to equate groups based on demographic information, and was not used to supplement another analysis (e.g., presenting a correlation matrix when the main analysis is a multiple regression model). Table 1 addresses our second goal by presenting a breakdown of the types of statistical analyses of which the 153 procedures were comprised.

INSERT TABLE 1 HERE

From Table 1, one can see that the most popular methods were ANOVA and *z* or *t*-tests. In fact, 41.2% of the analyses included a univariate mean comparison. Tests of univariate mean comparisons were highly representative of the articles in the *CJEP*. Analyses that examined

associations amongst variables (multiple regression, correlation, and chi square) were also frequently used (35% of the analyses used one of these three techniques). Multivariate and modeling techniques tended to be used less frequently and with a wide range of techniques employed (e.g., structural equation modeling, logistic regression, mixed effects models, and generalized linear models). Four of the analyses (2.6%) included only descriptive statistics (means, odds ratios, etc.) to answer their research question.

Types of Quantitative Information Reported

Whereas the prevalence of statistical methods is informative, investigating the types of statistical inference information presented from such analyses will provide information about reporting practices and areas for improvement. This was the focus of our third goal, and this information is presented in Table 2.

INSERT TABLE 2 HERE

Significance tests and effect sizes. Almost all of the articles presented significance tests with their analysis (88.9%), and these included inconsistencies with their reporting of *p* values. Dichotomous and exact *p* values were reported with almost equal frequency across the articles surveyed, although in many instances a researcher would report both dichotomous and exact *p* values within the same analysis. For example, when reporting the results of a multiple regression analysis, a researcher may have presented the exact *p* value for the overall model's significance test (e.g., $p = .023$), but then reported $p < .05$ from a predictor variable's hypothesis test. For this reason the percentage of significance tests that included dichotomous or exact *p* values does not

sum to 100% in Table 2. Few analyses (26.4%) reported the standard error associated with their test statistic and *p* value.

A large portion of articles presented an effect size. In fact, of the 136 analyses with a significance test, 86% included an effect size. Ten analyses included an effect size without a hypothesis test, and of these ten, only one included a CI on the effect. In fact, CIs were rarely employed regardless of whether a significance test was used since only 16 (10.1%) of all analyses included them.

One reason that the number is high is that a broad definition of effect size was used for the purposes of recording whether one was presented or not. Unstandardized and standardized effect sizes were both considered such that we defined 'effect size' as any measure of the magnitude of an effect. This could range from raw means to regression weights, to an adjusted R^2. We did not draw a distinction between standardized and unstandardized because we believe that researchers should present effect size information that they think conveys the strength or importance of their effect in the units that are most meaningful to the reader. This idea will be elaborated further in the discussion section.

Figures. Forty-six (31.4%) analyses included a visual representation of the data alongside their statistics. Of these 46, only 27 (56%) of them included an indication of variability such as error bars on the CI or standard error. In general the plots tended to be simple bar charts presenting a small number of group means.

Missing data and statistical assumptions. Only 30.1% of the analyses included explicit information about how much missing data was present and how the researcher dealt with this issue. That being said, examining the degrees of freedom from the analyses often allowed us to determine whether there was any missing data and if so, whether pairwise or listwise deletion

was used. In fact, it appeared as though many of the articles in the *CJEP* had complete cases. If information was presented about missing data, few articles reported a missing data strategy other than a simple deletion technique (listwise used more frequently) or mean substitution.

The number of analyses that included information about statistical assumptions was identical to those reporting on missing data (i.e., 30.1%). Despite only about a third of analyses including any information about statistical assumptions, only two analyses included information about whether *all* of their statistical assumptions were met. The other 44 included limited information and typically only addressed one of their statistical assumptions (e.g., were the data normally distributed when conducting an ANOVA?). In some cases, they attempted to address a statistical assumption, but did so incorrectly. One example of this is where a researcher failed to examine the normality of the regression residuals and instead examined the distribution shapes of the independent and dependent variables.

Appropriateness Ratings of Statistical Procedures

To address our fourth goal, we initially sought to provide information about whether authors implemented the most appropriate statistical analysis based on their research design or methodology and sample. The challenge was that few articles presented enough information in the results section to adequately assess whether their statistical choice was appropriate. The lack of transparency around statistical assumptions was one of the biggest issues for assessing whether articles are using appropriate methods. Almost all of the researchers chose statistical tools that adequately complemented their research design, but without information on statistical assumptions, it is impossible to provide reliable validation for an author's choice of statistical test. Thus, we decided against presenting data here on evaluating the adequacy of the research given that there was not enough information provided in the articles to allow for a reliable

assessment. Readers interested in seeing the informal evaluations we did make can request this information from the first author.

Discussion

The current research examined the methodological trends of psychology articles in the four major Canadian journals in psychology during 2013. We sought to provide information about the different types of articles such as the prevalence of theory and review papers in comparison to empirical papers. For the articles that included an empirical study we further classified them as using qualitative or quantitative methods (none used an integrative approach). The main focus was on the quantitative empirical studies, whereby we investigated in detail, the statistical methods used in the articles and information researchers reported alongside their analyses. Getting a view of the landscape in these articles can offer greater awareness of what analyses are currently being used in the Canadian literature and how this information is reported. Investigating this type of information may offer insight into what researchers are doing well and what could be done to improve the nature of inference in future studies.

Classification by Article Type

We found that about 70% of the articles included an empirical study, and of those that included an empirical study, authors overwhelming used quantitative approaches. Few articles included any qualitative data or information. While Gergen et al. (2015) argue that qualitative information allows for a more pluralistic or holistic view of individuals, we believe that both quantitative and qualitative methods have their own merits. Including a mixed methods approach that includes both quantitative and qualitative methods may provide a richer account of a particular phenomenon. Mixed or integrative methods were not observed in any of the articles surveyed.

A large number of articles were devoted to theory and review (30%). Most of these were within the *Canadian Psychology* journal. Theory and review papers are important for bridging together different areas of research within a subfield (e.g., developmental research in autism) or within the larger field (e.g., generalist articles). These papers provide a meta-view of the field to allow for more oversight into areas needing development or improvement. Review papers in Canadian journals also provides information that benefits Canadian researchers since some generalist papers about the state of research in institutions (such as the education system or hospitals) conducted in other countries may not be relevant or as applicable. Theory and review papers may also provide a new perspective on a topic or allow for concise information about a large number of studies in one paper.

Quantitative Methods Used in Empirical Articles

Of the articles under investigation, the types of statistical methods used were largely univariate in nature. In fact, the majority of the procedures included methods taught at an undergraduate level (i.e., *t*-tests, ANOVA, chi square, correlation and multiple regression). However, they also represented popularities in certain fields. For example, experimental articles (mostly in the *CJEP*) overwhelming used ANOVA. Observational studies typically included simple correlation matrices or multiple regression analyses. These findings contrast with those from Harlow, Korendijk, Hamaker, Hox, and Duerr (2013) who examined the extent of multivariate methods and statistical inference procedures used in eight European psychology journals. Their study found that 57% of the articles used multivariate methods. Whereas parsimony is important and few articles in the current study used a more complicated model when a simpler one would suffice, many articles would have benefited from incorporating their research hypotheses into a larger multivariate model. In general, researchers were much more

likely to conduct several univariate models than to include a multivariate. For example, instead of running several multiple regression models, a researcher could have used a path analysis model or structural equation model. Instances of running several ANOVA models instead of one MANOVA will reduce the number of significance tests thereby reducing multiplicity. Multivariate models also allow for a focus on a more cohesive, integrated understanding of the nature of the data with respect to the research questions asked. The challenge is that multivariate models tend to require larger sample sizes and the median sample size in the empirical articles surveyed was only 89 (but ranged from $N = 5$ to 44, 560).

Hypothesis Testing and Effect Sizes

Despite calls for reducing reliance on NHST, the majority of the articles surveyed used significance tests. However, the constant calls for reporting effect sizes appears to have had an effect on the Canadian psychology articles as almost 90% of the analyses that used a significance test included some sort of effect size. Few articles presented an effect size without hypothesis testing, and a small number of the analyses' results included a confidence interval. In fact, CIs typically were not reported as a supplement for NHST, but they were not presented when a researcher included an effect size with no significance tests either.

In coding whether an analysis included an effect size or not, we adopted a broad framework for effect sizes such that both unstandardized and standardized measures were included. We adopted this approach because the goal of presenting effect sizes is to provide readers with some measure of magnitude for an effect, such that readers can see the practical significance of the findings. This can be achieved in a number of different ways. As stated by the Task Force on Statistical Inference, "if the units of measurement are meaningful on a practical level (e.g., number of cigarettes smoked per day), then we usually prefer an unstandardized

measure (regression coefficient or mean difference) to a standardized measure (r or d)" (Wilkinson et al., 1999 p. 599). However, they go on to describe how it is important to include comments that place these effect sizes within a relevant theoretical context. We noticed that this is an area requiring improvement. Researchers are becoming more aware of the importance of presenting effect sizes, but they are not discussing them further or situating them within the larger body of literature.

Visual Displays of Data

Data visualization can be an incredibly useful tool for presenting statistical information. This study demonstrated that high quality informative graphics were not being utilized in the majority of these applied psychological research articles. Specifically, less a third of the analyses included a graphical representation of the data and only half of these included a measure of variability such as an error bar on the CI or standard error. For the figures that were included, many of them were unnecessary, presenting simple bar charts plotting means from *t*-tests or one-way ANOVAs. Given that in this particular investigation of psychology articles, the majority of analyses used univariate mean comparisons or simple correlation, presenting complex figures may not be as crucial. However, better visualization methods should be used. For example, it may be helpful to include boxplots instead of bar charts, since boxplots include information about distribution shape, central tendency, variability, and outliers.

General Transparency and Detailing Important Statistical Information

In general, the articles included a great amount of detail in the methods section, which allows other researchers to attempt replication, but the information provided in the results section was often minimal. The APA publication manual states:

> "In the results section, summarize the collected data and the analysis performed on those data relevant to the discourse that is to follow**.** Report the data in sufficient detail to justify your conclusions. Mention all relevant results, including those that are counter to expectation; be sure to include small effect sizes (or statistically nonsignificant findings) when theory predicts large (or statistically significant) ones. Do not hide uncomfortable results by omission." (APA, 2010, p. 32).

This statement makes it clear that researchers should aim to be as transparent as possible with their results. The majority of articles in the study included the information required by the publication manual such as test statistic, *df*, *p* value, effect size, but few articles presented their data and analyses in sufficient detail so that a reader could justify the authors' conclusions. For example, it was common for researchers to say that they conducted an ANOVA or *F*-test, without specifying which type. This term could refer to between subjects, within subjects, mixed effects, factorial, etc. If not explicitly stated, we used the model degrees of freedom and information from the design in the methods section to identify what type of ANOVA was used. Having to identify the type of ANOVA by degrees of freedom is particularly problematic when the *df* may have been adjusted due to missing data or robust alternatives (e.g., Greenhouse-Geisser epsilon), or may involve a typographical error. Readers should not have to go to such troubles just to figure out what type of analysis was used.

A related issue was the lack of information about how missing data and statistical assumptions were addressed. This poses a real problem for validating statistical decisions and was the biggest issue in trying to assess the appropriateness of a researcher's statistical test. As was discussed in the introduction, there are a number of potential reasons for why a researcher does not report on their statistical assumptions. If researchers are not testing for their statistical

assumptions, their choice of statistical test is likely problematic since research has suggested statistical assumptions in psychological research are frequently violated (Blanca, Arnau, Lopez-Montiel, Bono, & Bendayan, 2011; Keselman et al., 1998, Micceri, 1989). Using traditional parametric statistical tests with assumption violation has implications such as higher Type I or Type II error rates depending on the nature of the violations (Coombs, Algina, & Oltman, 1996; Cribbie, Fiksenbaum, Wilcox, & Keselman, 2012; Glass et al., 1972; Lix, Keselman, & Keselman, 1996). As researchers' conclusions, implications, and suggestions for future directions are all based on the results assuming valid parametric procedures, when statistical assumptions are violated but not addressed, one runs the risk of presenting useless, misleading, or potentially harmful results.

Why is all of this important information missing? In the majority of cases, we do not think that researchers are hiding data issues, but instead that applied researchers, reviewers, and editors do not immediately realize the importance of such information for critical evaluation of the work. With limited space in journals, some of the more gritty details of statistical tests may be lost; researchers focus their page space on the discussion and conclusions — what their results actually mean and why others should care. Further, researchers report what is required by the journal. If editors and reviewers do not require certain types of statistical information, it is unlikely to be reported since that page space can go to reporting other types of information that the researcher thinks is important. Another issue is that the importance of quantitative skills and training tends not to be emphasized enough in both undergraduate and graduate level training. In fact, most graduate students in North America are only required to take about one year of statistics courses by the end of their doctoral degree (Aiken, West, & Millsap, 2008). Given this limited training, many applied researchers rely on the information presented from software. For

example, if software does not report a CI on Cohen's *d*, it is unlikely that a researcher will calculate one his or herself. Recommendations by the publication manual and by journals may not be strong enough. Instead researchers should be required to discuss these types of information in their articles so that others have the necessary material to validate researchers' quantitative decisions.

Reporting Practices for a Better Science

The current study contributes to the field by providing an updated snapshot of the state of Canadian psychology articles with regards to statistical methods and inferential procedures used. It is important to monitor and report on practices and trends of a discipline to capitalize on strengths and address limitations. Some of the issues that arose in this study belong to a larger group of issues that need to be addressed. Transparency in reporting and research practices (e.g., Nosek, et al., 2012), replication (e.g., Anderson & Maxwell, 2016; Open Science Collaboration, 2015), and using valid and reliable methods and instruments are a few issues. Along with the APA publication manual and Task Force guidelines and recommendations, other researchers have published recommendations. For example, Funder and colleagues (2014) have outlined a number of recommendations put forth by the Society of Personality and Social Psychology Task force on Publication and Research Practices. Nuijten and colleagues (2015) examined articles from eight major journals from 1985 to 2013 and found a large percentage of reporting errors and include some recommendations for researchers in an effort to improve the dependability of psychological research. Cousineau (2014) discusses the importance and need for replication studies so that the field can work towards building a body of scientific knowledge rather than simply publishing articles with an acceptable *p* value. Recommendations are helpful but in order for the reporting practices of researchers, journals must insist on reporting certain information.

As Funder et al. (2014) note, "to make our field more amenable to these [recommended] practices, it is important for all of us, including editors, reviewers, and those who make hiring/promotion decisions, to educate ourselves about their value" (p. 9).

Recommendations and Conclusion

Reporting practices have come a long way in psychology. The changes from each edition of the APA publication manual highlight this progress, and journals now require more information from authors. That being said, we believe that it is not enough. Here we include a list of recommendations driven by the current study's results that will benefit both applied researchers and the reviewers and editors of journals. They are as follows:

1. Think about conceptualizing a larger model or using a multivariate method instead of running several univariate analyses. Larger models are not always necessary or feasible, but at the very least, researchers should consider whether their hypotheses can be answered by one model instead of running multiple smaller analyses.

2. Be explicit about which statistical test you have conducted. This can be as simple as stating that you conducted a paired samples *t*-test as opposed to reporting conducting simply a "*t*-test" or "mean comparison." Specifying the type of ANOVA (e.g., between groups factorial) or type of regression analysis (linear multiple regression) would improve the reader's comprehension and higher potential for reproducing results.

3. Report on the amount of missing data present and how you dealt with it in your analyses. In cases with a lot of missing data, or where missing data are not missing completely at random, consider strategies other than deletion methods or mean substitution (see Graham, 2009) to prevent biasing results.

4. Present information about whether statistical assumptions were met. If they were not, how were the violations addressed and were there other anticipated data complications?

5. Present the data graphically if visualization allows for readers to better see trends and patterns, but do not include graphics that are redundant or unhelpful (e.g., simple bar charts with two or three groups). We would always recommend presenting a figure for factorial ANOVA models that include an interaction; this allows readers to see the nature of the interaction, as this cannot easily be evaluated from the information provided by significance tests alone.

6. Always include some type of effect size and its associated confidence interval. This can be in the form of unstandardized units such as mean differences, or standardized units such as Cohen's *d*. Point estimates of effect size provide readers with some important information, but including the variability on the effect (e.g., CI) is often more useful. This information should also be discussed in the paper's results and discussion sections, not just whether a finding was statistically significant or not.

Overall, we found it encouraging that most of the articles we examined from these four Canadian journals in psychology reported effect sizes, along with information about statistical tests and the associated *p* values. Researchers should be encouraged to also include confidence intervals to highlight the degree of uncertainty around their effect sizes. Although not all computer programs provide CI information, online sources are available (e.g., Soper, 2006-2016). It would also be helpful to provide more information on missing data and assumptions to allow for more accurate assessment on the adequacy of the study and its findings. Finally, whereas we hope that our suggested recommendations can help researchers incorporate better reporting practices, we are aware that these implementations will take time and will depend on support from journal editors and research associations such as CPA and APA.

References

Abelson, R. P. (1997). The surprising longevity of flogged horses: Why there is a case for the significance test? *Psychological Science, 8*, 12–15. doi: 10.1111/j.1467-9280.1997.tb00536.x

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology. *American Psychologist, 63,* 32–50. doi: 10.1037/0003-066X.63.1.32.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage Publications.

American Psychological Association [APA]. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: APA.

Anderson, S., & Maxwell, S. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21,* 1-12. doi: 10.1037/met0000051

Baraldi, A. N., & Enders, C. (2013). Missing data methods. In T.D. Little's (Ed), *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2: Statistical Analysis* (pp. 635-664). New York, NY: Oxford.

Blanca, M. J., Arnau, J., Lopez-Montiel, D., Bono, R., & Bendayan, R. (2011) Skewness and kurtosis in real data samples. *Methodology, 92*, 78-84. doi: 10.1027/1614-2241/a000057

Bodner, T. E. (2006). Missing data: Prevalence and reporting practices. *Psychological Reports, 99*, 675−680. doi: 10.2466/PR0.99.3.675-680

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility*. Beverly Hills, CA: Sage.

Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997-1003. doi: 10.1037/0003-066X.49.12.997

Coombs, W. T., Algina, J., & Oltman, D. O. (1996). Univariate and multivariate omnibus hypothesis tests selected to control type I error rates when population variances are not necessarily equal. *Review of Educational Research, 66*, 137-179. doi: 10.3102/00346543066002137

Cousineau, D. (2014) Restoring confidence in psychological findings: A call for direct replication studies. *The Quantitative Methods for Psychology, 10*, 77-79.

Cribbie, R. A., Fiksenbaum, L., Wilcox, R. R. & Keselman, H. J. (2012). Effects of nonnormality on test statistics for one-way independent groups designs. *British Journal of Mathematical and Statistical Psychology, 65*, 56-73. doi: 10.1111/j.2044-8317.2011.02014.x

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*, 286–300. doi: 10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Denis, D. (2003). Alternatives to null hypothesis significance testing. *Theory & Science, 4*, 1-21. Available at: http://theoryandscience.icaap.org/content/vol4.1/02_denis.html.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Friendly, M. (2000). *Visualizing categorical data*. Cary, NC: SAS Institute, Inc. doi: 10.1002/jhbs.20078

Friendly, M. & Meyer, D. (in press). *Discrete data analysis with R: Visualization and modeling techniques for categorical and count data.* CRC Press.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review, 18*, 3–12. doi:10.1177/1088868313507536

Gergen, K. J., Josselson, R., & Freeman, M. (2015). The promises of qualitative inquiry. *American Psychologist, 70*, 1-9. doi: 10.1037/a0038597

Glass, G V, Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed-effects analysis of variance and covariance. *Review of Educational Research, 42*, 237-288.

Goertzen, J. R. (2008). On the possibility of unification: The reality and nature of the crisis in Psychology. *Theory Psychology, 18*, 829-852. doi: 10.1177/0959354308097260

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576. doi: 10.1146/annurev.psych.58.110405.085530

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*, 15-24. doi: 10.1037/0003-066X.52.1.15

Harlow, L. L. (2010). On scientific research: The role of statistical modeling and hypothesis testing. *Journal of Modern Applied Statistical Methods, 9*, 348-358. Available at: http://digitalcommons.wayne.edu/jmasm/vol9/iss2/4

Harlow, L. L., Korendijk, E., Hamaker, E. L., Hox, J., & Duerr, S. R. (2013). A meta-view of multivariate statistical inference methods in European psychology journals. *Multivariate Behavioral Research, 48,* 749-774. doi: 10.1080/00273171.2013.822784

Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (Eds.) (2016). *What if there were no significance tests?* Classic Edition. New York, NY: Routledge.

Harris, R. J. (1997). Significance tests have their place. *Psychological Science, 8*, 8-11. doi: 10.1111/j.1467-9280.1997.tb00535.x

Hoekstra, R., Kiers, H. A. L, & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology, 3*, 1-9. doi: 10.3389/fpsyg.2012.00137

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., … Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386. doi: 10.3102/00346543068003350

Kline, R. (2013). *Beyond significance testing: Reforming data analysis methods in behavioral research.* (2nd ed.). Washington, DC: American Psychological Association.

Little, R. J A., & Rubin, D. B. (2002). *Statistical analysis with missing data.* (2nd ed.). New York, NY: Wiley.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*, 579–619. doi: 10.3102/00346543066004579

Margolis, E., & Pauwels, L. (Eds.), (2011). *The SAGE handbook of visual research methods*. Thousand Oaks, CA: SAGE Publications.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166. doi: 10.1037/0033-2909.105.1.156

Mulaik, S. A., Raju, N. S. & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65-115). Mahwah, NJ: Erlbaum.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615-631. doi: 10.1177/1745691612459058

Nuijten, M. B., Hartgerink, C. H. J., Marcel A. L. M., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, October online. doi: 10.3758/s13428-015-0664-2

O'Neill, P. (2002). Tectonic change: The qualitative paradigm in psychology. *Canadian Psychology*, *43*, 190-194. doi: http://dx.doi.org/10.1037/h0086915

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349* (6251). http://dx.doi.org/10.1126/science.aac4716

Reinhart, A. (2015). *Statistics Done Wrong: The Woefully Complete Guide*. No Starch Press.

Rennie, D. L., Watson, K. D., & Monteiro, A. M. (2002). The rise of qualitative research in psychology. *Canadian Psychology*, *43*, 179-189. doi: http://dx.doi.org/10.1037/h0086914

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*, 1-12. doi: 10.1037/a0018326

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335-391). Mahwah, NJ: Erlbaum.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Eds.), *What if there were no significance testing?* (pp. 37-64). Mahwah, NJ: Erlbaum.

Soper, D. (2006-2016). Confidence interval calculators (retrieved online, March 28, 2016: http://www.danielsoper.com/statcalc/category.aspx?id=4).

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*, 423–432. doi: 10.1002/pits.20234

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

Wilkinson, L., & The APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals guidelines and explanations. *American Psychologist, 54*, 594–604. doi: 10.1037/0003-066X.54.8.594

Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials Review, 1*, 368−376. doi: 10.1191/1740774504cn032oa

Table 1

Prevalence of Statistical Analyses

Analysis	***N***	**%**
ANOVA	38	24.8
Z or *t* test on means	23	15.0
Multiple regression	21	13.7
Correlation	15	9.8
Chi square	15	9.8
Structural Equation Models	8	5.2
Logistic Regression	4	2.6
Factor Analysis (or Principle Components Analysis)	4	2.6
Descriptive Only	4	2.6
ANCOVA	3	2.0
Multilevel/Mixed Effects Models	3	2.0
Generalized linear models	3	2.0
MANOVA	2	1.3
Mann-Whitney U test	2	1.3
Z test on dependent correlations	2	1.3
Simulation study	2	1.3
Meta-analysis	1	0.7
Discriminant function analysis	1	0.7
Robust canonical correlation	1	0.7
MANCOVA	1	0.7
Total	153	100.0

Table 2

Content Reported from Statistical Analyses

Inference Information	***N***	**%**
Conducted significance test	136	88.9
Includes effect size	117	86.0
Reported dichotomous *p* values	75	*55.1
Reported exact *p* values	71	*52.5
Includes standard error	36	26.4
Effect size reported with no sig. test	10	6.5
Reported confidence interval on effect	16	10.5
Includes figure with data	48	31.4
Figure has error bar	27	56.3
Includes information on missing data	46	30.1
Includes information on at least one statistical assumption	46	30.1
Total Analyses	153	

Note: If information is indented, the percentage refers to the parent category and not the total number of analyses. For example, of the 48 analyses that included a figure, 27 included an error bar such that 56.3% included it, whereas only 17.6% of all analyses included a figure with an error bar.

* These do not add up to 100% because sometimes researchers reported both dichotomous and exact *p* values within the same analysis.