A commentary on reporting effect size and confidence intervals: Response to Palmer and Strelan (2014)☆

Sujay Dutta [a,1], Chris Pullig [b,*]

[a] *Department of Marketing, School of Business at Wayne State University, 5201 Cass Avenue, Detroit, MI 48202, United States*
[b] *Baylor University, United States*

ARTICLE INFO

ABSTRACT

Based on recalculations of effect size (Cohen's d) and confidence intervals around these estimates (ES&CI), Palmer and Strelan (P&S; forthcoming) re-interpret some of the findings from our 2011 *Journal of Business Research* article (Dutta and Pullig 2011). In response, we critique two aspects of P&S's conclusions, the use of ES&CI in a dichotomous nature and the assumed superiority of the ES&CI approach. In addition, we discuss the implications of the ES&CI results for the substantive domain of our research and findings. Specifically, the ES&CI results do not "overturn" our conclusions, but indicate that for two variables our results should be considered tentative while for two variables the ES&CI approach leads to similar conclusions. We conclude with a discussion of the implications of the larger debate between reporting traditional null hypothesis statistical testing (NHST), ES&CI, or both in behavioral research in marketing and business-related disciplines.

1. Introduction

Palmer and Strelan (P&S; forthcoming) re-interpret some of the findings in Dutta and Pullig (2011) based on their calculation of effect size (Cohen's "d") and the confidence interval for that effect size, for each of the four dependent variables in our main study (see P&S' Table 1). Additionally, they report the results of a meta-analysis across the dependent variables. Based on the traditional method of null hypothesis statistical testing (NHST), Dutta and Pullig (2011) report a failure to reject the null hypothesis assuming equality of the effectiveness of corrective action and reduction-of-offensiveness for three of the four dependent variables. Consequently, Dutta and Pullig conclude that based on the findings that these two responses may not lead to different effects, as evidenced by three of the four aspects of CBBI. In contrast, P&S conclude from their results of the effect size and confidence interval (ES&CI)-based approach that for these three aspects of CBBI, the two responses lead to different effects.

Thus, while Dutta and Pullig conclude that corrective action might not be more or less effective than reduction-of-offensiveness in addressing some brand-related concerns, P&S conclude that corrective action is more effective than reduction-of-offensiveness in addressing these concerns. Consequently, P&S claim that their re-analysis of Dutta and Pullig data "overturns" our conclusions.

In our response, we address P&S's conclusions and speak to the larger issue of the debate between NHST and the ES&CI-based approach. First, we critique two aspects of P&S' conclusion, viz., its dichotomous nature and assumed superiority. Second, we present an integrated interpretation of the two sets of results, NHST-based and ES&CI-based, and also discuss their implications for the substantive domain that our research addresses. Finally, we discuss the role that P&S' commentary and our response to it play in the debate surrounding NHST and reporting of ES&CI estimates.

2. Dichotomous conclusion

Based on the CIs that P&S computed for the effect sizes on the four dependent variables they have concluded that corrective action leads to an effect vis-à-vis reduction-of-offensiveness for social risk, brand attitude, and brand consideration, but not for psychological risk (P&S Table 1). Thus, their conclusion differs from ours for the first three of these variables and we are in agreement on the conclusion with respect to psychological risk. Although they conclude in favor of an effect for three of the four dependent variables, we note that the CI for each of these variables includes the value zero and hence one can also conclude

in favor of a lack of effect, as we did for these variables, based on NHST. If the presence of non-zero values in a CI can lead to a conclusion in favor of an effect (as did P&S) we can argue that the presence of zero in the CI ought to lead to a conclusion in favor of a lack of effect. Since a CI contains the plausible parametric values for the statistic concerned (e.g., Cumming & Fidler, 2010; Kirk, 1995), one should not conclude that an effect exists if zero is contained in a computed CI for Cohen's "d". Granted that one can gain more insight into the likelihood of an effect (or the lack thereof) by noting whether zero appears toward the center of the CI or toward the extremities, but to categorically conclude that an effect exists with zero anywhere in the CI is erroneous. Indeed, this is how classical CIs and bootstrap-based CIs are interpreted (Hayes, 2013; Thompson, 1993) and a similar interpretation is applicable to any point estimate, including Cohen's "d", if the objective is to draw a dichotomous conclusion about a null hypothesis (e.g., Cumming & Fidler, 2010; Cumming & Finch, 2005). In summary, if drawing a dichotomous conclusion, i.e., whether or not an effect exists, was the objective, we can argue that P&S' results are consistent with our conclusion. However, the problem is deeper than this because we infer from the literature that stating a dichotomous, definitive conclusion from the ES&CI-based approach defeats one of its intended purposes.

One criticism of NHST is that it leads to oversimplification of results because researchers tend to definitively conclude in a dichotomous "Yes/No" fashion, that is, either reject the null hypothesis (typical conclusion: an effect exists) or not (typical conclusion: an effect does not exist), based on an arbitrary value for Type-I error (e.g., Kirk, 1996; Rosnow & Rosenthal, 1989). Use of confidence intervals (CIs) has been forwarded as an alternative or better approach to help avoid this trap of definitive dichotomous conclusions partly because CIs accommodate a broader set of interpretations of the results from a single sample, help generate more insight into the research question at hand, and encourage replication that might lead to a body of results over time that provide more insight into a research question than do typical single-sample NHST-driven dichotomous conclusions (Balleurka, Gómez, & Hidalgo, 2005; Cumming, 2012; Kline, 2013).

Consequently, proponents of confidence intervals likely would advise against stating dichotomous, definitive conclusions and would recommend broader interpretation of those intervals. Nonetheless, P&S state their confidence interval-based conclusion in a dichotomous and definitive fashion, specifically, by concluding in favor of an effect. As we discuss subsequently, the true value of the ES&CI-based approach arises from the avoidance of a dichotomous conclusion and the embrace of a broader style of interpretation.

3. Assumption of superiority

P&S' claim to overturn Dutta and Pullig's conclusions is most likely based on their assumption that ES&CI is an unequivocally superior approach to analysis than NHST. We do not agree that their ES&CI-based conclusions overturn or supplant our NHST-based conclusions for two reasons. First, we infer from the nature of the discussion in the literature that many of the cited problems of NHST appear to be concerns that may not necessarily occur in all instances of its application and hence research efforts must be scrutinized on a case-by-case basis to infer what specific shortcomings they entail (e.g., Gliner, Leech, & Morgan, 2002; Kline, 2004, 2013). For instance, one of the likely problems is researchers' propensity to state conclusions in deterministic rather than probabilistic terms and this can be overcome by making conclusions sound more probabilistic than deterministic (Carver, 1993), as we did in our statement of the conclusions in our paper. Indeed, our treatment of NHST successfully endured the rigorous reviewing process of a respectable journal such as *Journal of Business Research* and deserves some degree of confidence because of this reason. However, P&S do not state the exact problems they see with our application of the NHST and summarily reject the conclusions based on that approach. Indeed,

questionable application of confidence intervals is also quite possible. For instance, Kline (2004; pg. 11) writes, "… proposed alternatives, such as effect size estimation and interval estimation in individual studies and the use of meta-analysis to synthesize these results across studies, have their own potential problems. Thus, alternatives to statistical tests should not be uncritically endorsed." Given this state of affairs, a summary rejection of a specific application of NHST and unqualified acceptance of an alternative approach whose mere application does not guarantee a remedy seems highly unwarranted.

Second, the debate surrounding NHST has been raging for many decades and there are abundant viewpoints in favor of and against both NHST and the various suggested remedies, particularly the use of effect size and/or confidence intervals (Balleurka et al., 2005). Thus, just as criticism of NHST (e.g., Cohen, 1994; Murphy, 1990), advocacy for effect size reporting (e.g., Carver, 1993; Cohen, 1994), and advocacy for use of confidence intervals (Kirk, 1996, 2001; Meehl, 1997) exist in the literature, so does the defense of NHST and opposition of its criticism (e.g., Cortina & Dunlap, 1997; Hagen, 1997; Markus, 2001), and criticism of or difficulties with the suggested remedies (e.g., Abelson, 1997; Cortina & Dunlap, 1997; Fern & Monroe, 1996; Hagen, 1997; Hayes, 1998). Recently, Morey, Rouder, Verhagen, and Wagenmakers (2014; pg. 1289) comment on Cumming (2014), "… the benefits of estimation have been overstated, and the mistaken idea that estimation is superior to hypothesis testing is unfortunately becoming the conventional wisdom in psychology."

Further, despite strong criticism against it for decades, and despite some notable attempts at reform, NHST continues to be prevalent in research in psychology and other fields (Kline, 2013). Given the continued adherence to NHST despite an abundance of caution against it and the quagmire of viewpoints for and against this approach and its suggested remedies from experts in statistics to whom researchers in substantive domains largely look for guidance, we are hesitant to categorically accept the superiority of the ES&CI-based approach and P&S' conclusions therefrom. To be fair, also we would refrain from assuming the superiority of our use of NHST and the conclusions based on it. At this point in the history of the debate on the two approaches, perhaps we gain more by allowing them to operate in a collaborative, as opposed to a competitive manner.

4. Implications

We offer two sets of implications of P&S' commentary and our response to it. The first set of implications is in the specific context of the substantive domain of the research reported in our paper. The second set of implications is in the general context of the debate between the two analytical approaches, NHST and ES&CI.

4.1. Implications for substantive domain

Earlier, we stated our opinion that a dichotomous conclusion of the ES&CI-based results is unwarranted. Hence, as an illustration of the true potential of this approach as a counterpoint to dichotomous conclusions, we provide a broader interpretation of these results, largely based on the guidelines provided by Cumming (2014) and Cumming and Fidler (2010), and with complete awareness of the pitfalls inherent in interpreting CIs (e.g., Belia, Fidler, Williams, & Cumming, 2005; Hoekstra, Morey, Rouder, & Wagenmakers, 2014). A more elaborate reporting and interpretation of CIs is possible (e.g., Cumming, 2012; Cumming & Finch, 2005), but given the primarily illustrative purpose of our response, we present a very simplified set of interpretations. We also calculated ESs and CIs using Stata 13 statistical software (StataCorp, 2013) and found very similar results. Given that these estimates are similar, our discussion, for purposes of illustration, is based on the estimates presented in P&S (2014).

For social risk and brand attitude, we note that although zero is included in the CIs, it is relatively closer to the lower extremity than