Will the ASA’s Efforts to Improve Statistical Practice be Successful? Some Evidence to the Contrary

Raymond Hubbard

Department of Marketing, College of Business and Public Administration, Drake University, Des Moines, IA

ABSTRACT
Recent efforts by the American Statistical Association to improve statistical practice, especially in countering the misuse and abuse of null hypothesis significance testing (NHST) and p-values, are to be welcomed. But will they be successful? The present study offers compelling evidence that this will be an extraordinarily difficult task. Dramatic citation-count data on 25 articles and books severely critical of NHST’s negative impact on good science, underlining that this issue was/is well known, did nothing to stem its usage over the period 1960–2007. On the contrary, employment of NHST increased during this time. To be successful in this endeavor, as well as restoring the relevance of the statistics profession to the scientific community in the 21st century, the ASA must be prepared to dispense detailed advice. This includes specifying those situations, if they can be identified, in which the p-value plays a clearly valuable role in data analysis and interpretation. The ASA might also consider a statement that recommends abandoning the use of p-values.

“Much has been written about the misunderstandings and misinterpretations of p-values. The cumulative impact of such criticisms in statistical practice and on empirical research has been nil” (Berry 2017, p. 896).

1. Introduction
The American Statistical Association recently launched a number of unprecedented initiatives aimed at improving statistical practice. These include the ASA statement on p-values (Wasserstein and Lazar 2016), the October 11–13, 2017 Symposium on Statistical Inference in Bethesda, MD, which in turn spawned this TAS Special Issue, “Statistical Inference in the 21st Century: A World Beyond $p < 0.05$,” an online, permanently open access issue of The American Statistician. For these efforts, the ASA is to be warmly congratulated. But will they be successful?

Leaving aside the tricky definition as to what constitutes “successful,” and the time horizon this might involve, it is clear that the ASA faces a daunting uphill battle. As Berry’s (2017) introductory quotation spells out, previous attempts to tackle this problem, especially the rampant misuse and abuse of null hypothesis significance testing (NHST) and its ubiquitous p-value, have been totally ineffective.

The present contribution provides empirical support for this claim. More specifically, it is shown how a series of highly cited articles and books exposing the fundamental weaknesses associated with NHST, and its detrimental effects on knowledge development, nevertheless were unable to prevent its inexorable rise in the social and management sciences over the period 1960–2007. This does not bode well for the ASA’s initiatives.

2. Data on the Spread of NHST, 1960–2007
Data on the spread of NHST in the social and management sciences were obtained from Hubbard (2016, pp. 16–30). He content-analyzed a randomly selected issue of leading journals in the social (geography, political science, psychology, and sociology) and management (accounting, economics, finance, management, and marketing) sciences for every year from 1960 through 2007 to estimate the incidence of empirical research using NHST in these areas. Leading journals were targeted since they would be expected to feature best research practices. For each social science discipline, these journals are given in parentheses: Geography (Annals of the Association of American Geographers, Economic Geography, Professional Geographer); Political Science (American Journal of Political Science, American Political Science Review, Public Administration Review); Psychology (Journal of Applied Psychology, Journal of Comparative Psychology, Journal of Consulting and Clinical Psychology, Journal of Educational Psychology, Journal of Experimental Psychology: General, Psychological Bulletin, Psychological Review); and Sociology (American Journal of Sociology, American Sociological Review, Social Forces).

This examination resulted in a total of 5541 empirical articles employing NHST across the four disciplines. The percentage of empirical articles employing NHST in these four disciplines, on a decade-by-decade basis (the final “decade” being 2000–2007), is given in column 10 of Table 1.

The above analysis was repeated for the management disciplines. Being younger disciplines, some journals did not span the entire 1960–2007 time period. In these cases, their inaugural publication dates are supplied: Accounting (The Accounting Review, Journal of Accounting and Economics, 1979, Journal of Accounting Research, 1963); Economics (American Economic

CONTACT Raymond Hubbard drabbuyan@aol.com Department of Marketing, College of Business and Public Administration, Drake University, Des Moines, IA 50311.
Table 1. Citation counts of selected articles and books critical of NHST, 1960–2017, and the percentage of empirical social and management science research employing it, 1960–2007.

| Years       | Article/Book                  | Citation counts | 1960–1969 | 1970–1979 | 1980–1989 | 1990–1999 | 2000–2009 | 2010–2017 | 1960–2017 | Social Sciences | Management Sciences |
|-------------|-------------------------------|----------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-------------------|---------------------|
| 1960–1969   | Rozeboom (1960)               |                | 28        | 55        | 71        | 104       | 213       | 283       | 754       | 56%               | 52%                 |
|             | Bakan (1966)                  |                | 27        | 138       | 215       | 219       | 192       | 301       | 1092      |                   |                     |
|             | Meehl (1967)                  |                | 9         | 95        | 85        | 134       | 220       | 291       | 834       |                   |                     |
|             | Lykken (1968)                 |                | 8         | 137       | 200       | 215       | 226       | 376       | 1162      |                   |                     |
|             | Total                         |                | 72        |           |           |           |           |           |           |                   |                     |
| 1970–1979   | Morrison and Henkel (1970)    |                |           | 121       | 124       | 167       | 193       | 176       | 781       | 72%               | 80%                 |
|             | Greenwald (1975)              |                |           | 54        | 194       | 215       | 193       | 272       | 928       |                   |                     |
|             | Carver (1978)                 |                |           | 7         | 208       | 267       | 334       | 342       | 1158      |                   |                     |
|             | Meehl (1978)                  |                |           | 9         | 225       | 475       | 517       | 568       | 1794      |                   |                     |
|             | Total                         |                | 616       |           |           |           |           |           |           |                   |                     |
| 1980–1989   | Leamer (1983)                 |                |           | 204       |           | 427       | 903       | 876       | 2410      | 84%               | 89%                 |
|             | Oakes (1986)                  |                |           | 18        |           | 216       | 332       | 243       | 809       |                   |                     |
|             | Berger and Sellke (1987)      |                |           | 38        |           | 230       | 328       | 440       | 1036      |                   |                     |
|             | Gigerenzer and Murray (1987) |                |           | 14        |           | 282       | 254       | 166       | 776       |                   |                     |
|             | Gigerenzer et al. (1989)      |                |           | 5         |           | 244       | 503       | 430       | 1182      |                   |                     |
|             | Rosnow and Rosenthal (1989)   |                |           | 2         |           | 224       | 344       | 393       | 963       |                   |                     |
|             | Total                         |                | 1603      |           |           |           |           |           |           |                   |                     |
| 1990–1999   | Cohen (1990)                  |                |           |           | 469       | 786       | 890       | 2445      | 92%       | 92%               |                     |
|             | Cohen (1994)                  |                |           |           | 543       | 1590      | 2070      | 4203      | 3012      |                   |                     |
|             | Kirk (1996)                   |                |           |           | 69        | 620       | 802       | 1491      |           |                   |                     |
|             | Schmidt (1996)                |                |           |           | 186       | 576       | 447       | 1209      |           |                   |                     |
|             | Harlow, Mulaik, and Steiger (1997) |            |           |           | 39        | 374       | 287       | 700       |           |                   |                     |
|             | Wilkinson et al. (1999)       |                |           |           | 12        | 1410      | 1590      | 3012      |           |                   |                     |
|             | Total                         |                | 4737      |           |           |           |           |           |           |                   |                     |
| 2000–2009   | Nickerson (2000)              |                |           |           |           | 321       | 582       | 903       | 69%       | 92%               | 93%                 |
|             | Gigernzer (2004)              |                |           |           |           | 82        | 446       | 528       |           |                   |                     |
|             | Kline (2004)                  |                |           |           |           | 298       | 734       | 1032      |           |                   |                     |
|             | Ziliak and McCloskey (2008)   |                |           |           |           | 75        | 681       | 756       |           |                   |                     |
|             | Total                         |                | 10,884    |           |           |           |           |           |           |                   |                     |
| 2010–2017   | Wasserstein and Lazar (2016)  |                |           |           |           | 762       | 762       | 14,448    | 32,360    |                   |                     |
|             | Total                         |                |           |           |           |           |           |           |           |                   |                     |

NOTE: The figures listed in the last two columns show the percentage of empirical work using NHST. They are from Hubbard (2016, p. 233).
Review, Economic Journal, Journal of Political Economy, Quarterly Journal of Economics, Review of Economics and Statistics; Finance (Journal of Finance, Journal of Financial Economics, 1974, Journal of Financial and Quantitative Analysis, 1966, Journal of Money, Credit and Banking, 1969); Management (Academy of Management Journal, Administrative Science Quarterly, Human Relations, Journal of Management, 1975, Journal of Management Studies, 1964, Organizational Behavior and Human Decision Processes, 1966, Strategic Management Journal, 1980); and Marketing (Journal of Consumer Research, 1974, Journal of Marketing, Journal of Marketing Research, 1964).

All told, some 7762 empirical articles in the management sciences relied on NHST. Column 11 of Table 1 shows this as a percentage on a decade-by-decade basis.

### 3. Citation Analysis

Table 1 presents the Google Scholar citations earned by 25 works (18 articles, 7 books) from the dates of their initial publication through December 10, 2017. It reveals that these ostensibly influential publications highly critical of the uses and abuses of NHST were nevertheless unable to reduce its prevalence. On the contrary, every decade from 1960 to 2007 saw an increase, or no diminution, in its adoption.

#### 3.1. Decade-by-Decade Citation Analysis

During the 1960s, when approximately 52%–56% of empirical studies in the management and social sciences employed NHST, four of the earliest and most damning indictments of how its use retards good science appeared in the literature. The articles by Rozeboom (1960), Bakan (1966), Meehl (1967), and Lykken (1968) ignited a debate that continues to this day. Between them they garnered 72 citations in this decade.

The beginning of the 1970s saw the publication of Morrison and Henkel's (1970) anthology, The Significance Test Controversy, which included the four articles mentioned above, draw further attention to this issue. Later that decade, Greenland (1975), Carver (1978), and Meehl (1978) expressed concerns over the damaging effects of NHST. These four efforts attracted 191 citations for 1970–1979. The cumulative citations earned by the eight works listed in Table 1 for 1960–1979 now stood at a respectable 688. Yet 1970–1979 revealed a substantial increase over 1960–1969 in the proportion of empirical research using NHST in both the social (from 56% to 72%) and management (from 52% to 80%) sciences (see Table 1).

Six further prominent critiques of NHST appeared in the 1980s, most of them in the latter half. During this time they gathered 281 citations, with Leamer (1983), at 204, responsible for the lion’s share of these. For 1980–1989, a total of 1603 citations accrued to the 14 contributions in Table 1 decrying NHST. Meanwhile, the cumulative citations of these reached 2291. Despite this, the percentage of empirical research based on NHST forged ahead to 84% in the social, and 89% in the management, sciences.

The 1990s witnessed the arrival of two juggernauts by Cohen (1990, 1994) on the damage to scientific progress caused by NHST. At 543, his 1994 article gained the most citations for this decade (or any before it), while his 1990 publication, at 469, took third place. Sandwiched between these, on 475, was Meehl (1978). A total of 4737 citations of research challenging the ubiquity of NHST occurred during 1990–1999. Cumulatively, this total rose to 7028. By now, almost predictably, the incidence of NHST featured in empirical research moved ahead in both the social, 92%, and management, 92%, sciences.

Four additional, noteworthy entrants lamenting the deleterious consequences of NHST for advancing good science debuted in 2000–2009. Consisting of two articles, Nickerson (2000) and Gigerenzer (2004), and two books, Kline (2004) and Ziliak and McCloskey (2008), these acquired some 776 citations during that time. Elsewhere for 2000–2009, the numbers increased dramatically. Weighing in with 1590, Cohen (1994) again probed the most citations for this decade, while Wilkinson et al. (1999), on 1410, claimed the second spot. Leamer (1983), at 903, captured third place. All told, the 24 contributions in Table 1 gained 10,884 citations for 2000–2009; and 17,912 cumulatively. At 92% and 93%, respectively, the proportion of empirical research relying on NHST (which data cover only the period 2000–2007) in the social and management sciences may (or may not) have plateaued.

The largest number of citations attracted by the 25 studies in Table 1 is for the (partial) decade 2010–2017. Over this period, these works were cited 14,448 times. Cohen (1994) and Cohen (1990) took the first and third places with 2070 and 890, in turn. Wilkinson et al. (1999), on 1590, are runners-up. A recent newcomer to this debate warrants special attention. This is Wasserstein and Lazar's (2016) “The ASA’s Statement on p-Values: Context, Process, and Purpose.” In less than 2 years this statement has gone viral, from a citations-generated perspective, with 762 of them. Incredibly, the cumulative total number of citations won by the 25 articles and books in Table 1 for the period 1960–2017 is 32,360.

In closing this section, it is instructive to note that over these same decades very few articles were published attempting to defend the practice of NHST. A notable exception is Hagen (1997). Another is Wainer (1999), although he could only muster “One cheer for null hypothesis significance testing.” And while virtually all of the 25 works listed in Table 1 are severely critical of NHST and its baneful consequences for scientific progress, some, for example, Nickerson (2000, p. 241), remark “that when applied with good judgment it [NHST] can be an effective aid to the interpretation of experimental data.” The only problem is that, in a 60-page journal article, Nickerson never provides us with examples showing how.

#### 3.2. Further Insights from Table 1

Other patterns in the data in Table 1 deserve notice. The first is that some 14 of the 24 (58%) entries show monotonic increases in their citation rates over the decades (Wasserstein and Lazar, 2016, were excluded because there is no prior decade for comparison). This is especially impressive for the earlier-published articles, such as Rozeboom (1960), Lykken (1968), Carver (1978), and Meehl (1978).

Second, for 17 of the 24 (71%) entries, the period 2010–2017 is the one recording their maximum number of citations. This latest surge in numbers, which is evident also in the years preceding 2010–2017, possibly could reflect greater concern
with bad research practices. On the other hand, it is far more likely that this growth in citations is attributable to the pro-
fusion of journals, the development of more extensive and
sophisticated citation-tracking systems, and manipulations to
inflate them, over recent decades.

Reiterating from Section 3.1, across the period 1960–2017
the 25 works listed in Table 1 have amassed a total of 32,360
citations, or an average of 1294 apiece. These figures are stag-
ggering when it is acknowledged that most research goes uncit-
(Hubbard 2016, p. 238). The five most-cited pieces in Table 1
are Cohen (1994), 4203; Wilkinson et al. (1999), 3012; Leamer
(1983), 2410; Cohen (1990), 2145; and Meehl (1978), 1794.
Amazingly, the late Jacob Cohen authored 40% of these! Not
surprisingly, given that they were among the earliest adopters—
and subsequent critics—of NHST (Hubbard 2016, p. 21), 80%
of this group are psychologists. These percentages are mirrored
in Table 1 as a whole, with 19 of the 25 (76.0%) entries written
by scholars trained in psychology, the remainder by economists
(Leamer 1983; Ziliak and McCloskey 2008), sociologists (Mor-
rison and Henkel 1970), statisticians (Berger and Sellke 1987;
Wasserstein and Lazar 2016), and an interdisciplinary mix
(Gigerenzer et al. 1989).

5. Conclusions

This article has furnished overwhelming evidence support-
ing Berry’s (2017) introductory quotation regarding the abject
failure to rein in the use and abuse of NHST and p-values.
With 32,360 citations between them—the kind of publicity only
dreamed of in academic circles—25 publications severely criti-
cal of NHST have not been able to arrest, never mind reverse, its
growing popularity in the social and management sciences. In
fact, I would not be at all surprised to learn that its percentage
usage in empirical work in these areas has actually inched ahead
of the 92% and 93% highs for 2000–2007 in more recent years.

If the ASAs late intervention in this decades-long saga to
improve statistical practice is to have any hope of success, its
advice must be both specific and possibly radical. By specific I
mean identifying those, if any, situations where p-values make
a genuine contribution. As amply demonstrated, general state-
ments have been spectacularly unsuccessful. By radical, I mean
that the ASA might contemplate issuing a statement banning the
use of p-values. Such actions will be difficult, but necessary, if
the statistics profession is to be relevant in both the classroom, and
in scientific method, in the 21st century.

References

Bakan, D. (1966), “The Test of Significance in Psychological Research,” Psy-
chological Bulletin, 77, 423–427. [32,33]
Berger, J. O., and Sellke, T. (1987), “Testing a Point Null Hypothesis: The
Irreconcilability of p Values and Evidence,” Journal of the American Sta-
tistical Association, 82, 112–122. [32,34]
Berry, D. (2017), “A p-Value to Die For,” Journal of the American Statistical
Association, 112, 895–897. [31,34]
Briggs, W. M. (2017), “The Substitute for p-Values,” Journal of the American Statistical
Association, 112, 897–898. [34]
Carver, R. P. (1978), “The Case Against Statistical Significance Testing,”
Harvard Educational Review, 48, 378–399. [32,33]
Cohen, J. (1990), “Things I Have Learned (So Far),” American Psychologist,
45, 1304–1312. [32,33,34]
——— (1994), “The Earth is Round (p < .05),” American Psychologist, 49,
997–1003. [32,33,34]
Cumming, G. (2014), “The New Statistics: Why and How,” Psychological
Science, 25, 7–29. [34]
Gelman, A., and Carlin, J. (2017), “Some Natural Solutions to the p-Value
Problem—and Why They Won’t Work,” Journal of the American Statistical
Association, 112, 899–901. [34]
Gigerenzer, G. (2004), “Mindless Statistics,” Journal of Socio-Economics,
33, 587–606. [32,33]
Gigerenzer, G., and Murray, D. J. (1987), Cognition as Intuitive Statistics,
Hillsdale, NJ: Lawrence Erlbaum. [32]
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., and Krüger,
L. (1987), The Empire of Chance: How Probability Changed Science and
Everyday Life, Cambridge, UK: Cambridge University Press. [32,34]
Greenwald, A. G. (1975), “Consequences of Prejudice Against the Null
Hypothesis,” Psychological Bulletin, 82, 1–20. [32,33]
Hagen, R. L. (1997), “In Praise of the Null Hypothesis Statistical Test,” Ame-
rican Psychologist, 52, 15–24. [33]
Harlow, L. L., Mulaik, S. A., and Steiger, J. H. (eds.) (1997), What If There
Were No Significance Tests? Mahwah, NJ: Lawrence Erlbaum. [32]
Hubbard, R. (2016), Corrupt Research: The Case for Reconceptualizing
Empirical Management and Social Science, Thousand Oaks, CA: SAGE
Publications, Inc. [31,34]
Kirk, R. (1996), “Practical Significance: A Concept Whose Time Has Come,”
Educational and Psychological Measurement, 56, 746–759. [32]
Kline, R. B. (2004), Beyond Significance Testing: Reforming Data Analy-
sis Methods in Behavioral Research, Washington, DC: American Psychological Association. [32,33]
Leamer, E. E. (1983), “Let’s Take the Con Out of Econometrics,” American Economic Review, 73, 31–43. [32,33,34]
Lykken, D. T. (1968), “Statistical Significance in Psychological Research,” Psychological Bulletin, 70, 151–159. [32,33]
Matthews, R. (2017), “The ASA’s p-Value Statement One Year On,” Significance, April, 38–40. [34]
McShane, B. B., and Gal, D. (2017), “Statistical Significance and the Dichotomization of Evidence,” Journal of the American Statistical Association, 112, 885–895. [34]
Meehl, P. E. (1967), “Theory-Testing in Psychology and Physics: A Methodological Paradox,” Philosophy of Science, 34, 103–115. [32,33]
——— (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. [32,33,34]
Morrison, D. E., and Henkel, R. E. (eds.) (1970), The Significance Test Controversy: A Reader, Chicago, IL: Aldine. [32,33,34]
Nickerson, R. S. (2000), “Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy,” Psychological Methods, 5, 241–301. [32,33]
Oakes, M. (1986), Statistical Inference: A Commentary for the Social and Behavioural Sciences, Chichester, UK: Wiley. [32]
Rosnow, R. L., and Rosenthal, R. (1989), “Statistical Procedures and the Justification of Knowledge in Psychological Science,” American Psychologist, 44, 1276–1284. [32]
Rozeboom, W. W. (1960), “The Fallacy of the Null-Hypothesis Significance Test,” Psychological Bulletin, 57, 416–428. [32,33]
Schmidt, F. L. (1996), “Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers,” Psychological Methods, 1, 115–129. [32]
Trafimow, D., and Marks, M. (2015), “Editorial,” Basic and Applied Social Psychology, 37, 1–2. [34]
Wainer, H. (1999), “One Cheer for Null Hypothesis Significance Testing,” Psychological Methods, 4, 212–213. [33]
Wasserstein, R. L., and Lazar, N. A. (2016), “The ASA’s Statement on p-Values: Context, Process, and Purpose,” The American Statistician, 70, 129–133. [31,33,34]
Wilkinson, L., and American Psychological Association Task Force on Statistical Inference (1999), “Statistical Methods in Psychology Journals: Guidelines and Explanations,” American Psychologist, 54, 594–604. [33,34]
Ziliak, S. T., and McCloskey, D. N. (2008), The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives, Ann Arbor, MI: University of Michigan Press. [32,33,34]