ARE LARGER EFFECT SIZES IN EXPERIMENTAL STUDIES GOOD PREDICTORS OF HIGHER CITATION RATES? A BAYESIAN EXAMINATION.

Jesper W. Schneider[1] and Dorte Henriksen[2]

[1] *jws@cfa.au.dk*
Danish Centre for Studies in Research and Research Policy, Department of Political Science and Government, Aarhus University, Bartholins Álle 7, DK-8000 Aarhus C (Denmark)

[2] *dh@cfa.au.dk*
Danish Centre for Studies in Research and Research Policy, Department of Political Science and Government, Aarhus University, Bartholins Álle 7, DK-8000 Aarhus C (Denmark)

Abstract
Effect sizes are perhaps the most important quantitative information in statistical inferential studies. Recently, the hypothesis that rational citation behaviour in general ought to give credit to studies that successfully apply a treatment and detect greater effects, resulting in such studies being cited more frequently among comparable studies. Hence, it is predicted that larger effect sizes increases study relative citation rates.
Two recent studies in biology provide contradictory results on this hypothesis. The present study investigates the same hypothesis but in different research areas and with a more credible model selection procedure.
Using meta-analyses, we identify comparable individual experimental studies (n=259) from five different research specialties. Effect sizes are compared to the citation rates of the individual studies and impact factors for the journals where the studies are published. Contrary to the previous findings, and in fact most studies in scientometrics, we examine the hypothesis with a Bayesian model selection procedure. This is advantageous, as we thereby are able to quantify the statistical evidence for both hypotheses, H_0 and H_1. This is not possible in classical statistical inference, though the implicit inferential decision made by most researchers when they fail to reject H_0 is to accept it. This is a flawed logic. Given uniform priors for the two hypotheses, the result from the present data set is posterior odds of 13/4 to 1 in favor of the null models examined. Consequently, the study give positive evidence to the claim made by Lortie et al. (forthcoming) that effect sizes do not predict citation rates and are as such poor proxies for the quantitative merit of a given experimental treatment.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2) and Sociological and Philosophical Issues and Applications (Topic 13)

Introduction

In a forthcoming study, Lortie et al. (doi: 10.1007/s11192-012-0822-6) hypothesize that if citation behavior is supposed to be rationale, articles that report larger biological effect sizes from successful treatments in ecology and evolutionary biology studies, should generally also have higher relative citation rates. The hypothesis is apparently not supported by their data and the conclusion is that citations are a poor proxy for quantitative merit of a given treatment in ecology and evolutionary biology. A similar hypothesis, also from a sample of ecology studies, is investigated by Barto and Rillig (2012). Contrary, to Lortie et al., Barto and Rillig (2012) do identify a positive relationship between effect size and citation rates. The importance of effect size in reference behavior was also previously indicated in a survey by Shadish et al. (1995). The hypothesis is interesting, though not without problems, and warrants replication for other research areas. In this study, we examine the hypothesis in five different research specialties (in psychiatry, clinical psychology, brain research, psychotherapy and educational research) in order to further examine if and how the magnitude of effect sizes and citation rates are related.

The general hypothesis has some merit. It seems reasonable to assume that rational reference behavior in quantitative experimental domains would entail that in specialized research areas, studies that, *ceteris paribus*, demonstrate larger effect sizes will also generally be more cited. The notion of empirical science being cumulative warrants such an assumption. Also, higher impact journals should on average publish studies with larger effect sizes if they do indeed differentiate for stronger evidence (Song, Eastwood & Gilbody, 2000). At face value, effect sizes are very important in quantitative studies that rely on statistical inference. It is well known that statistical significance tests are flawed, seriously misused and misinterpreted (e.g., Berkson, 1942; Oakes, 1986; Cohen, 1994; Nickerson, 2000; Kline, 2004; and for a scientometric perspective Schneider, 2013). *P* values cannot quantity the importance of a result, but effect sizes with confidence limits can (e.g., Goodman, 1999a; Goodman, 2008; Ellis, 2010; Cumming, 2012). Reporting effect sizes are also important for meta-analytic purposes. The latter basically serves as a formal tool of evidence, where effect sizes from comparable studies are evaluated statistically. Notice, the latter is certainly not without its problems (Berk & Freedman, 2003; Berk, 2007).

However, a straightforward relation between the magnitudes of effect sizes and reference behavior is doubtful. The question is whether effect sizes alone are sufficient to warrant a reference. For example, often large effect sizes (relatively to the phenomenon studied) are reported in the earliest studies within a domain (Barto & Rillig, 2012). Often such findings cannot be replicated and the subsequent effect sizes become more moderate. Also, samples size and quality of the study design are crucial elements in relation to effect sizes and their reference potential. Large effect sizes from a non-experimental study with a relative small sample size are generally considered less robust and causally inferior and thus

have *de facto* lower evidence. Consequently, other rational epistemic factors may be of more importance to the citing author when he or she decides to reference an experimental study. Clearly, references are given (or not given) for a whole number of reasons, some rational and sound, others haphazardly or perfunctory, and still others suspicious, self-promoting and political, and citations are perceived differently among researchers (for overviews see for example Bornmann & Daniel, 2008; Aksnes & Rip, 2009). At the same time, numerous studies have tried to identify citation predictors for articles in restricted settings (van Dalen & Henkens, 2005; Stremersch, Verniers & Verhoef, 2007; Mingers & Xu, 2010). Common for many of these studies is that their model specification and subsequent fitting procedure is done on the same data set. Further, a preponderance of the proxy variables specified and tested seems to be easily quantifiable document attributes from the bibliographic records retrieved from a citation database. Hence, what we are left with is an ordinal knowledge base about the potential influence - on average - upon citations to articles from indicators such as journal status, document type, number of authors and similar proxies. The meaning and validity of the proxies are seldom discussed. The influence of more cognitive aspects of documents, i.e., the content that ought to stimulate reactions from peers, whether positive, neutral or negative, is not well established quantitatively. Clearly more effort is needed to analyze cognitive and epistemic patterns relating to reference behaviors. In that respect, effect sizes in quantitative experimental studies are interesting. The aim of experimental studies is to investigate treatment effects and the most important quantitative entity when reporting the results is the estimated effect size (standardized or non-standardized) and its margin of error.

Consequently, this study further examines the hypothesis that somehow effect sizes ought to influence citation rates and show a relation to journal impact factors. Contrary to other studies, we take a Bayesian approach, where we provide statistical evidence for the hypotheses investigated. The next section explains the methods and materials used, including our Bayesian perspective; subsequently we report on our results, and end with a discussion of the results.

Methods and materials

We basically follow the same data collection strategy as Lortie et al. Our aim in this study is to explore whether Lortie et al.'s claims are discernible in other domains, or alternatively, to find support for the claims by Barto and Rillig (2012). Hence, we have not set-up a strict data collection procedure for a specific domain. An initial search was conducted in Thompson Reuters' *Web of Science*© (WoS) with various forms of the term 'meta-analysis'. The result was restricted to meta-analyses published from 2003-2012. From the large set of meta-analyses identified (≈ 26.000), five was chosen based on the following inclusion criteria 1) a random selection procedure selected five different WoS subject categories; 2) 25 meta-analyses were randomly selected within each of the categories; 3) these meta-analyses were scanned to see if they reported individual standardized effect

sizes as well as sample sizes for the studies analyzed. Among those meta-analyses eligible, only the ones where all studies analyzed were experiments with random procedures was selected in order to have some control of the study design quality. Among these, one meta-analysis was chosen randomly for each of the five categories resulting in 259 individual effect sizes (i.e., the five selected meta-analyses are Willcutt et al., 2005; Sommer et al., 2008; Beck et al., 2012; Furtak et al., 2012; Oldham et al., 2012). Effect sizes from the individual studies (e.g., Glass'Δ, Hedges'*g* and Pearson's *r*) were transformed to one scale Cohen's *d*. In studies where multi-effect sizes were reported only the largest reported effect size was included, effectively favoring the hypothesis investigated. Citation statistics for each of the 259 studies were obtained from WoS, as well as 5-year Impact Factors (JIF) from journals where the studies were published. JIFs were calculated so that they matched the years immediately after publication of the study.

Though random elements are used in the selection process, the overall sampling frame cannot be considered a probability sample. However, the sampling frame ensures that we can analyze fairly homogenous studies across domains. Firstly, the choice of meta-analyses as pointers to individual studies ensures that we identify a restricted set of articles that presumably study the same phenomenon often with similar approaches (i.e., the meta-analysis has already enforced strict inclusion criteria); secondly, reporting of standardized effect sizes entail a common scale so that comparison of effect sizes across domains is possible. The requirement that sample sizes should be reported (i.e., not just shown as confidence limits) enable us to control for sample size when predicting the potential influence of effect sizes on citation rates (i.e., large sample size, *ceteris paribus*, produce more stable effect sizes).

Contrary to Lortie et al., as data are continuous, we apply simple OLS as our primary models to explore the hypothesized positive linear trend between effect sizes and citation rates of individual studies, as well as impact factors at the journal level, i.e., larger effect sizes tend to be published in higher impact journals (individual article citation rates and 5-year JIFs are log-transformed). To mimic Lortie et al., we also specified Poisson models. It may be reasonable to model citation rates as counts, despite the fact that data are continuous, since $y \geq 0$ (e.g., Wooldridge, 2002). Even so, the GLM models provide the same interpretations as the logged-*y* OLS-models, but with less convincing diagnostics.

The individual studies (n = 259) are collapsed into one sample, as sensitivity analyses revealed no discernable effects relevant for this study. For example, field normalization with logarithm-based citation z-scores (Lundberg, 2007) does not alter the pattern of relations compared to simple mean annual citation scores when all studies are collapsed. Likewise, there was no discernible difference when using mean annual citation scores for all years versus a 5-year period after publication. What is of importance is whether higher effect sizes tend to

influence citation rates, minor differences in general citation activity between domains does not affect this aim.

Bayesian hypothesis testing

Contrary to most studies in scientometrics and the social sciences, we take a Bayesian approach to statistical evidence. Inference by *p* values, in the frequentist amalgam "null hypothesis significance testing" (NHST), is nearly ubiquitous despite longstanding serious criticisms concerning its logical flaws, rote use, misunderstandings and misuses (see some good introductory references in the introduction section out of literally hundreds). Two critical issues are important for this study. NHST does not allow researchers to state evidence *for* the null hypothesis (e.g., Hubbard & Lindsay, 2008), nevertheless, this is more or less the implicit inferential decision made by most researchers when they fail to reject H_0; as an example, Lortie et al., base their claims of no effect on the failure to reject H_0. Further, it has been clearly demonstrated that *p* values themselves overstate the evidence against the null hypothesis, i.e., a rejection of H_0, especially in the *p*-interval from .05 to .01 (Jeffreys, 1961; Berger & Sellke, 1987; Goodman, 1999b; Sellke, Bayarri & Berger, 2001).

Bayes factors (Jeffreys, 1961; Kass & Raftery, 1995) have been advocated as superior to *p* values for assessing statistical evidence in data (Edwards, Lindman & Savage, 1963; Raftery, 1995; Wagenmakers, 2007; Rouder et al., 2009). We entirely concur with this claim. The Bayes factor computes the probability of the observed data under H_0 *vis-a-vis* H_1. Notice, in contrast to the frequentist *p* value, the Bayes factor allows researchers to quantify evidence in favor of H_0. In the Bayesian model selection procedure, the ultimate objective is to compute a probability reflecting which model is more likely to be correct, on the basis of the obtained data and the core concept is Bayes' theorem.

We use Bayes factors as the model selection procedure in this study. The two models examined are H_1, that effect sizes predict citation rates, against H_0 of no or a minuscule relation. Although the Bayes factor is conceptually straightforward, its use is not widespread in the social sciences. Bayesian models require specification of priors. Like, NHST it is uncomplicated to calculate $p(D|H_0)$, however, H_1 does not specify one particular a priori value for the effect in question. Rather, H_1 is associated with a distribution of possible effect sizes, and the value of the Bayes factor depends on the nature of that distribution. Therefore, exact computation of the Bayes factor quickly becomes complex, involving integration over the space of possible effect sizes using procedures such as Markov chain Monte Carlo methods. This is complicated and no general commercial software package enables Bayesian modeling.

In this study we apply a more practical alternative suggested by Raftery (1995) and Wagenmakers (2007), where we approximate the Bayes factor using the Bayesian Information Criterion (BIC) (Raftery, 1995). BIC is often used to quantify the goodness of fit of a model to data, accounting for the number of free

parameters in the model. BIC is easy to compute and for some models, popular statistical computer programs already provide the raw BIC numbers, so that in order to perform an approximate Bayesian hypothesis test, one only needs to transform BIC values for two competing models, H_0 and H_1, to posterior probabilities (for details, see e.g., Raftery, 1995; Glover & Dixon, 2004; Wagenmakers, 2007).

Some assumptions and limitations are in order. Obviously, the Bayes factor is sensitive to the shape of the prior distribution, but the use of BIC does not require the researcher to specify his or her own prior distribution. This is appealing, but also the main drawback of using BIC. BIC implicitly assumes the *unit information prior* (Kass & Wasserman, 1995) and it has been argued that this prior is too wide, resulting in a decrease of the prior predictive probability of H_1, and therefore makes H_0 appear more plausible than it actually is. In this sense, the BIC estimate of the Bayesian posterior probabilities should be considered somewhat conservative with respect to providing evidence for the alternative hypothesis (Raftery, 1999). Thus, the drawback of the BIC is that it does not incorporate substantive information into its implicit prior distribution; the virtue of the BIC is that the specification of the prior distribution is completely automatic. Another limitation of BIC is that its approximation ignores the functional form of the model parameters, focusing exclusively on the number of free parameters. A full-blown Bayesian analysis is sensitive to the functional form of the parameters because it averages the likelihood across the entire parameter space. Although the issue of functional form is important, it is much more important in complicated nonlinear models than it is in standard linear statistical models.

The Bayes factor plays a crucial role in establishing the relative evidential support for H_0 and H_1. The Bayes factor (BF) can be estimated using the following transformation of the difference in BIC values for two competing models:

$$\mathbf{BF} \approx \frac{\boldsymbol{p}\mathbf{BIC}(\mathbf{D}|\mathbf{H_0})}{\boldsymbol{p}\mathbf{BIC}(\mathbf{D}|\mathbf{H_1})} = \boldsymbol{e}^{(\Delta \mathbf{BIC}/\mathbf{2})} \quad (1)$$

where $\Delta BIC = BIC(H_1) - BIC(H_0)$. The resulting estimate of the Bayes factor yields the odds favoring the null hypothesis, relative to the alternative hypothesis. BF can then be converted to the posterior probability that the data favor the null hypothesis as follows (assuming equal priors):

$$p\mathrm{BIC}(H_0|D) = \frac{\mathbf{BF}}{\mathbf{BF+1}} \quad (2)$$

With only two competing models, the posterior probability that the data favor the alternative hypothesis is just the complement of Equation 2:

$$p\mathrm{BIC}(H_1|D) = 1 - p\mathrm{BIC}(H_0|D) \quad (3)$$

In the present study we use total sum of squares and the sum of squares of the error term to derive $BIC(H_1)$ and $BIC(H_0)$ (e.g., Wagenmakers, 2007). Finally,

based on Jeffreys' (1961) rules of thumb for interpreting Bayes factors, Raftery (1995) has provided descriptive terms for strength of evidence as follows: pBIC(H_i|D) .50-.75 (weak), .75-.95 (positive), .95-.99 (strong), and >.99 (very strong).

Results

First we present some figures that explore the data set. Figure 1 below shows the distribution of standardized mean effect sizes for the individual studies in the five meta-analyses (MA1, MA2, MA3, MA4 and MA5).

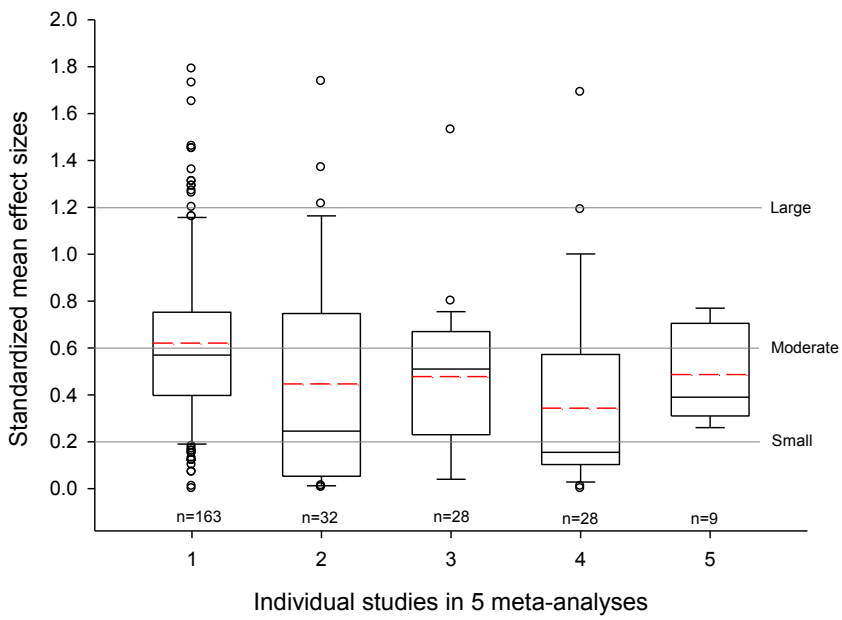

Figure 1. Box plot of standardized mean effect sizes in individual studies reported in 5 meta-analyses (n=259). Solid lines in boxes show median effect sizes and dotted lines average effect sizes for studies included in the meta-analyses.

If we apply Cohen's reference categories for interpreting effect sizes (Cohen, 1988), we can see that median effect sizes for all but one meta-analysis (MA4) can be considered small, whereas MA4's is trivial. We also see that MA1, with its large n come closest to a Gaussian distribution, whereas the other meta-analyses show considerable skewness. Three meta-analyses have rather long whiskers (at the high end) and four meta-analyses have outliers, which corresponds to large effect sizes.

Figure 2 below shows the distribution of citation scores for the five meta-analyses. MA1 and MA4 have the largest median "mean annual citation scores", 6.5 and 5.1 respectively. The other three meta-analyses have considerable lower citation activity. All distributions are skewed, but skewness for MA1 is considerably lower than the others (except MA5, but *n* here is only 9 and scarcely robust). We see some marked outliers in MA2 and MA4; the outlier in MA4 is an article published in *Nature* with an annual mean citation score of 38.2.

Figure 3 below is a plot of mean annual citation scores for individual studies, as well as journal impact factors, as a function of the magnitude of effect sizes. It is clear that the concentration of observations is in the reference categories trivial (*n*=52), small (*n*=98) and medium (*n*=91). The outlier on the border to the large category is the *Nature* article (high annual citation scores and obviously a high JIF).

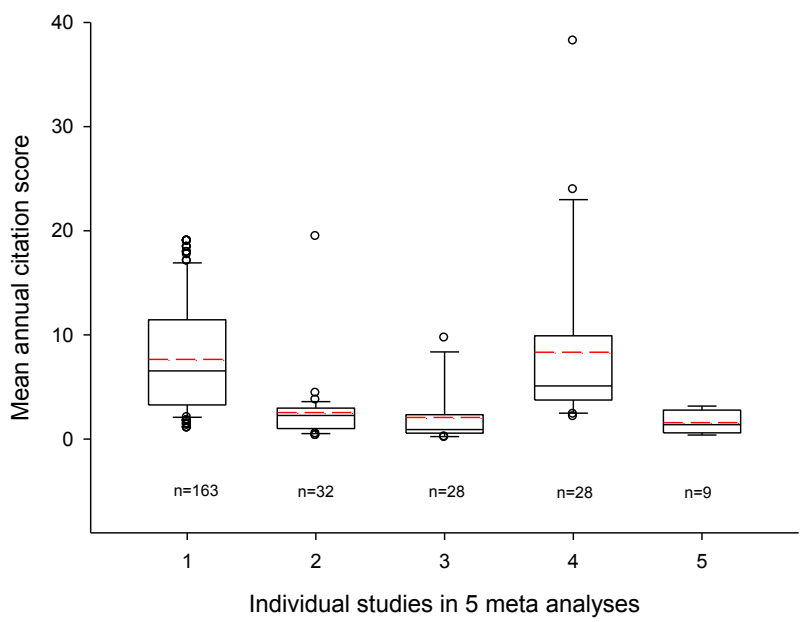

Figure 2. Box plot of mean annual citation scores for individual studies reported in 5 meta-analyses (*n*=259). Solid lines in boxes show median citation scores and dotted lines average citation scores for the studies included in the meta-analysis.

Finally, we group the individual experimental studies according to their reference category as defined by Cohen (1988) and plot this against mean annual citation scores as illustrated in Figure 4 above. This box plot reveals almost identical central tendencies in citation activity across the four reference groups. If the

predicted hypothesis of a linear trend was true, then the boxes should be staggered so that the trivial box was at the bottom, followed by the small and medium boxes, ending with the large box at the top. Clearly this is not so, hence we can expect support for the null hypothesis.

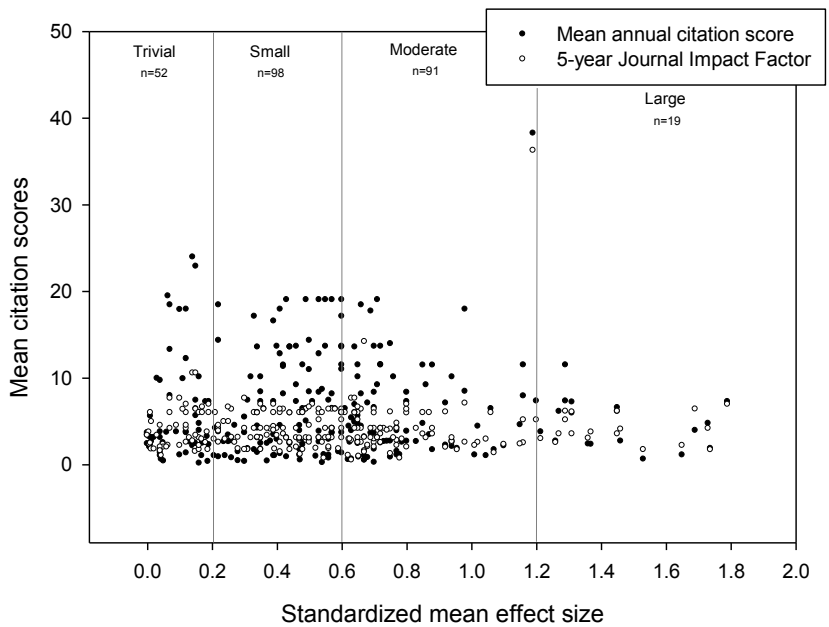

Figure 3. Plot of mean citation scores (mean annual citation scores for individual studies and journal impact factors) as a function effect sizes (N = 259). Vertical grey lines show Cohen's reference categories for interpretation of effect sizes and *n* indicate the number of studies in each reference.

Table 1. Scatter matrix of relationships between effect size, mean citation score, journal impact factor and sample size.

	Mean annual citation Score	5-year journal impact factor	Sample size
Effect size	-0.001	0.072	-0.296
Mean annual citation Score		0.456	0.253
5-year journal impact factor			0.065

In Table 1 below, we report Pearson correlation coefficients between effect sizes, mean annual citation scores, 5-year journal impact factors and sample sizes; rank correlations give similar correlations.
As one would suspect from inspecting Figure 3 and 4, there are close to no linear relation between effect sizes and citation rates for individual articles or JIFs from the journals where these articles are published. However, there is a negative relation between effect size and sample size. The relation is moderate and not surprising. To some extent larger sample sizes in studies result in relatively lower effect sizes. Large sample sizes reduce variability and thus provide more stable effect sizes. Citation rates and JIFs are also moderately correlated, though this is uninteresting in this context. What is more important is that citation rates and sample size have a small correlation. This may indicate that to some vague degree citing authors are aware of the importance of sample size for the robustness of results.

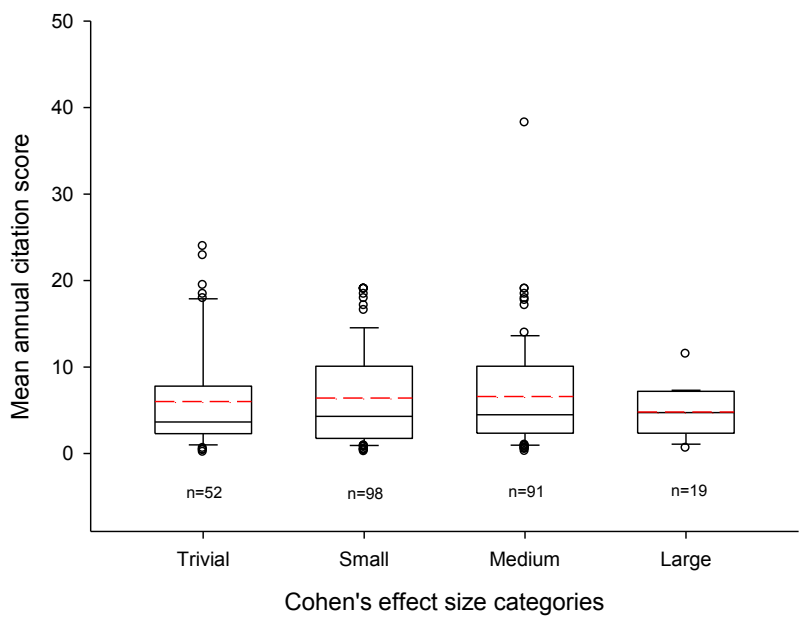

Figure 4. Box plot of mean annual citation scores distributed according to Cohen's reference categories for interpretation of effect sizes. Solid lines in boxes show median citation scores and dotted lines average citation scores for the studies included in the reference category.

Like Lortie et al., we use a simple model with one predictor (effect size). However, unlike Lortie et al. we apply Bayes factors to assess the evidence for the two competing hypotheses. Unlike *p* values and NHST, we are therefore able to

quantify the evidence *for* H_0. The exploratory data analysis has already indicated that we should expect a slope close to zero (H_0). Model 1, where effect sizes should predict log-transformed mean citation scores, results in a Bayes factor of 12.9, which, with equal priors, gives a posterior probability for the null model of $p(H_0|D)$ =.93 versus $p(H_1|D)$ =.07 for the linear model. The result qualifies in Raftery's (1995) descriptive terms as positive evidence in favor of the null hypothesis. Model 2, where effect size should predict log-transformed journal impact factors, results in a Bayes factor of 13.7, which, with equal priors, also gives a posterior probability for the null model of $p(H_0|D)$ =.93 versus $p(H_1|D)$ =.07 for the linear model, also qualifying as positive evidence in favor of the null hypothesis. The result is clear, with the given equal priors, we have positive evidence, approximately 13/4 to 1, that the data are clearly most probable under the null model.

As the data analyses suggest, a model specification where sample size and journal impact factor act as controls, brings nothing. Likewise, controlling for potential differences between studies, brings nothing. Consequently, effect size is no predictor of citation rates in the present data set. Of curiosity, a model where sample size is a predictor of log-transformed citation rates yields a Bayes factor of 2.5 in favor of H_0 and posterior probabilities of $p(H_0|D)$ =.72 versus $p(H_1|D)$ =.28. The *F*-test for the model is .0558; some would declare this statistical significant at the 10% level, others will have a hard time explaining why .0499 means a statistical significant model, whereas .0558 does not. But all will fail to appreciate that the evidence against the null hypothesis is only .28 and that the odds in fact favors H_0. This is an example of the Lindley paradox (Lindley, 1957) where *p* values overstate the evidence against H_0.

Discussion

The present study supports the overall claims by Lortie et al. that the effect size of a given study in general does not directly predict its subsequent citation rate, and at an aggregated level, populations of effect sizes associated with journals did not predict the impact of journals. The findings therefore suggest that for the present data set, across five research domains, citing authors do not generally use effect sizes of a given study directly when they find primary motivation for citing an experimental study. Other epistemic factors play a role, one of them may be sample size. Considering, the numerous factors and motivations suggested that may influence citing behavior, it is perhaps not surprising that effect size alone is no good predictor of citation rates. Other studies have for example shown that in some fields studies were cited more often when results were statistically significant (Kjaergard & Gluud, 2002; Leimu & Koricheva, 2005; Etter & Stapleton, 2009). A fact Lortie et al. also stress from their findings. Nevertheless, the *de facto* zero correlation and the clear positive quantitative evidence supporting H_0 are surprising. Usually in the social sciences, we can detect the "crud factor", i.e., that "everything is related to everything else", with some reasonable samples size (Meehl, 1990). In the case of effect sizes and

citation rates in the present data set this is apparently not the case and this is surprising.

Contrary to Lortie et al., the present study is able to present numerical evidence *for* the null hypothesis (as well as the alternative hypothesis). Given equal priors, both null hypotheses are supported with approximate odds of 13/4 to 1. We find the Bayes factor superior to *p* values for assessing statistical evidence and as such our result is important. While an inspection of Lortie et al.'s Figure 1 may indicate their claim of no relation, their ritual inferential procedure is faulty. *P* values are conditional probabilities of the data given the null hypothesis and they cannot provide support for a null hypothesis, as Fisher himself pointed out (Fisher, 1934). Given that contradictory claims were present in the literature, we find it reasonable to commence our exploration with a uniform prior. Two apparent Bayesian opportunities arise from these results, further studies with uniform priors that can confirm the present findings and/or a move to a full-blown Bayesian analysis where the current finding can be used to inform the priors, meaning that H_0 should have a higher prior probability compared to H_1 and a spectrum of different priors should then be analyzed.

In the current data set, we need to investigate what may be the primary reasons for citing authors to give references to the highly cited articles, now that effect sizes apparently seems not to be a principal reason, even though they could be, given their epistemic importance in experimental studies. The issue is essential because it touches upon the question of citations' relation to research quality (aka importance). If a citation network depicts the temporal and cumulative nature of science, it is reasonably to imagine that the highly cited articles in the network, for a large part, are important nodes, where importance *also* embraces the explicit quantitative statements about the phenomenon under study such as effect sizes. If this generally seems not to be the case in the social, behavioral and medical sciences, we clearly need to examine what then characterizes these fields as cumulative when it comes to citations. Instead of positing novel far-fetched models to investigate potential citation predictors among these highly cited articles in this study, citation context analysis may be more fruitful for this restricted purpose.

Finally, it is important to examine whether in the long run, meta-analyses, with their aggregated effect sizes, will eventually be more cited on average compared to the individual studies they set out to evaluate. This may not be a foregone conclusion, as inclusion criteria, comparability of studies and aggregated effect sizes are controversial issues in the debate about meta-analyses and the purported evidence they claim.

References

Aksnes, D. W., & Rip, A. (2009). Researchers' perceptions of citations. *Research Policy, 38*(6), 895-905.

Barto, E. K., & Rillig, M. C. (2012). Dissemination biases in ecology: Effect sizes matter more than quality. *Oikos, 121*(2), 228-235.

Beck, N. N., Johannsen, M., Stoving, R. K., Mehlsen, M., & Zachariae, R. (2012). Do postoperative psychotherapeutic interventions and support groups influence weight loss following bariatric surgery? A systematic review and meta-analysis of randomized and nonrandomized trials. *Obesity Surgery, 22*(11), 1790-1797.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis - the irreconcilability of p-values and evidence. *Journal of the American Statistical Association, 82*(397), 112-122.

Berk, R. A. (2007). Statistical inference and meta-analysis. *Journal of Experimental Criminology, 3*(3), 247-270.

Berk, R. A., & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In T. G. Blomberg & S. Cohen (Eds.), *Punishment and social control* (pp. 235-254). New York: Walter de Gruyter.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37*(219), 325-335.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45-80.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003.

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

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193-242.

Ellis, P. D. (2010). *The essential guide to effect sizes. Statistical power, meta-analysis, and the interpretation of research results*. Cambridge, UK: Cambridge University Press.

Etter, J. F., & Stapleton, J. (2009). Citations to trials of nicotine replacement therapy were biased toward positive results and high-impact-factor journals. *Journal of Clinical Epidemiology, 62*(8), 831-837.

Fisher, R. A. (1934). *Statistical methods for research workers* (5th ed.). London: Oliver & Boyd.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research, 82*(3), 300-329.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review, 11*(5), 791-806.

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine, 130*(12), 995-1004.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The bayes factor. *Annals of Internal Medicine, 130*(12), 1005-1013.

Goodman, S. N. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology, 45*(3), 135-140.

Hubbard, R., & Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory and Psychology, 18*(1), 69-88.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773-795.

Kass, R. E., & Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association, 90*(431), 928-934.

Kjaergard, L. L., & Gluud, C. (2002). Citation bias of hepato-biliary randomized clinical trials. *Journal of Clinical Epidemiology, 55*(4), 407-410.

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

Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution, 20*(1), 28-32.

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44*(1-2), 187-192.

Lortie, C., Aarssen, L., Budden, A., & Leimu, R. Do citations and impact factors relate to the real numbers in publications? A case study of citation rates, impact, and effect sizes in ecology and evolutionary biology. *Scientometrics*, 1-8.

Lundberg, J. (2007). Lifting the crown-citation z-score. *Journal of Informetrics, 1*(2), 145-154.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often interpretable. *Psychological Reports, 66*(1), 195-244.

Mingers, J., & Xu, F. (2010). The drivers of citations in management science journals. *European Journal of Operational Research, 205*(2), 422-430.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Oldham, M., Kellett, S., Miles, E., & Sheeran, P. (2012). Interventions to increase attendance at psychotherapy: A meta-analysis of randomized controlled trials. *Journal of Consulting and Clinical Psychology, 80*(5), 928-939.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology 1995, Vol 25, 25*, 111-163.

Raftery, A. E. (1999). Bayes factors and bic - comment on "a critique of the bayesian information criterion for model selection". *Sociological Methods & Research, 27*(3), 411-427.

Rouder, J. N., Speckman, P. L., Sun, D. C., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237.

Schneider, J. W. (2013). Caveats for using statistical significance tests in research assessments. *Journal of Informetrics, 7*(1), 50-62.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of rho values for testing precise null hypotheses. *The American Statistician, 55*, 62 - 71.

Shadish, W. R., Tolliver, D., Gray, M., & Sengupta, S. K. (1995). Author judgments about works they cite - 3 studies from psychology journals. *Social Studies of Science, 25*(3), 477-498.

Sommer, I. E., Aleman, A., Somers, M., Boks, M. P., & Kahn, R. S. (2008). Sex differences in handedness, asymmetry of the planum temporale and functional language lateralization. *Brain Research, 1206*, 76-88.

Song, F., Eastwood, A. J., & Gilbody, S. (2000). Publication and related biases *Health Technological Assessments* (Vol. 4, pp. 1-115).

Stremersch, S., Verniers, I., & Verhoef, P. C. (2007). The quest for citations: Drivers of article impact. *Journal of Marketing, 71*(3), 171–193.

van Dalen, H. P., & Henkens, K. (2005). Signals in science - on the importance of signaling in gaining attention in science. *Scientometrics, 64*(2), 209-233.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779-804.

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry, 57*(11), 1336-1346.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Boston, MA: MIT Press.