EFFECT SIZES AND CONFIDENCE INTERVALS FOR MULTIVARIATE ANALYSIS: HOW COMPLETE ARE PUBLISHED ACCOUNTS OF RESEARCH IN PSYCHOLOGY?

Fiona Fidler[1], Lisa L. Harlow[2], Geoff Cumming[1] and Jacenta Abbott[1]
[1]School of Psychological Science, La Trobe University, Australia
[2]University of Rhode Island, United States of America
f.fidler@latrobe.edu.au

Effect sizes (ESs) and confidence intervals (CIs) are widely advocated in psychology and other disciplines. To date most expository articles have focused only on univariate analyses, despite there being similarly good reasons for reporting and interpreting ESs and CIs following multivariate analyses. We surveyed articles published in leading psychology journals in 2008 to discover: a) which multivariate methods were in common use, b) what types of ESs accompany typical multivariate reports, c) whether CIs on ESs were routinely reported d) whether error bars are reported in figures and e) what software authors were using to conduct these analyses. Our results revealed varying traditions of ES reporting for different multivariate techniques, but CIs were in all cases rare. These results highlight areas for software development and for increased educational efforts.

STATISTICAL REFORM: EFFECT SIZES AND CONFIDENCE INTERVALS

Advocates of statistical reform have long argued that complete reporting of results involves not just test statistics and *p* values, but also Effect Sizes (ESs) and Confidence Intervals (CIs) (e.g., Harlow, Mulaik & Steiger, 1997; Kline, 2004). There are literally hundreds of articles calling for increased reporting of ESs and CIs in psychological research (Fidler, 2005). There has also been increasing institutional support for this position, in particular, from the American Psychological Association (APA). The most recent edition of the APA *Publication Manual* (2010) states: "APA stresses that NHST [Null Hypothesis Significance Testing] is but a starting point and that additional reporting elements such as effect sizes, confidence intervals, and extensive description are needed" (p. 33).

Many journal surveys have examined statistical reporting practices in psychology, especially practices associated with NHST and alternatives advocated by reformers (e.g., Kirk, 1996). Even relatively recently, adoption of reform practices appears relatively slow (Cumming et al., 2007). One objection to changing practice is what has been called elsewhere the 'pragmatic argument' (Grayson, Pattison & Robins, 1997). The pragmatic argument highlights the fact that advocates of ESs and CIs have often relied too heavily on oversimplified research scenarios to fulfil their rhetorical needs. The examples reformers provide to demonstrate how p values are easily replaced by ESs and CIs are usually of simple two-group independent t tests or similar. How to affect this substitution in complex multivariate designs is not necessarily straightforward. With the majority of statistical developments over the last half century being made within a significance testing framework, the scope of application for ESs and CIs remains comparatively narrow.

Our journal survey, reported below, was designed to uncover which multivariate techniques are in common use and what, if any, ES and CI reporting practices exist in these traditions. Our aim was to identify areas that further technical and software development and statistics education may address.

REPORTING PRACTICES IN MULTIVARIATE RESEARCH

We surveyed statistical reporting practices in articles with multivariate designs in two leading psychology journals, *Journal of Abnormal Psychology* and *Journal of Social and Personality Psychology*. (Our survey is ongoing, and will eventually include more journals and more time periods.) We used advance library search techniques to search the full text of all articles published in 2008 issues of these journals. This search identified 198 articles using multivariate techniques, which were subsequently coded for reports of the items shown in Tables 1 to 3.

Multiple Regression (MR) and Hierarchical linear models (HLM) were the most commonly reported multivariate techniques with n=47 and n=46 articles using these procedures respectively. Structural Equation Modeling (SEM) and Factor Analysis (FA) were also common (n=37 and n=36).

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

Other analysis recorded included: ANCOVA (n=32), Logistic Regression (LR, n=24), Longitudinal methods (Long, n=21), MANOVA (n=8) and Taxonomic methods (Tax, n=5).

We found strikingly different reporting cultures for the different multivariate methods. For some, overall test statistics and *p* values are routinely reported, for others, these reports are virtually never made. Similar variability existed in ES reporting. The only consistent finding was an unfortunate one: CIs are exceptionally rare!

Overall Test Statistics and P Values

In ANCOVA and MANOVA an overall test statistic was routinely reported (88% and 100% respectively). In FA, overall test statistics were rarely provided (6%). Other methods fell somewhere in between these extremes, as Table 1 shows. Reporting rates of *p* values for overall tests were surprisingly low, given the general dominance of NHST in psychology. *P* values were most often reported in SEM articles, where they appeared in 60% of articles. Statistical power reporting was predictably low (under 5%) with two exceptions: 30% of HLM and 24% of Long papers reported power.

Table 1. Percentages of articles with various multivariate methods reporting Overall Analysis and Estimation Statistics

	ANCOVA	**MANOVA**	**LR**	**MR**	**FA**	**Tax**	**HLM**	**SEM**	**Long**
	n=32	n=8	n=24	n=47	n=36	n=5	n=46	n=37	n=21
% of total (N=198)*	16	4	12	24	18	3	23	19	11
***Overall Analysis* (% of n)**									
Test Statistic	87.5	100.0	58.3	27.7	5.6	0.0	32.6	59.5	38.1
P value	50.0	50.0	45.8	21.3	5.6	0.0	30.4	59.5	38.1
Statistical Power	3.1	0.0	4.2	0.0	2.8	0.0	30.4	0.0	23.8
***Estimation* (% of n)**									
At least one ES	62.5	87.5	79.2	100.0	91.7	60.0	84.8	91.9	38.1
-With *p* value (% of ES)	85.0	75.0	36.8	80.9	15.2	1.0	69.2	32.4	37.5
-With CI (% of ES)	0.0	0.0	21.1	2.1	0.0	0.0	7.7	8.8	0.0
-With SE (% of ES)	10.0	0.0	21.1	17.0	0.0	0.0	33.3	0.0	12.5

* Note: %s do not sum to 100% as some articles include more than one type of method.

Estimation Statistics: Effect Sizes, and their accompanying p values, CIs, SEs

ES reporting was similarly variable over the different method types, again emphasizing the different traditions associated with different methods. ES reporting was most noticeably low in Long (38%) and ANCOVA (62%). It was remarkably high in MANOVA (87%), FA and SEM (92%), and 100% in MR. Table 1 shows the percentage of articles of each method type reporting at least one ES.

If ESs were accompanied by an inferential statistic, it was more often a *p* value than a CI or SE, but there was considerable variability in this *p* value reporting too. For example, in ANCOVAs, 85% of ESs were accompanied by a *p* value; in MR, ESs with *p* values were similarly common (81%). Yet, for other methods, *p* values only accompany ESs about one third of the time (SEM, 32%; LR, 37%; Longitudinal, 38%).

ESs were rarely reported with CIs, under 10% in every case and 0% for half the methods (see Table 1). The only exception was LR, where ESs were accompanied by CIs 21% of the time. It is tempting to think that elevated rate in LR is a carry-over of the strong tradition of reporting CIs with Odds Ratios in Medicine. However, it is important to keep in mind the low numbers here: 4 out of 19 ESs associated with LR were accompanied by CIs. For some methods there is a reasonably convincing tradition of reporting SEs with ESs: 33% of HLM report ESs with SEs, 21% of LR and 17% of MR reported SEs.

Common types of ESs, with their reporting rates as a percentage of the overall number of articles, are shown in Table 2. Regression weights and correlations (category = B, β, or *r*) were by far the most commonly reported ESs. (Percentages in Table 2 do not sum to 100% because some articles reported more than one type of analysis and/or ES.)

Table 2. Percentage of different types of effect sizes reported in various multivariate methods

Effect Size % of N=198	ANCOVA	MANOVA	LR	MR	FA	Tax	HLM	SEM	Long	TOTAL
η2	6.1	3.0								**9.1**
Cohen's *d*	2.0	0.5					0.5			**3**
f or f^2 (Cohen's)	0.5									**0.5**
B, β, or r	1.5		5.6	18.2	3.5	0.5	12.6	4	1.5	**47.4**
R^2 or shared variance			1	5.1	3		2	1.5	0.5	**13.1**
Odds Ratio or expB			3				0.5			**3.5**
R					0.5					**0.5**
Factor loading					8.6			4		**12.6**
Fit index: CFI, TLI etc						0.5		3	0.5	**4**
RMSEA							0.5	2.5	0.5	**3.5**
Likelihood Ratio							0.5	0.5	0.5	**1.5**
AIC or BIC									0.5	**0.5**
Other				0.5	1.0	0.5	3	1.5		**6.5**

Table 3. Frequency of reported software use

Software	ANCOVA	MANOVA	LR	MR	FA	Tax	HLM	SEM	Long	TOTAL (Freq)
AMOS				1	2			7	1	**11**
EQS	1									**1**
MacANOVA	1									**1**
STATA				1			2			**3**
MPlus					1			4		**5**
SYSTAT						1				**1**
SPSS	1		1	1			4	3		**10**
HLM							12			**12**
SAS							11	4	1	**16**
LISREL								3		**3**

Figures

We also recorded whether or not an article reported a figure with empirical data, and in particular, whether figures included error bars. Fifty-eight percent (58%, 115 of 198) of articles reported a figure. Of those, 16.5% included SEs, 5.2% CIs and 4.3% error bars that were unlabelled. Reporting error bars in figures is a widely recommended practice, and one endorsed by the APA *Publication Manual.* It is surprising, though consistent with other journal surveys in psychology (e.g., Cumming, Fidler, Leonard et al, 2007), that use of error bars is so low. This is an area we identify as in need of development, both in terms of software and education.

Software

Finally, we recorded, where available, the software package authors used. Only 32% (62 of 198) of the sampled articles reported the software used. Table 3 shows the relevant frequencies.

CONCLUSION

ES and CI reporting is widely advocated in psychology, and many journal surveys have been conducted in psychology to assess reporting rates of *p* values, statistical power, ESs, CIs and other statistical procedures. Relatively little, if any, research has focused specifically on statistical reporting practices of multivariate methods.

We found reporting traditions varied greatly between different multivariate methods. For some methods, ES reporting rates are promisingly high. However, we offer the same caution in interpreting those rates as Kirk (1996) did. Kirk surveyed four psychology journals and found particularly high ES reporting in the *Journal of Applied Psychology*. He cautioned: "Before anyone concludes that authors of articles in the *Journal of Applied Psychology* are more aware of the limits of null hypothesis significance testing, remember that these authors are more likely to use regression and correlation procedures. Computer packages routinely provide R^2 for these procedures." (p.754). This is equally true of the most common type of ES reported in our sample:

B, β, or *r*. It would not be surprising if these ESs were reported without recognition that they are in fact ESs, and without relevant interpretation (Note: We did not code for recognition or interpretation.)

In all cases CI reporting was low. The reporting of error bars in figures was also low. We identify these areas, in particular, as in need of development. Instructional and software resources are necessary to encourage more widespread use of these highly desirable practices.

REFERENCES

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

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J., McMenamin, N., & Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18*, 230-232.

Fidler, F. (2004). From Statistical Significance to Effect Estimation: Statistical Reform in Psychology, Medicine and Ecology. *Unpublished doctoral dissertation, University of Melbourne.*

Grayson, D., Pattison, P., & Robins, G. (1997). Evidence, inference and the "rejection" of the significance test. *Australian Journal of Psychology, 49*, 64-70.

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

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

Kine, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioural research.* Washington, DC: American Psychological Association.