DICHOTOMOUS THINKING: A PROBLEM BEYOND NHST

Jerry Lai
School of Psychological Science, La Trobe University, Australia
Kj2lai@students.latrobe.edu.au

The cliff effect—a sudden drop of confidence that a real effect exists just above p=0.05—captures the way many researchers and students interpret p-values. It is consistent with dichotomous judgements based exclusively on statistical significance (SS). Many have argued that CI can overcome over-reliance on SS. In our study, 172 researchers rated the strength of evidence against the null hypothesis as a function of 8 p-values crossed with 2 sample sizes. A further 86 received the same results presented as CIs. Although the cliff was sometimes found with p-values (23% of 172), it was more frequent with CIs (32% of 86). Thus, the argument that CIs can reduce over-reliance on SS may be overstated. Students, and also researchers, should be trained to think in terms of (or to ask) quantitative (how much A and B differ) rather than dichotomous research questions, whether analysis relies on SS or CIs.

INTRODUCTION

Suppose you obtain a *p* of .04, how strong is the evidence against H_0? What about .06? Many statistics textbooks suggest the latter provides much weaker evidence although the difference is in fact minuscule (Huberty, 1993; Gigerenzer, 2004). This type of reasoning is common in the interpretation of null hypothesis significance testing (NHST) *p*-values, and is known as dichotomous thinking (DT). It often results in an over-reliance on *p*-values and the rejection of dichotomous statistical hypotheses. Effect sizes (ES), and confidence intervals (CIs) or other quantified measures of uncertainty are often neglected. Such thinking severely undermines researchers' incentive to ask better research questions that require quantitative answers (Meehl, 1978). Consequently, students, like their teachers are taught to focus exclusively on *p*-values and statistical significance (SS).

The cliff effect is one way to measure DT. When interpreting *p*-values, the cliff effect is a sudden drop of confidence that a real effect exists just above p=0.05. Rosenthal and Gaito (1963) first demonstrated the cliff in 10 graduates and nine educators in psychology who rated their amount of confidence in research findings as a function of 14 gradually increasing *p*-values, paired with two *n*'s (n=10, n=100). Their findings were replicated by Nelson et al. (1986) with 85 psychologists, despite the inclusion of ESs. However, Poitevineau and Lecoutre (2001) questioned the robustness of the cliff effect. In their replication, they concluded that the cliff effect may have been caused by the few (4 of 18) respondents with an all-or-none interpretation of statistical significance (SS). The first aim of our study is to follow up their findings and evaluate the robustness of the cliff effect.

NHST critics argue that confidence intervals (CIs) can reduce DT, even when used as significance tests (e.g. Schmidt & Hunter, 1997). Following this logic, using CIs in place of *p*-values should reduce the cliff effect. Despite increasing support for CIs in research and education, their proposed advantages require further research. So the second aim of our study is to explore the extent to which CIs can reduce the cliff effect, and thereby DT.

METHOD

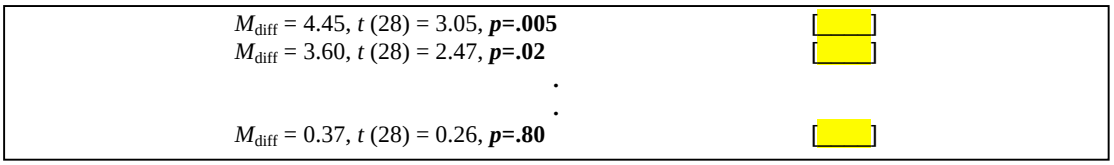
Authors of journal articles published in psychology (Psych) and medicine (Med) received one of our two surveys: the NHST, and CI surveys. Each begins by presenting a fabricated experimental scenario comparing a treatment and control group (Figure 1).

> Suppose you conduct an experiment comparing a treatment and a control group, with ***n*=15 in each group.** The null hypothesis states there is no difference between the two groups. Suppose a two-sample *t* test was conducted and a two-tailed ***p* value** calculated. [Suppose the difference between the two group means is calculated, and a 95% confidence interval placed around it].

Figure 1. Research scenario presented in the NHST and CI (in square brackets) survey

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

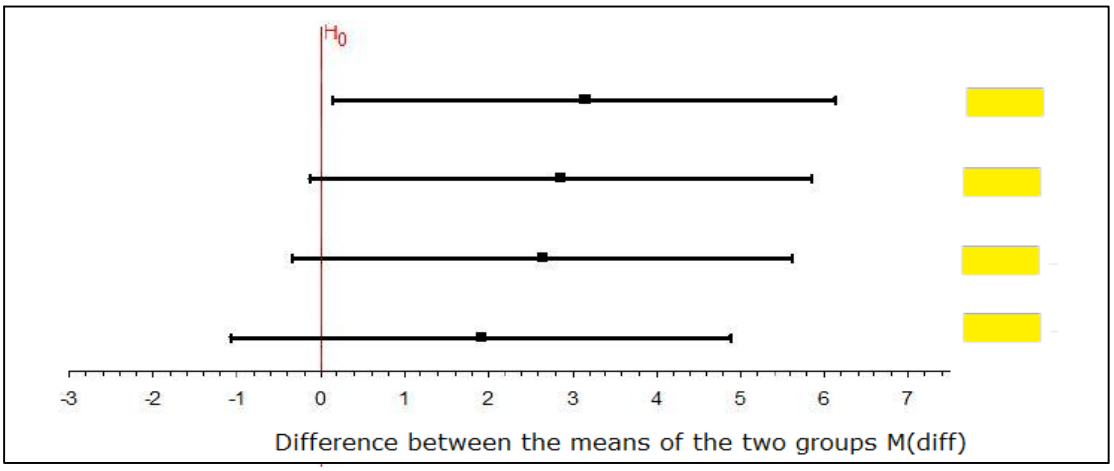
All surveys asked two main questions. For questions 1 and 2, a set of possible results of the fabricated experiment were presented as eight significance levels (p = .005, .02, .04, .06, .08, .20, .40, .80) over two sample sizes (n = 15 and n = 50, for each treatment group). All combinations assumed equal variances (SD_{pool} = 4). In the NHST survey, the two sets of hypothetical results were summarized as typical t-test outputs as advised in the APA Publican Manual (2001).



Reponses were entered in the yellow spaces

Figure 2. Snapshot of part of the NHST survey

The CI survey was in a web-based format since graphical presentation of CIs was not possible via emails. It is equivalent to the NHST survey, except the eight hypothetical results were presented as 95% CIs for the difference between the means of the two groups (M_{diff}). The width of the CIs was determined by n = 15 or n = 50, with SD_{pool} = 4.



Reponses were entered in the yellow spaces

Figure 3. Snapshot of part of the CI survey

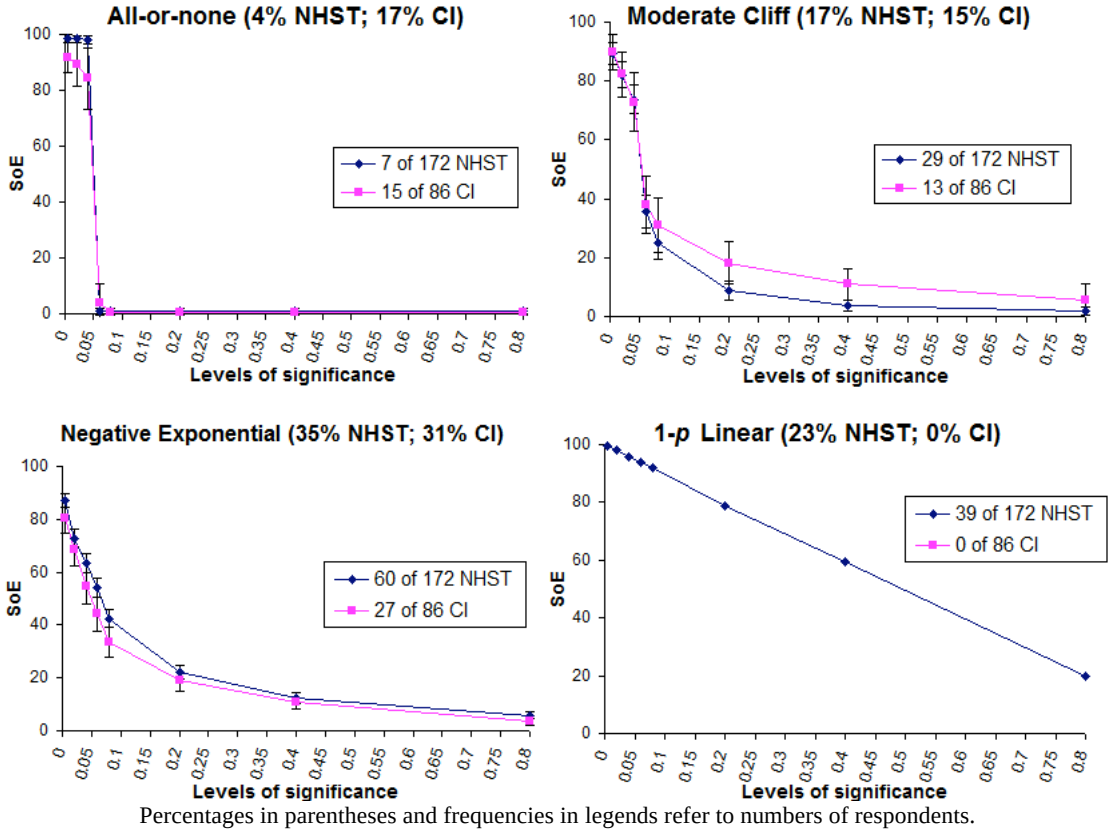
For each survey, respondents were asked to rate, for each possible result, their perceived strength of evidence against the null hypothesis of no difference. This rating scale ranged from 0 (weakest possible evidence against the null hypothesis) to 100 (strongest possible evidence against the null hypothesis). Finally respondents wrote brief comments on how they approached the questions in general.

RESULTS

Analysis of the two surveys began by the calculating the cliff ratio (CR), for each response set. This was calculated by dividing the decrease in the rated strength of evidence (SoE) from [p=0.04 to p=0.06] by the average SoE decrease from [p=0.02 to p=0.04] and [p=0.06 to p=0.08] (i.e., $[SoE_{p=.04-p=.06}] / (0.5[SoE_{p=.02-p=.04} + SoE_{p=.06-p=.08}])$).

All responses were then inspected, and clustered manually upon the similarity of their overall shape and the size of the CR (i.e. potential cliff denoted as CR > 2). The responses were initially matched with the three models identified by Poitevineau and Lecoutre (2001): all-or-none ($y = a$ if $p < 0.05$, $y = b$ *otherwise*), 1-p linear ($y = a + bp$), negative exponential ($y = exp(a + bp)$). However, further inspection of data suggested that a moderate cliff model is highly plausible. A considerable number of responses had a CR greater than 2 with a cliff clearly visible in graphs but were not as extreme as the all-or-none model. Responses not belonging to any of these groups were currently regarded as unclassified, and will not be discussed in this paper.

Figure 1 summarizes the mean SoE for each of the 8 *p*-values within the four identified models, for the two disciplines and *n*'s combined. Data inspection suggested that were not substantial differences between *n*=15 and *n*=50, and the two disciplines; analyses reported here were conducted with the two disciplines and n's combined. Despite the high similarity between the moderate cliff and negative exponential model, the former has a CR of 3.5 and 4.2 for NHST and CI; the latter has 0.9 for both NHST and CI. Hence, it is plausible that Poitevineau and Lecoutre (2001) have grouped the moderate cliff responses with the negative exponential since analyses of CR were not preformed.



Percentages in parentheses and frequencies in legends refer to numbers of respondents.

Figure 4. Mean SoE as a function of the 8 *p*-values for the four identified models and for each survey, combining the two *n*'s and two disciplines. Error bars represent 95% CIs

The proportion of responses falling into the four main categories is summarized in Table 1. There were large variations in how researchers interpret both NHST and CIs. The cliff effect was found in 22% of NHST respondents (33% for CI), and the 1-*p* linear model was found only in the NHST survey, and was often accompanied by common *p*-value misinterpretations. For example, one respondent stated that *p*-values are "the likelihood that the observed difference occurred by chance" (i.e., Odds against chance fallacy), another claimed that "I have estimated…based on probability that the null hypothesis is false" (i.e., Inverse fallacy). The negative exponential model is another common model, capturing approximately 30% of both NHST and CI responses. Also, a considerable percentage of respondents gave responses that could not be explained by the other three models.

DISCUSSION

Our findings reveal a large variation in the way NHST and CIs are interpreted. The high prevalence of the 1-*p* linear and negative exponential models implies that NHST does not necessarily entail DT. Many respondents have shown a Fisherian use of *p*-values, as a continuous measure of evidence against H_0. However, it must be stressed that the cliff models are still accountable for approximately 21 and 33% of responses, in NHST and CI interpretation respectively. Hence, the cliff is unlikely to be, as Poitevineau and Lecoutre (2001) concluded, a byproduct of the few extreme responses from the all-or-none subgroup. Furthermore, our results

reveal that only a minority of NHST responses (all-or-none; 4%) were fully consistent with the Neyman-Pearson decision-making approach (i.e. reject or do-not-reject). A much higher proportion was compatible with the hybrid logic of NHST (moderate cliff model; 17%), where SS was inappropriately used as both a decision-making criterion and a measure of evidence (see Gigerenzer, 2004, for detailed critique).

Table 1. Frequencies and percentages of NHST and CI respondents in each model, for disciplines and n's combined. 95% CIs of the percentages are presented in parentheses

Models	NHST *n*=172			CI *n*=86		
Cliff (All-or-none & Moderate)	36	21%	(.16, .28)	28	33%	(.24, .43)
Negative Exp	60	35%	(.28, .42)	27	31%	(.23, .42)
1-*p* Linear	39	23%	(.17, .30)	0	0%	(.00, .04)
Unclassified	38	22%	(.17, .29)	32	37%	(.28, .48)

Also, sample size was found to have little impact on researchers' interpretation of SS in both NHST and CI, which is inconsistent with previous research. Qualitative data revealed mixed opinions about the role of *n*. For example, one respondent asserted that: "A low *p*-value with a small sample size is indicative of greater differences between groups". Some suggested that a larger *n* enhances precision of estimates, hence gives greater SoE. Others believed that SoE is conditioned only on *p*, regardless of *n*.

Surprisingly, the cliff effect prevalence in CI interpretations was more than 50% higher than that of NHST. So Schmidt and Hunter's (1997) speculation that CIs will reduce DT may be too optimistic. DT occurs regardless the choice of statistical methods; the cliff effect occurred in both *p*-value and CI. However, the adoption of CIs is surely beneficial in the long run as they provide an integrated summary of ESs and the level of uncertainty. But this information will likely be ignored if future researchers are trained only to rely on SS and the rejection of H_0 rather than the evaluation of ESs.

Because the rejection of H_0 is so ingrained in the research tradition, replacing *p*-values with CIs alone is insufficient to overcome DT. Meehl (1978) pointed out that statistical hypotheses are often misused as theory-driven research questions, so a dichotomous conclusion is indeed a minimally sufficient answer. Therefore, other possible ways to overcome DT in both NHST and CI, are to teach future researchers to (1) formulate better research questions that require quantitative answers (i.e., to what extent A is better than B?), and (2) think and communicate in terms of ES estimation rather than dichotomous statistical hypotheses.

REFERENCES

Gigerenzer, G. (2004). Mindless statistics. The *Journal of Socio-Economics, 33,* 587-606.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman—Pearson views in textbooks. *Journal of Experimental Education, 61*(4), 317-333.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806-834.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41*(11), 1299-1301.

Poitevineau, J., & Lecoutre, B. (2001). Interpretation of significance levels by psychological researchers: The .05 cliff effect may be overstated. *Psychonomic Bulletin & Review, 8*(4), 846-850.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology, 55,* 33-38.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S, A. Mulaik & J. H. Steiger (Eds.). *What if there were no significance tests? (*pp. 37-64). Mahwah, NJ: Lawrence Erlbaum.