Running Head: EFFECT SIZE DISTORTION

Multiple Trials May Yield Exaggerated Effect Size Estimates

Andrew Brand

King's College London, Institute of Psychiatry

Michael T. Bradley, Lisa A. Best and George Stoica

University of New Brunswick

Please send correspondence to:
Dr. Lisa A. Best
Department of Psychology
P.O. Box 5050
University of New Brunswick
Saint John, New Brunswick
E2L 4L5
Canada

Brand, A., Bradley, M. T., Best L. A., & Stoica, G. (2011). Multiple trials may yield exaggerated effect size estimates. *The Journal of General Psychology, 138(1),* 1-11.

Link to published version: http://dx.doi.org/10.1080/00221309.2010.520360

Abstract

Published psychological research attempting to support the existence of small and medium effect sizes may not have enough participants to do so accurately and, thus, repeated trials or the use of multiple items may be used in an attempt to obtain significance. Through a series of Monte-Carlo simulations, this paper describes the results of multiple trials or items on effect size estimates when the averages and aggregates of a dependent measure are analyzed. The simulations revealed a large increase in observed effect size estimates when the numbers of trials or items in an experiment were increased. Overestimation effects are mitigated by correlations between trials or items but remain substantial in some cases. Some concepts such as a P300 wave or a test score are best defined as a composite of measures. Troubles may arise in more exploratory research where the interrelations of amongst trials or items may not be well described.

Keywords: effect size, regression/correlational designs, multiple items, repeated measures,

Multiple Trials May Yield Exaggerated Effect Size Estimates

Despite criticisms, null hypothesis significance testing (NHST) remains the norm in psychological research. Although the flaws of relying solely on NHST are increasingly recognized (i.e., Gliner, Leech, and Morgan, 2002; Hubbard & Lindsay, 2008; Trafimow, 2003; Trafimow & Rice, 2009), many researchers rely solely on these techniques. To counter this reliance, The American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson & the APA Task Force on Statistical Inference, 1999, p. 399) suggested that researchers should "Always provide some effect size estimate when reporting a *p* value."

The effect size is a descriptive measure of the magnitude of the association between two variables that can be used to complement inferential findings. Effect sizes show how much manipulating an independent variable changes a dependent variable. A commonly calculated and cited effect size is the Cohen's *d*, (Cohen. 1969). Cohen's *d* is a standardized effect size. It expresses the magnitude of the difference between two independent group means in standardized deviation units that are metric free. Standardized effect sizes such as Cohen's d can be used to compare effects across an entire literature in which different researchers have used different measures of the dependent variable. An implication of the Wilkinson (1999) recommendation is that effect sizes such as Cohen's *d* may become a central statistical finding of a study since it reports the magnitude of an effect as opposed to whether it merely exists. It also promotes meta-analytical thinking (i.e., statistical comparisons across studies). Thus accurate precise measures of effect size and an understanding of variables that influence effect sizes are important considerations.

Ideally, in an unbiased fashion, effect sizes should reflect the relationship between levels of a variable (drug doses) and a dependant variable (i.e. well being). Unfortunately, due to the typically small sample sizes in psychology research and a prevalent bias for journals to publish research that has obtained statistically significant results ($p < 0.05$), only part of the distribution of potential effect sizes is published. Stated more formally the statistical power of published psychological research to detect small and medium size effects has been shown to be low (Cohen, 1962; Rossi, 1990; Sedlmeier & Gigerenzer, 1989) and, as a consequence, published psychological research tends to overestimate true effect sizes (Brand, Bradley, Best, & Stoica, 2008; Kraemer, Gradner, Brooks, & Yesavage, 1998). Furthermore, meta-analyses and power-analyses using overestimates of effect size will produce misleading results. The meta-analyses will overestimate population effect sizes and the subsequent power-analyses will underestimate sample sizes required to obtain statistically significant results (e.g., Muncer, Craigie, & Holmes, 2002).

When using NHST, researchers want to maximize the likelihood that they will be able to reject the null hypothesis and find support for their alternative hypothesis. One way to increase the likelihood of obtaining statistical significance is to simply increase sample size. Sometimes, however, the number of participants locally available is limited and the sample size required to obtain sufficient statistical power is large (sizeable), this method may not always be feasible. For example, for a two-group between subjects design to have 80% power and detect a small effect ($d = .20$) at the $p = .05$ level, 788 participants would be required. Researchers could also approach significance by decreasing the error variance of the dependent measure. Typically, to decrease the error variance researchers design experiments that include stringent control over potentially influential but theoretically irrelevant variables (e.g. differing times of day for

groups) or culling of items of low relevance. This approach is not always practical because it can require substantial pre-testing of various experimental designs and measures.

Although the effects of increasing sample size on measurement and controlling for extraneous variability have been well documented, a third way to increase the likelihood of rejecting the null hypothesis is to increase the number of data points by having each participant complete multiple trials or respond to multiple items of the dependent measure. The effect that this tactic has on obtaining significance is dependent upon how the repeated measures are handled. An analysis using trial data as a within subjects factor in a split-plot ANOVA will not result in increased statistical power with additional trials (see, Bradley & Russell 1998; Overall 1996); however, researchers more commonly aggregate or average item or trial data. For example, in the August 2008 issue of *Journal of Experimental Psychology: General*, for all of the experiments in which a dependent measure was repeated for each participant, statistical analyses were conducted on either the aggregate or average of the dependent measures. Moreover, if researchers aggregate or average participants' data across all the trials and then use a between subjects design (i.e., different subjects in different conditions) with these aggregates and averages as a single measure of the dependent variable, there will be an augmentation in the estimate of effect size.

Statisticians in the medical sciences have reported that the apparent augmentation of effect size with a subsequent increase in power is influenced by correlations between data from different trials (Frison & Pocock, 1992; Vickers, 2003). For instance, when the correlation between trial data is high, in that the magnitude of the dependent measure is consistent across trials (i.e., participants who score high on a trial will score high on the others trials and conversely participants who score low on one trial will score low on the others trials) effect size

is not augmented by increasing the number of trials. Whereas, when the correlation between trial data is low, in that the magnitude of the dependent measure is inconsistent across trials (i.e., participants who score high on a trial will not necessarily score high on the others trials and conversely participants who score low on one trial will not necessarily score low on the others trials) effect size is augmented by increasing the number of trials. Although, the effect that increasing the number of trials has upon statistical power has been described for various trial data correlations, the effect that increasing the number of trials has upon effect sizes estimation has not. Moreover, such a description is necessary since researchers could, in following the recommendations of Wilkinson and the APA Task Force on Statistical Inference (1999), report exaggerated effect sizes without a full awareness of the size of the exaggeration.

If significance testing diminishes in importance then the reporting of effect sizes may become the primary result of interest. If the effect size is a legitimate reflection of the phenomena under study, valid interpretations could be drawn, whereas, distortions in effect size due to repetitions could result in interpretation problems. The problems are twofold: first, increasing the number of trials or items in a single study can lead to distortions within that study; and, second, if the distortions vary across studies, it would be difficult to compare the results of several studies. Given the widespread practice of averaging or aggregating trial/item data, the point of this paper is to more precisely outline, through the use of Monte-Carlo simulations, how this method of analyzing multiple trial/item data affects effect size estimations over a variety of correlations amongst the trials. Our primary goal is to describe the joint effects of repetition and intercorrelation on effect size estimates.

The Monte-Carlo Simulations

The simulation was conducted using the statistical package R (R Development Core Team, 2009, http://www.r-project.org/)[1]. The simulation was based on Cohen's *d* standardized effect size. Cohen's *d* is calculated by dividing the mean difference between two independent groups by the pooled standard deviation that is a weighted average of the standard deviation of the two groups. The simulation involved varying 3 factors: true effect size (e.g., $d = 0.20$, $d = 0.50$ and $d = 0.80$), the average correlation between trial data (e.g., 0, 0.20, 0.50, 0.80 and 1) and the number of trials in an experiment (e.g., 1, 5, 10, 20 and 30). For each of the 76 different combinations of the factors 100,000 experiments were simulated. This resulted in a total of 7,600,000 simulated experiments.

A control distribution with a standard normal distribution of 1,000,000 values, a mean of 10, and a standard deviation of 2 was generated for each trial in the experiment. The values from the control distributions are then placed in a matrix, in which a row represented each participant and a column represented each trial in an experiment. The cells in the columns (i.e., data for a given trial) were then re-arranged so that the average correlation between columns (i.e., trial data) was obtained. To achieve this, all the values in a column were placed in ascending order and then a given number of values were randomly rearranged. To create the experimental data this procedure was repeated using different means to produce the different effect sizes (e.g. $M = 10.40$ when $d = 0.20$; $M = 11.00$ when $d = 0.5$; and $M = 11.60$ when $d = 0.8$).

Based on the results of survey of statistical power of psychological research and typical sample sizes used in psychological research (Rossi, 1990), a simulated experiment involved 76 participants. To simulate an experiment, 38 rows from the control data matrix and 38 rows from

[1] The R scripts are available from the author upon request.

the experimental data matrix were randomly selected and the aggregates (sum) and (mean) averages for these rows were calculated. The observed effect size from a simulated experiment was calculated. For each set of 100,000 simulated experiments, the means for the observed effect size, the percentage of the differences between the mean observed effect size, and the true effect size were calculated. These results are summarized in Tables 1, 2, and 3. The results from multi-trial experiments in which aggregates were used as the dependent measure were identical to the results from those in which averages were used the dependent measure and, thus, the values in Tables 1, 2, and 3 refer to multi-trial experiments in which *either* averages and aggregates of the dependent measure were analyzed.

Insert Table 1,2 and 3 About Here.

The Monte-Carlo simulation revealed that when a between subjects analysis was used with no correlation between trials the observed effect size was equivalent to the effect size put into the design for single trial experiments and was substantially higher in multiple-trial experiments than in single trial experiments. For the aggregate measure, this occurs because the mean difference between the control group and the experimental group (i.e., the numerator of the effect size) increased as the number of trials increased whereas for the average measure the effect is due to the fact that the pooled standard deviation (i.e., the denominator of the effect size) decreases as the number of trials increases. Tables 1, 2, and 3 not only shows that the observed effect sizes from multi-trial experiments are larger than those seen with a single trial experiment but that the magnitude of the observed effect is moderated by the correlation between trial data. Although it is not necessarily surprising that items that are not perfectly correlated result in an inaccurate

estimate, it is important to note that the effect size distortion was *always* an overestimation and *never* an underestimation. Table 1 shows that an inter-item correlation of zero results in extreme distortion (460%). It must be noted, however, that a zero correlation is extremely unrealistic because repeated measurements, either on the same measure or on related items, are likely to be correlated. When r = .2 and r =. 5 distortions ranging from 35% to 115% occur. When the correlation is high (r =.8), the distortions are minimal and, finally, at r =1.0, there are essentially no observed effect size increments, although in many cases we might expect a high correlation between trials and items researchers do not routinely report inter-trial and inter-item correlations. Furthermore, even in cases in which the inter-item and inter-trial is high (r = .8) , there is still a 14 - 15% distortion in the effect sizes. Although this distortion is not as large as for the lower correlations it is still sizeable and should be of concern to researchers.

It is noteworthy that this pattern of increase was repeated across all three effect sizes with the same proportional increase over the trials. For example, considering only conditions with r = 0 and regardless of the true effect size, the percentage of distortion increased from 129% for 5 trials experiments to 460% for 30 trials experiments. As a result the observed effect sizes from multi-trial experiments with a large number of experimental trials substantially overestimate the true effect based on the raw value input. It should be noted that for simple effect size measures based solely on the mean difference there is no distortion in effect size when the average measure is employed for a multi-trial experiment. However, even though this has been encouraged by Baguley (2009) and Bond, Wiitala and Richard, (2003), researchers rarely report simple effect sizes or conduct meta-analysis using simple effect sizes. Furthermore, the effect size distortion will still occur for unstandardized effect sizes measure if the aggregate measure is employed. This is because increasing the number of trials, when the correlation between trial

data is not 1, will increase the mean difference.

Implications of the Monte-Carlo Simulations

The current results are both profound and mundane depending upon the sophistication of the researcher and readers of such studies. Part of the profundity of the results for the current authors was the sheer magnitude of the increase in the observed effect size as the number of trials increase. To appreciate this, we outline the implications for different research scenarios in which a multiple number of trials or test items are used. In one scenario, researchers arbitrarily decide upon the number of trials or test items; for example, a face recognition experiment may have 60 test faces or a word categorization experiment may have 80 test words. Not only will the observed effect sizes based on aggregates or averages overestimate the true underlying effect sizes to the degree that trials or test items are increased, there is a very good chance that these study results will be incomparable to other results simply because various experimenters have arbitrarily chosen different numbers of trials or test items.

Another incomparable aspect across studies could involve intercorrelations either between trials or items. Researchers who use well-established questionnaires may report the inter-item reliabilities (Cronbach's alpha) and generally accept that if the inter-item reliability is greater than .8, the test is a reliable measure of an underlying construct. In fact, if the test items are part of a reliable and validated measure in which interrelated items measure various aspects of same concept, the aggregate scores provide a more accurate measure of the phenomenon being investigated and, hence, analyses using these measures are not distorted. For example, the score on a Beck's Depression Inventory (Beck, Ward, & Mendelson, 1961) or Eysenck's Personality

Inventory (Eysenck, & Eysenck, 1975) reflect complex concepts not adequately explicable in a one item questionnaire but definable through several interrelated items.

With a trials effect across a broad range of intercorrelations, effect size changes will influence power such that they are more likely to be found statistically significant and, therefore, be published. As a result, researchers surveying the literature will be under the impression that an effect is considerably larger than it is and meta-analyses based on the published effect sizes of such a literature will be greatly distorted.

As already mentioned**,** some concepts are the sum or average of individual interrelated elements. Experimental designs with inherently noisy EEG recordings or reaction times may demand the aggregation or averaging of trial-by-trial data and specific concepts emerge only through repetition. For example, an Evoked Response Potential (ERP) only becomes visible from Electroencephalographic (EEG) recordings through repeated estimates. The trial-by-trial repetition is necessary and reduces variability associated with single trial data. Trial data is averaged to determine the most probable waveform. The ERP waveforms that were defined by repetition or more exactly through replication are the important results of the study and often graphical comparisons replace inferential analyses. In physiological studies such as these the averaging of trial by trial data is legitimate and is necessary for the evaluation of results. Furthermore, researchers who use these measures typically report the number of trials that are averaged. Ultimately researchers may come to a consensus as to the optimal number of repetitions for different experimental designs.

It is worth repeating that a mitigating factor involves the correlation between trials and items. We included, to describe the universe of possible outcomes, the improbable situation where the correlation between trials is 0. As correlations strengthen the augmentation in effect

size decreases such that augmentation is zero when $r = 1.0$. It is particularly interesting that the effect is strongest in the first five trials and thereafter is strongly muted. That is, there is a 65% growth in effect size with an $r = .2$ in 5 trials followed by a 25% growth for 10 trials. At an input effect size of .5, the 5-trial increment is approximately 35% and this increases to 45% with 30 trials. At an $r = .8$, the effect size increment is 15% by 30 trials with most of that increment (10%) in the first 5 trials.

We would suggest that one solution to the problems associated with correlated items and trials would be for researchers to simply report the numbers of trials that are being aggregated or averaged and the average correlation between those items or trials. In doing so, readers and other researchers would have a strong sense of the distortion in effect sizes that could occur and provide their audience with an indication of the strength of their results. Given this practice, it might follow that the number of trials or items could be standardized so that there could be guidelines about the number of trials or items that are optimal for specific types of research.

Conclusions

In summary, increasing the number of trials or test items in an experiment distorts observed effect size through increases if aggregates or averages of the dependent measure are analyzed. This article describes this distortion and notes that the increase in effect size creates some complex situations. These situations require judgment because the cost of more trials or test items is the potential distortion of estimates. Hence, the value of analyzing aggregates or averages of a dependent measure from multi-trial experiments depends upon the conceptualization that the researcher has in mind. If the concept is best reflected by an aggregate

of measures then the researcher has to respect the idea that they have an exceptionally powerful design and may want to analyse the results with both a liberal test and the more conservative split-plot design to gain a full understanding of a range of potential effect sizes. In the same spirit, researchers need to be made aware that analyzing aggregates or averages of a dependent measure in multi-trial experiments distorts effect size estimates. Knowing this means researchers they can be cautious when interpreting results from multi-trial experiments and meta-analyses that are based on multi-trial experiments. The mitigating factor is the degree of correlation between the items or trials but this factor brings in additional complications related to the coherence or incoherence of the items included in measurement.

References

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*, 603-617.

Beck, A.T., Ward, C., & Mendelson, M. (1961). Beck Depression Inventory (BDI). *Archives of General Psychiatry, 4*, 561-571.

Bond, C. F., Jr., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8*, 406-418.

Bradley, D.R., & Russell, R.L. (1998). Some cautions regarding statistical power in split-plot designs. *Behavior Research Methods, Instruments, & Computers, 30*, 462-477.

Brand, A., Bradley, M. T., Best L. A., & Stoica, G. (2008). Accuracy of effect size estimates from published psychological research. *Perceptual and Motor Skills, 106*, 645-649.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 69*, 145-153.

Eysenck, H. J. & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire*. London: Hodder and Stoughton.

Frison, L. and Pocock, S. J. (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implication for design. *Statistics in Medicine, 11*, 1685-1704.

Gliner, J. A., Leech, N. L., and Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education, 71*(1), 83-92.

Hubbard, R., & R. M. Lindsey. (2008). Why P values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology 18*, 69-88.

Kraemer, H. C., Gradner, C., Brooks, J. O., & Yesavage, J.A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3*, 23-31.

Muncer, S. J., Craigie, M., & Holmes, J. (2003). Meta-analysis and power: Some suggestions for the use of power in research synthesis. *Understanding Statistics, 2*, 1-12.

Overall, J.E. (1996). How many repeated measurements are useful? *Journal of Clinical Psychology, 52*, 243-252.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years?

Journal of Consulting and Clinical Psychology, 58, 646-656.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review, 110*, 526-535.

Trafimow, D., & Rice, S. (2009). What if social scientists had reviewed great scientific works of the past? *Perspectives in Psychological Science, 4*, 65-78.

Vickers, A. J. (2003). How many repeated measures in repeated measures designs? Statistical issues for comparative trials. *BMC Medical Research Methodology, 3*, 22-31.

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

Table 1: Mean Observed Effect Size and the Mean Percentage of the Effects Size Distortion as a function of the Correlation between Trial Data and the Number of Trials in the Experiment when the True Effect Size is Small (d = 0.20).

	Trials in the Experiment				
	1	5	10	20	30
Correlation Between Trial Data = 0					
Mean Observed effect size	0.20	0.46	0.65	0.92	1.12
% of Effect Size Distortion	0	130	225	360	460
Correlation Between Trial Data = 0.2					
Mean Observed effect size	0.21	0.34	0.39	0.42	0.43
% of Effect Size Distortion	5	70	95	110	115
Correlation Between Trial Data = 0.5					
Mean Observed effect size	0.20	0.27	0.28	0.28	0.29
% of Effect Size Distortion	0	35	40	40	45
Correlation Between Trial Data = 0.8					
Mean Observed effect size	0.20	0.22	0.23	0.23	0.23
% of Effect Size Distortion	0	10	15	15	15
Correlation Between Trial Data = 1.0					
Mean Observed effect size	0.20	0.20	0.21	0.20	0.20
% of Effect Size Distortion	0	0	5	0	0

Table 2: Mean Observed Effect Size and the Mean Percentage of the Effects Size Distortion as a function of the Correlation between Trial Data and the Number of Trials in the Experiment when the True Effect Size is Medium (d = 0.50).

	Trials in the Experiment				
	1	5	10	20	30
Correlation Between Trial Data = 0					
Mean Observed effect size	0.51	1.14	1.62	2.29	2.80
% of Effect Size Distortion	2	128	224	358	460
Correlation Between Trial Data = 0.2					
Mean Observed effect size	0.51	0.85	0.97	1.04	1.07
% of Effect Size Distortion	2	70	94	108	114
Correlation Between Trial Data = 0.5					
Mean Observed effect size	0.51	0.66	0.69	0.71	0.72
% of Effect Size Distortion	2	32	38	42	44
Correlation Between Trial Data = 0.8					
Mean Observed effect size	0.51	0.56	0.57	0.57	0.57
% of Effect Size Distortion	2	12	14	14	14
Correlation Between Trial Data = 1.0					
Mean Observed effect size	0.51	0.51	0.51	0.51	0.51
% of Effect Size Distortion	2	2	2	2	2

Table 3: Mean Observed Effect Size and the Mean Percentage of the Effects Size Distortion as a function of the Correlation between Trial Data and the Number of Trials in the Experiment when the True Effect Size is Large (d = 0.80).

	Trials in the Experiment				
	1	5	10	20	30
Correlation Between Trial Data = 0					
Mean Observed effect size	0.82	1.83	2.59	3.66	4.48
% of Effect Size Distortion	2	129	224	358	460
Correlation Between Trial Data = 0.2					
Mean Observed effect size	0.82	1.37	1.55	1.67	1.72
% of Effect Size Distortion	2	71	94	109	115
Correlation Between Trial Data = 0.5					
Mean Observed effect size	0.82	1.06	1.11	1.14	1.15
% of Effect Size Distortion	2	33	39	42	44
Correlation Between Trial Data = 0.8					
Mean Observed effect size	0.82	0.89	0.90	0.91	0.91
% of Effect Size Distortion	2	11	12	14	14
Correlation Between Trial Data = 1.0					
Mean Observed effect size	0.82	0.82	0.82	0.82	0.82
% of Effect Size Distortion	2	2	2	2	2

Note: Cohen's d is a biased estimate when n is small (Hedges & Olkin, 1985).