The Journal of General Psychology, 2012, *139*(4), 260–272

More Voodoo Correlations: When Average-Based Measures Inflate Correlations

ANDREW BRAND
King's College London

MICHAEL T. BRADLEY
University of New Brunswick

ABSTRACT. A Monte-Carlo simulation was conducted to assess the extent that a correlation estimate can be inflated when an average-based measure is used in a commonly employed correlational design. The results from the simulation reveal that the inflation of the correlation estimate can be substantial, up to 76%. Additionally, data was re-analyzed from two previously published studies to determine the extent that the correlation estimate was inflated due to the use of an averaged based measure. The re-analyses reveal that correlation estimates had been inflated by just over 50% in both studies. Although these findings are disconcerting, we are somewhat comforted by the fact that there is a simple and easy analysis that can be employed to prevent the inflation of the correlation estimate that we have simulated and observed.

Keywords: correlation, effect size, averaging, reliability, inflation

CENTRAL TO THE INTEGRITY OF PSYCHOLOGICAL RESEARCH is the accuracy of its reported effect size estimates. Due to the dominance of Null Hypothesis Significance Testing (NHST), the reporting of the effect size was at times somewhat ignored and its importance de-emphasized. This problem has changed with recommendations by Wilkinson and the APA Task Force on Statistical Inference (1999). These recommendations have encouraged and promoted the reporting of effect sizes. Consequently, the accuracy of observed effect sizes and their interpretability should now be of primary concern to researchers.

The emphasis on effect size reporting has resulted in a wider understanding amongst researchers that published effect sizes are likely to be inflated estimates

The authors would like to thank Cherl McCormick and Carl Hodgetts for providing us with their raw data.

Address correspondence to Dr. Andrew Brand, King's College London, Psychology, 16 De Crespigny Park, London SE5 8AF, UK; andrew.brand@kcl.ac.uk (e-mail).

of the true effect size when statistical power is low and results obtain statistical significance (Brand, Bradley, Best, & Stoica 2008; Lane & Dunlap, 1978). Perhaps less widely appreciated are effect size distortions that result from specific details associated with experimental design and the statistical analysis of a study. For instance, Pearson's correlation coefficient (r), which can be regarded as a standardized effect size estimate, can be distorted when the range of a measure is restricted (Baguley, 2010). For example, when only the extreme values of a measure are sampled the correlation estimate will be inflated (Preacher, Rucker, MacCallum & Nicewander, 2005). Alternatively, a correlation estimate can be inflated or deflated by non-representative sampling of the experimental stimuli (Fielder, 2011). In other words, researchers may, inadvertently or advertently, select stimuli to maximize the correlation between two measures rather than try and select stimuli that are representative of the stimuli population. High-profile examples of distorted correlations are from the inflated so called "Voodoo" correlations in social neuroscience. These result from non-independent sampling of volume elements or voxels in fMRI studies (Vul, Harris, Winkielman, & Pashler, 2009) where reported correlations were derived from only the voxels that were strongly correlated with the behavioral measure.

Equally alarmingly and noteworthy, especially due to its prevalence, is effect size distortion resulting from the common practice of averaging or aggregating data across multiple trials. Brand and colleagues (2011), in their simulation, showed that Cohen's d effect size estimates would inflate as a joint function of the number of trials and the average inter-trial reliability. For instance, when the number of trials in an experiment is 10, and the average inter-trial reliability is .50, the reported effect size estimate is inflated by approximately 39%.

A potentially troublesome correlation design that is employed by researchers involves correlating a measure (X) that is based on the average ratings from n raters with a measure (Y) based on a single measurement. For instance, Feinberg and colleagues (2008) reported that the average attractiveness of 123 female voices based on the ratings of 10 males is correlated ($r = .34$) with the pitch of the female voices. This correlation value gives a fairly definite impression that a noteworthy effect size is present. But this also raises the question: is this a reflection of individual estimates or is it predominately an artifact of averaging group values?

It is not our purpose to review the many other recent examples as intriguing and topical as the research conducted by Feinberg et al. (2008), but we will mention a few published in high impact journals to show that this is a common approach. Carré and colleagues (2009) reported that the correlation between the average perceived aggressiveness of 24 male faces based on the ratings of 31 raters and the width to height ratio of the male faces was .59. Coeztee and colleagues (2010) observed a negative correlation of $-.61$ between the average perceived body weight of 43 female faces based on the ratings of 26 raters and the cheek to jaw width ratio of the 43 females faces. Hodgetts and colleagues (2009) reported a negative correlation of $-.86$ between the average similarity judgment of 81 object

pairs based on 42 raters and the transformational distance between the objects of the pair. Savani and colleagues (2011) obtained a negative correlation of −.63 between the average victim blame rating of 54 raters based on their ratings across 6 vignettes and their rated political orientation. Rule and Ambady (2008) observed a correlation of .30 between the average Chief Executive Officer leadership rating based on the ratings of 50 raters and the profits of 43 companies.

We point out these particular studies because these findings, in high impact journals, especially when they are on potentially topical issues, are likely to be widely disseminated. Whenever, as in the above studies, ratings are averaged or aggregated prior to producing correlations they will be inflated in comparison to the correlations produced at the individual level and then averaged. If the distinction is not clear, a sizable section of the research community may mistake the inflated value from the group-based measure as an accurate reflection of the individual-based measure.

Monin and Oppenheimer (2004) and Nickerson (1995) identify the problem associated with these types of findings. They highlight and demonstrate the importance of distinguishing between correlated averages and averaged correlations. Monin (2003) had demonstrated a particular phenomenon through the use of correlated averages. Later Monin and Oppenheimer (2004) re-analyzed their earlier data to demonstrate that their earlier correlated averages were substantially larger than the averaged correlations. More importantly, they cogently argued that because researchers are often interested in the correlation between two measures at an individual level and not at an, arbitrary defined, group level, it is the averaged correlations that should be calculated and reported. They clearly stated that two problems arise from the calculation of the correlated average, one is the inflation we have focused on and the other is that inflation is related not only to the number of ratings averaged but also the reliability of the average-based estimates.

In this article we extend on the work of Monin and Oppenheimer (2004) and Nickerson (1995), and use the Monte-Carlo simulation approach by Brand, Bradley, Best, and Stoica (2011) to define the range of inflation that correlated averages may yield over average correlations. Firstly, we conducted a Monte-Carlo simulation to assess the extent that average-based measures inflate correlation estimates. Secondly, we re-analyzed raw data from two recently published studies that adopted a correlation design. The Monte-Carlo approach maps out the potential range of distortion and the empirical approach provides a concrete example of inflated correlations that confirm the correctness of Monin and Oppenheimer's reanalysis and observation.

The Monte-Carlo Simulation

The simulation was conducted using the statistical package R (R Development Core Team, 2011, http://www.r-project.org/). The simulation involved varying

3 factors: the true effect size (e.g., r = .10, r = .30 and r = .50), the average inter-rater reliability (e.g., .30, .50 and .80) and the number of raters in the study (e.g., 1, 5, 10, 20 and 30). Note that the values used for the true effect size correspond to Cohen's (1988) small, medium and large effect size benchmarks. Further note that the true effect size refers to the average of the single rater correlations. For each of the 45 different combinations of the factors 10,000 studies were simulated. This resulted in a total of 450,000 simulated studies. The design of the correlation study involved correlating a measure X that is based on the mean average rating with a measure Y that was based on a single measure. The correlation was based on 40 test items, where participants' ratings for measure X were on a 1 to 8 Likert scale. It is important to note that additional simulations were conducted that showed that the percentage of inflation of the correlation is not affected by the number of test items that are correlated, nor the means or standard deviations of the measure X and measure Y. So the generalizability of the findings from this simulation is not affected by the arbitrary selection of these parameters.

Forty integers from a normal distribution with a mean of 100 and a standard deviation of 20 were randomly generated for the measure Y. Data for measure X was generated as follows. For each rater in the experiment, 40 integers in the range 1 to 8 were generated from a normal distribution with a mean of 4 and a standard deviation of 2. These values were then placed in a matrix in which a row represents each of the 40 test items and a column represents each rater. The cells in the columns (i.e., data for a given rater for a given test item) were then re-arranged so that the average correlation between columns (i.e., the average inter-rater reliability) was obtained. To achieve this, all the values in a column were placed in ascending order and then a set number of values in a column were randomly rearranged. Next the 40 values of measure Y were reordered by placing them in ascending order and then a given number of values were randomly rearranged so that on average the average correlation between the measure X and the measure Y for each rater is equivalent to the true effect size. Note that this procedure will yield more accurate results than Brand et al. (2011), because the specified average inter-rater reliability (i.e., .30, .50 and .80) is obtained for each individual sample as opposed to just the entire population.

To simulate a study, the (mean) average rating for each of the 40 test items was calculated to derive the averaged measure X, and then that measure X was correlated with measure Y. For each set of 10,000 simulated experiments, the mean of the correlation coefficients (r), the percentage of the difference between the mean observed correlation and the true effect size were calculated. Additionally, in order to help identify why the correlation coefficient is being distorted, for each set of 10,000 simulated experiments we calculated, the standard deviation of X, the standard deviation of Y, the covariance between X and Y (i.e., the numerator of the correlation coefficient), and the standard deviation X multiplied by the standard deviation Y (i.e., the denominator of the correlation coefficient).

The results of the Monte-Carlo simulation are summarized in Table 1.[1] The results show that increasing the number of raters increases the percentage of distortion in the observed correlation estimate. The magnitude of the distortion is affected by the average inter-rater reliability, in that the lower the inter-rater reliability the greater effect size distortion. It is also important to note that the pattern in the percentage of distortion of the observed correlation is equivalent across the three different true effect sizes. The degree that the correlation estimate is distorted is most pronounced for the first 5 raters and thereafter becomes increasingly muted. The magnitude of the distortion of the correlation is fairly substantial. For instance looking at a correlation of .30, when there are 10 raters and the inter-rater reliability is .50 the inflated correlation is .40 for a percentage distortion of 35%. Moreover, if we square the inflated correlation estimate to obtain the coefficient of determination (R^2), a commonly used effect size index that equates to the proportion of variance accounted for, we subsequently conclude that the correlation accounts for 16% of the variance rather than 9%. This is a distortion of 78%! Note that though R^2 is commonly used as an effect size index, some researchers have argued that its use is inappropriate (D'Andrade & Dart, 1990; Ozer, 1985).

The results recorded in the table also enable us to identify the cause of the distortion. The inflation of the correlation coefficient ($r = COV(X,Y) / SD(X) * SD(Y)$) is due to the denominator of the correlation coefficient decreasing as a combined function of the number of raters and the average inter-rater reliability. The decrease in the denominator of the correlation is due to the standard deviation of the average ratings of the test items decreasing as a combined function of the number of raters and the average inter-rater reliability. In other words, assuming that the average inter-rater reliability is not 1, the average ratings for each test item will converge to some degree as the number of raters increases. The degree of convergence will be dependent upon the average inter-rater reliability, in that the lower the average inter-rater reliability the greater the convergence.

One method Brand et al. (2010) suggested for avoiding effect size inflation due to averaging in a multiple trial experiments is simply to report the unstandardized effect size (i.e., the mean difference) instead of the standardized effect size (i.e., Cohen's d). This is because averaging in a multiple trial experiment inflates the standardized effect size but does not inflate the unstandardized effect size. Similarly, Baguley (2009) suggests reporting the unstandardized effect size (i.e., unstandardized slope of the linear regression line) as opposed to the standardized effect size (i.e., the correlation estimate) when the range of a measure is restricted in a correlational study. However, adopting this strategy will not work for the correlational study we have simulated. This is because the unstandardized effect size, which is the unstandardized slope of the linear regression line ($b = COV(XY) / SD(X)^2$) is also inflated when an average ratings are used. Since, the denominator of the unstandardized slope of the linear regression line, which is based on the average ratings of the test item, also decreases when the number of raters increases and the average inter-rater reliability is less than one.

TABLE 1. Results From the Monte-Carlo Simulation

	True Effect (r) = 0.10 Raters in the Experiment					True Effect (r) = 0.30 Raters in the Experiment					True Effect (r) = 0.50 Raters in the Experiment				
	1	5	10	20	30	1	5	10	20	30	1	5	10	20	30
Inter-Rater Reliability = 0.30															
Mean SD(x)	1.66	1.17	1.09	1.02	1	1.52	1.12	1.01	1	1.04	2.02	1.19	1.04	1	0.99
Mean SD(y)	92.92	97.83	97.1	100.08	99.47	97.5	100.35	99.78	97.9	95.22	104.03	99.55	98.28	100.45	99.38
Mean COV(x,y)	3.43	3.44	3.44	3.44	3.44	10.33	10.32	10.32	10.31	10.3	17.22	17.2	17.23	17.15	17.17
Mean SD(x)*SD(y)	34.3	22.85	20.92	19.91	19.58	34.42	22.84	20.94	19.9	19.52	34.44	22.84	20.97	19.87	19.54
Mean r	0.1	0.15	0.16	0.17	0.18	0.3	0.45	0.49	0.52	0.53	0.5	0.75	0.82	0.86	0.88
% Distortion of r	0	51	64	73	76	0	51	64	73	76	0	51	64	73	76
Inter-Rater Reliability = 0.50															
Mean SD(x)	2.02	1.44	1.4	1.32	1.26	1.9	1.42	1.36	1.27	1.28	1.7	1.27	1.26	1.25	1.27
Mean SD(y)	97.78	94.53	97.78	101.47	107.7	100.17	101.38	98.15	103.5	97.42	100.75	97.45	101.08	100.12	97.6
Mean COV(x,y)	3.44	3.44	3.43	3.44	3.44	10.3	10.31	10.3	10.31	10.31	17.18	17.19	17.19	17.2	17.19
Mean SD(x)*SD(y)	34.39	26.64	25.48	24.9	24.7	34.34	26.65	25.47	24.91	24.72	34.36	26.64	25.51	24.93	24.72
Mean r	0.1	0.13	0.13	0.14	0.14	0.3	0.39	0.4	0.41	0.42	0.5	0.65	0.67	0.69	0.7
% Distortion of r	0	29	35	38	39	0	29	35	38	39	0	29	35	38	39
Inter-Rater Reliability = 0.80															
Mean SD(x)	1.82	1.58	1.64	1.52	1.54	1.85	1.51	1.54	1.6	1.6	1.62	1.57	1.64	1.53	1.58
Mean SD(y)	102.75	101.1	103.72	104.47	98.67	97.03	105.05	98.47	105.17	98.28	99.08	98.85	100	105.08	100.03
Mean COV(x,y)	3.43	3.44	3.44	3.44	3.44	10.3	10.31	10.32	10.33	10.31	17.21	17.18	17.19	17.19	17.22
Mean SD(x)*SD(y)	34.31	31.52	31.13	30.96	30.86	34.34	31.49	31.14	30.98	30.88	34.41	31.5	31.14	30.94	30.93
Mean r	0.1	0.11	0.11	0.11	0.11	0.3	0.33	0.33	0.33	0.33	0.5	0.55	0.55	0.56	0.56
% Distortion of r	0	9	10	11	11	0	9	10	11	11	0	9	10	11	11

Fortunately, there is a simple and easy method to avoid inflating the correlation coefficient when the correlational study involves multiple raters: researchers should just calculate the correlation of interest for each rater and then compute the average of these correlations instead of calculating the average rating across raters and then computing the correlation of interest based on the average ratings.

To extend these findings to actual examples from the literature we re-analyzed raw data from two previous published studies. These studies had adopted a correlation design with an averaged based measure. Bearing in mind that it is conceivable that the average inter-rater reliability could be high (i.e., .80 or higher) and consequently the inflation in the correlation coefficient due to using an averaged based measure will be relatively minor, this approach determines whether inflation has occurred and provides an estimate of the approximate extent of the inflation.

Case Study 1: Carré et al. (2009)

In the study by Carré and colleagues (2009), 31 participants rated 24 faces for perceived aggression using a 1 to 7 Likert scale. They then correlated the mean aggression ratings (across raters) for each face with the width-to-height ratio (WHR) reporting $r = .70$. Cronbach's alpha for perceived aggression ratings was encouragingly high at .95. Cronbach's alpha, however, is not an adequate index of uni-dimensonality (see, Cortina, 1993; Schmidt, 1996; Shevlin, Miles, Davies & Walker, 2000). Hence, a high Cronbach's alpha does not guarantee that the average inter-rater correlation will also be high, and this was in fact the case. The average inter-rater correlation for the perceived aggression rating data was .40.

To re-analyze the raw data, we used R (R Development Core Team, 2011, http://www.r-project.org/) to compute the correlation between the average aggression rating and the face WHR as the number of raters increase for 10,000 randomly generated sequences of raters. Basing our re-analysis on 10,000 randomly generated sequences of raters minimizes the effect that the order in which an individual rater's data is entered on the outcome of the analysis. Figure 1 shows the mean magnitude of the correlation between the average aggression rating and WHR as a function of the number of raters. The correlation estimate is 55% greater when there are 31 raters as opposed to one rater. We do not, of course, know the true size of the correlation, as we did in the Monte-Carlo simulation, but we do know, from the results of the simulation, that the mean correlation based on only one rater will provide the best (least biased) estimate of the true correlation from the available data. Hence the observed discrepancy between 1 rater and 31 raters is a good indication of the inflation of the correlation coefficient. In this case with 55% inflation of the estimate, the distortion is substantial.

It is interesting to note that Carré and colleagues (2009) did actually report the mean correlation ($r = .38$) between perceived aggression and WHR (i.e., they calculate the correlation estimate for each rater and then compute the mean

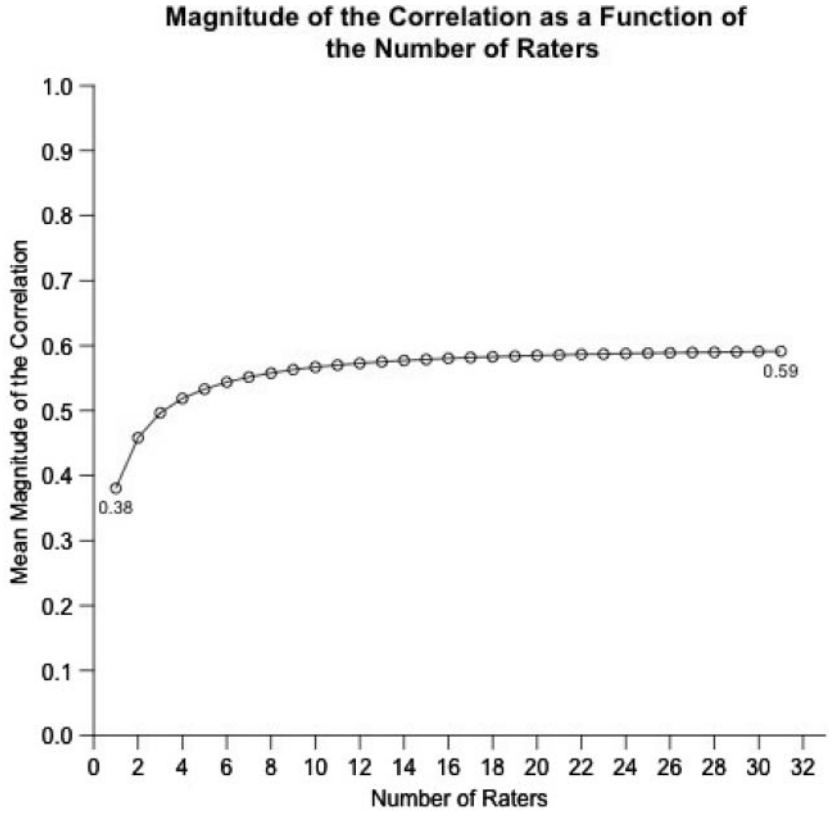

FIGURE 1. Re-analysis of Data from Carré et al. (2009).

average of these correlations) but not in the main text of the article. It was reported peripherally in text accompanying a figure.

Case Study 2: Hodgetts et al. (2009)

In Hodgetts et al. (2009) study, 42 raters rated the similarity of object sets using a 1 to 6 Likert scale. They then correlated the mean similarity rating (across raters) for each object set with the number of transformations required to transform the base pair of stimuli to a target pair of stimuli. Unlike, Carré and colleagues (2009), they did not calculate and report the Cronbach's alpha for the similarity ratings. We calculated the average inter-rater correlation for the similarity ratings and found it to be relatively low, .42. Out of curiosity, we also calculated the

Cronbach's alpha to see whether there is a large discrepancy between average inter-rater correlation and Cronbach's alpha and indeed it was large, Cronbach's alpha = .97.

We used the same procedure as we used to re-analyze the data for Carré and colleagues (2009). We computed the correlation between the average similarity rating and the number of transformations as the number of raters increase for 10,000 different sequences of raters. Figure 2 shows the mean magnitude of the correlation between the average similarity rating and the number of transformations as a function of the number of raters. The unbiased correlation estimate was 0.56. Hence, the correlation estimate is 52% greater when there are 42 raters as opposed to one rater. Hence, the inflation of the correlation coefficient due to using an average-based measure is again substantial. Furthermore, the R^2 effect size index

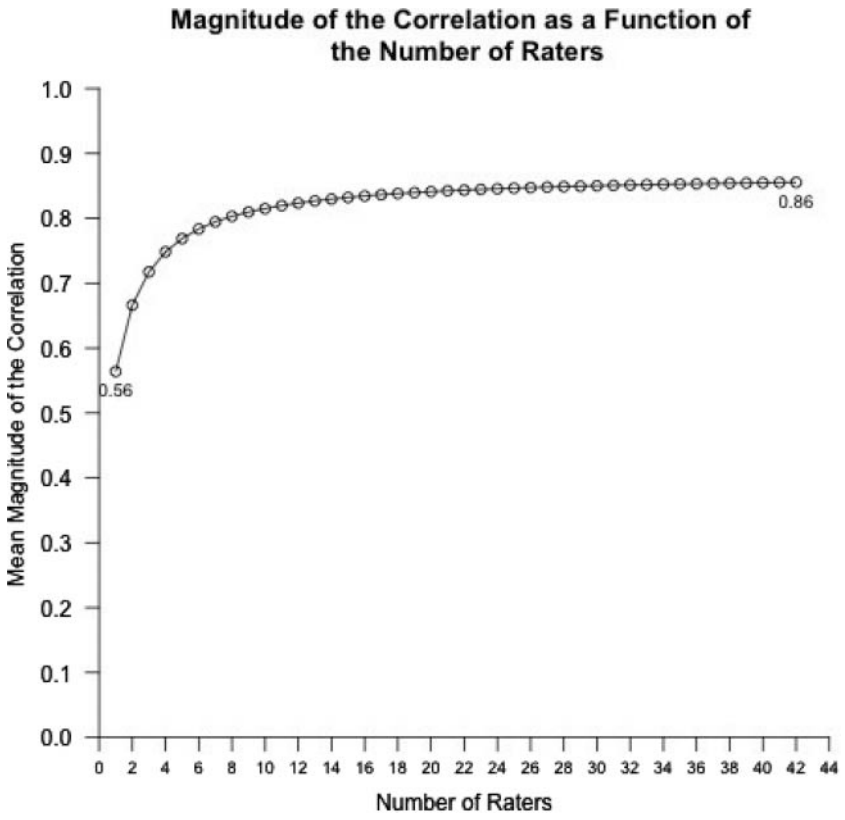

FIGURE 2. Re-analysis of data from Hodgetts et al. (2009).

reported by Hodgetts and colleagues (2009) will be distorted by approximately 130%!

Conclusions

The results from the Monte-Carlo simulation reflect a similar pattern of results as obtained by Brand and colleagues (2011). The results from the simulation essentially show if the average inter-rater correlation is low, the correlation estimate that is based on an average rating from multiple raters will be inflated. Moreover, the inflation of the correlation estimate can be substantial and the inflation of the R^2 effect size index even more so. Our re-analysis of data from previously published studies is in concordance with Monin and Oppenheimer (2004) and shows that sizeable inflation of the correlation estimate will occur in studies adopting a correlation design with a measure based on average ratings. It is also important to note from our re-analyses that the inflation in the correlation estimate can even occur when the Cronbach's Alpha for the ratings is very high. This is because Cronbach's Alpha is not an appropriate measure of average inter-rater reliability, and as the Monte-Carlo simulation demonstrated, true average inter-rater reliability is the prime factor influencing the inflation of the correlation estimate.

As we have mentioned above, the inflation of the correlation estimate can be easily avoided. Researchers should calculate the correlation of interest for each rater and then compute the average of these correlations instead of calculating the average rating across raters and then computing the correlation. It is worth noting that the unstandardized effect size, the unstandardized slope of the linear regression line (b), is also inflated when the predictor variable (X) is based on an average. Hence, researchers must be cautious when interpreting both the unstandardized effect size (i.e., the unstandardized slope of the linear regression line) and the standardized effect size (i.e., the correlation coefficient) when the predictor variable (X) is based on an average. Ideally, researchers should obtain the raw data from the original study and calculate the desired effect size (either standardized or unstandardized) for each rater and then compute the average of these effect sizes. Overall, we suggest great caution in accepting a positive position towards correlations from averaged data since the size of the correlation is dependent on the number of ratings averaged and the inter-rater reliability.

From a classical test theory perspective, increasing the number of raters will enhance the reliability of the averaged-based measure, and as a result the correlation involving an averaged-based measure will become increasingly more accurate. This is because the averaged measure is comprised of a true component and an error component. That error component that consists of random responses, transient, and specific factor errors (see, Schmidt, Le & Ilis 2003) will diminish in impact as the number of raters increase. However, we believe the classical test

theory interpretation that the averaged-based measure will yield a more accurate as opposed to an inflated estimation of the correlation as problematic, since it depends on what the researcher wishes to describe. In the typical application of classical test theory, a test consists of a set of test items, and the individual test score is the aggregate of the individual's test item responses, and as the number of test items increases, the reliability of the individual's test score will also increase.

However, in the particular cases we have documented, the researchers are treating individual raters as analogous to test items. As such, a measure based on the average of the individual ratings is a measure specific to that group of raters, and therefore a correlation involving the average-based measure will also be specific to that group of raters. But researchers are typically only interested in the correlation between two measures at an individual level and not at an arbitrary defined group level. For example, Carré and colleagues (2009) are interested in the relationship between the individual's rating of a face's aggression and the face's width-to-height ratio and not the relationship between a group's rating of a face's aggression and the face's width to height ratio. Researchers are therefore making a cross-level inference if the correlation they calculate involves an averaged based measure derived from a group of raters and draw conclusions concerning the correlation at the individual level and not at the group level. Or in other words, the researchers are organizing the data (i.e., grouping and averaging the ratings from individuals) in one way but the conclusion they draw from their analysis of the data assumes that the data is organized in a different way (Nickerson, 1995). Even if a cross-level inference is not made, a correlation based on an averaged based measure has the undesirable property of varying as a function the number of raters and the average inter-rater reliability, consequently replication and interpretation of a correlation based on an averaged based measure will be problematic unless the number of raters and reliability are equivalent.

In summary, we have extended the valuable observation made by Monin and Oppenheimer (2004). Through simulation methods we have highlighted the circumstances under which correlation estimates (i.e., the standardized effect sizes) can be markedly inflated when an average-based measure is employed, and we have provided a range of inflated values from varying inter-rater reliability and rater number. The size of the inflations we have simulated and have observed in two published studies warrants particular concern, since the examples we used were published after Monin and Oppenheimer (2004). Hopefully, this article directly mapping out the parameters of the problem rather than presenting an instance of it will make researchers aware that correlations based on the average produced by multiple raters can be markedly inflated. With this information, it is hoped that in future studies researchers will analyze their data appropriately and with studies currently available interpret the findings of research correlating average-based measures with caution.

NOTE

1. Anonymous reviewers have pointed out that the results from the Monte-Carlo simulation can also be derived analytically by applying the Spearman–Brown prophecy formula (Spearman, 1910): $R(n) = R(1)^*n/[1+(n-1)^*R(1)]$, where $R(n)$ = the reliability of the mean of n raters, n = number of raters and $R(1)$ = the average inter-item correlation and the disattenuation formula (Spearman, 1904): $R_{xy} = r_{xy} / SQRT(r_{xx}{}^*r_{yy})$, where R_{xy} = the true correlation between x and y, r_{xy} = observed correlation, r_{xx} = reliability of x and r_{yy} = reliability of y.

AUTHOR NOTES

Andrew Brand is a software developer for the Department of Psychology at King's College London. He is also the creator of iPsychExpts (www.ipsychexpts.com), a website that encourages and promotes the use of web experiments for conducting psychological research. **Michael T. Bradley** teaches cognition and social psychology at the University of New Brunswick. He has conducted research on information and memory tests with the polygraph and has a long-standing interest in effect size and power issues.

REFERENCES

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

Baguley, T. (2010). When correlations go bad. *The Psychologist*, *23*, 122–123.

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.

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*, 1–11.

Carre, J., McCormick, C., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behaviour. *Psychological Science*, *20*, 1994–1998.

Coetzee, V., Chen, J., Perrett, D. I., & Stephen, I. D. (2010). Deciphering faces: Quantifiable visual cues to weight. *Perception*, *39*, 51–61.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences (2nd Edition)*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cortina. J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104.

D'Andrade, R. G. and Dart, J. (1990). The interpretation of r versus r^2 or why percent of variance is a poor measure of size of effect. *Journal of Quantitative Anthropology*, *2*, 47–59.

Feinberg, D. R., DeBruine, L. M., Jones, B. C., & Perrett, D. I. (2008). The role of femininity and averageness of voice pitch in aesthetic judgments of women's voices. *Perception*, *37*, 615–623.

Fiedler, K. (2011). Voodoo correlations are everywhere—not only in neuroscience. *Perspectives on Psychological Science*, *6*, 163–171.

Hodgetts, C. J., Hahn, U., & Chater, N. (2009) Transformation and alignment in similarity. *Cognition, 113*, 62–79.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*, 107–112.

Monin, B. (2003). The warm glow heuristic: When liking leads to familiarity. *Journal of Personality and Social Psychology*, *85*, 1035–1048.

Monin, B., & Oppenheimer, D. (2005). Correlated averages vs. average correlations: Demonstrating the warm glow heuristic beyond aggregation. *Social Cognition*, *23*, 257–278.

Nickerson, C. A. E. (1995). Does willingness to pay reflect the purchase of moral satisfaction? A reconsideration of Kahneman and Knetsch. *Journal of Environmental Economics and Management*, *28*, 126–133.

Ozer, D. J. (1985). Correlation and the Coefficient of Determination. *Psychological Bulletin*, *97*, 307–315.

Preacher, K. J., Rucker, D. D., MacCallum, R. C. & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, *10*, 178–192.

R Development Core Team (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.r-project.org.

Rule, N. O., & Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological Science*, *19*, 109–111.

Savani, K., Stephens, N. M., & Markus, H. R. (2011). The unanticipated interpersonal and societal consequences of choice victim blaming and reduced support for the public good. *Psychological Science*, *22*, 795–802.

Schmidt, F. L., Le, H., & Ilis, R. (2003). Beyond alpha: An experimental examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods*, *8*, 206–224.

Schmidt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350–353.

Shevlin, M., Miles, J. N. V., Davies, M. N. O., & Walker, S. 2000). Coefficient alpha: A useful indicator of reliability? *Personality and Individual Differences*, *28*, 229–237.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.

Vul, E., Harris C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI Studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

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

Original manuscript received February 9, 2012
Final version accepted June 12, 2012