

ARTICLES

Reform of statistical inference in psychology: The case of *Memory & Cognition*

SUE FINCH

University of Melbourne, Melbourne, Victoria, Australia

and

GEOFF CUMMING, JENNIFER WILLIAMS, LEE PALMER, ELVIRA GRIFFITH, CHRIS ALDERS,
JAMES ANDERSON, and OLIVIA GOODMAN

La Trobe University, Melbourne, Victoria, Australia

Geoffrey Loftus, Editor of *Memory & Cognition* from 1994 to 1997, strongly encouraged presentation of figures with error bars and avoidance of null hypothesis significance testing (NHST). The authors examined 696 *Memory & Cognition* articles published before, during, and after the Loftus editorship. Use of figures with bars increased to 47% under Loftus's editorship and then declined. Bars were rarely used for interpretation, and NHST remained almost universal. Analysis of 309 articles in other psychology journals confirmed that Loftus's influence was most evident in the articles he accepted for publication, but was otherwise limited. An e-mail survey of authors of papers accepted by Loftus revealed some support for his policy, but allegiance to traditional practices as well. Reform of psychologists' statistical practices would require more than editorial encouragement.

Critics have long argued that null hypothesis significance testing (NHST) is often inappropriately used by psychologists to make binary decisions about whether an effect exists or not (e.g., Bakan, 1966; Carver, 1978, 1993; Cohen, 1990, 1994; Hammond, 1996; Harlow, Mulaik, & Steiger, 1997; Loftus, 1993a, 1993b, 1996; Lykken, 1968; Meehl, 1978; Oakes, 1986; Schmidt, 1992; Tukey, 1969). Such binary decisions ignore other information in the data, such as the direction and size of effects and the range of parameter estimates that are consistent with the data obtained. Until recently, there has been little official response to this critique, and NHST remains the dominant technique for drawing conclusions from data (Finch, Cumming, & Thomason, 2001; Finch, Thomason, & Cumming, 2002; Nickerson, 2000).

However, a report from the American Psychological Association (APA) Task Force on Statistical Inference (TFSI; Wilkinson & the TFSI, 1999) and the latest edition of the APA's *Publication Manual* (APA, 2001) support

the use of alternatives to NHST. Recommendations include the reporting of exact p values, effect sizes, and confidence intervals (CIs) and use of graphical displays to investigate the data, and communicate information about their distribution and inferential statistics. Wilkinson and the TFSI stated: "In all figures, include graphical representations of interval estimates whenever possible" (p. 601), and the use of CIs is strongly recommended by the *APA Manual* (p. 22). The recommendations of Wilkinson and the TFSI and the new *APA Manual* represent a major change of emphasis from the tools psychologists have been using for the past 50 years to draw inferences from data.

Reform advocates have stated that journal editors have an important role to play in reform; some have even argued that editors are the "only force that can effect change" (Sedlmeier & Gigerenzer, 1989, p. 315). Kirk (1996) suggested that changes in editorial policies "would cause a chain reaction" (p. 757) of changes to courses, textbooks, and journal authors' inference strategies.

A case study of reform instigated by an editor can be found in a psychology journal in the 1990s. As editor-elect of *Memory & Cognition*, in 1993 Geoffrey Loftus published an editorial describing his guidelines on data analysis (Loftus, 1993a). Loftus's goal was to "try to decrease the overwhelming reliance on hypothesis testing as the major means of transiting from data to conclusions" (p. 3). He proposed to "emphasize the increased use of figures depicting sample means along with standard error bars" (p. 3), and offered the following guidelines:

This research was supported by the Australian Research Council. We thank Geoffrey Loftus for consultation via e-mail, and Liora Pedhazur Schmelkin and Bruce Thompson for comments on a draft. Reviews from Peter Dixon, Geoffrey Loftus, and Michael Masson resulted in a much improved manuscript. Correspondence about this article may be addressed to either S. Finch, Statistical Consulting Centre, University of Melbourne, Melbourne, Victoria 3010, Australia, or G. Cumming, School of Psychological Science, La Trobe University, Melbourne, Victoria 3086, Australia (e-mail: suefinch@hotmail.net.au, g.cumming@latrobe.edu.au).

1. By default, data should be conveyed as a figure depicting sample means *with associated standard errors, and/or where appropriate, standard deviations*.

2. More often than not, inspection of such a figure will immediately obviate the necessity of any hypothesis-testing procedures. In such situations, presentation of the usual hypothesis-testing information (F values, p values, etc.) will be discouraged. (p. 3, emphasis in the original)

Loftus presented his guidelines as just that, and stated that he would “happily consider whatever technique by which an author believes this goal can best be accomplished” (p. 3). He believed that

overreliance on the impoverished binary conclusions yielded by the hypothesis-testing procedure has subtly seduced our discipline into insidious conceptual cul-de-sacs that have impeded our vision and stymied our potential. . . . [T]here are often better ways of trying to convey what the data from an experiment are trying to tell us. (p. 3)

In this article, we evaluate the success of Loftus’s attempted reform of statistical analysis and reporting practices. We present data describing the kinds of statistical analyses that appeared in journal articles and also report on authors’ opinions of the proposed reform.

Our first goal is to examine the response to Loftus’s recommendations by describing the way data and statistics were presented in articles published in *Memory & Cognition* before, during, and after Loftus’s time as editor. We assess the extent to which articles conformed to his proposals. We also present comparative data describing other articles of authors who published in *Memory & Cognition* under Loftus, and articles published in similar journals. The results are presented in Phases 1–3. Our second goal is to describe authors’ experience of publishing under Loftus and their views of his editorial guidelines; data are described in Phase 4. Our broader purposes are to provide a characterization of current statistical practice and to consider implications for the reform of statistical practices in psychology.

PHASES 1–3

Studies of Statistical Reporting Practices in Published Articles

Inclusion Criteria

In Phases 1–3, we examined published journal articles that described empirical studies reporting original analy-

ses of sample means, including sample proportions and percentages (*empirical articles*). We excluded theoretical articles and meta-analyses.

Reporting Practices

We identified and described a number of practices relating to the presentation of statistical information. Trials of independent coding, followed by discussion, helped us identify seven practices that we could code reliably; these allow a description of how well articles followed Loftus’s recommendations. We examined each article and noted whether we could find at least one example of a practice.

The first practices we examined relate directly to Loftus’s (1993a) instructions:

Practice 1. Displaying sample means in a figure.

Practice 2. Displaying error bars on sample means in a figure. We recorded the kinds of bars: standard errors (SEs), standard deviations (SDs), CIs (including the level of confidence), and unspecified.

Practice 3. Displaying results as means with error bars but without an accompanying NHST. This is a subset of Practice 2.

Other practices consistent with Loftus’s recommendations were:

Practice 4. Reporting CIs presented outside of figures. We recorded the level of confidence.

Practice 5. Mention of error bars in text. This practice referred to any description or discussion of error bars beyond a simple statement that they were shown in a figure.

We also examined practices contrary to the spirit of Loftus’s recommendations:

Practice 6. Any use of NHST for means.

Practice 7. Use of NHST for means reported in a figure (a subset of Practice 6).

Types of Articles

We classified the articles into the following four mutually exclusive types. Table 1 describes the definitions of the four types in terms of the seven practices.

1. In an *interval-inclusive* article, a figure with error bars was used (Practice 2) and/or CIs were reported (Practice 4); other examined practices may or may not have been used.

2. An *NHST-with-figure* article relied on NHST (Practice 6) but included a figure of means without error bars

Table 1
Four Types of Empirical Articles Defined in Terms of Seven Practices

Practice	Type of Article			
	Interval Inclusive	NHST With Figure	NHST Only	Noninferential
1. Means displayed in a figure	Optional	Necessary	Absent	Optional
2. Error bars displayed in a figure	Sufficient*	Absent	Absent	Absent
3. Means and error bars displayed in a figure, no corresponding NHST	Optional	Absent	Absent	Absent
4. CI in a table or text	Sufficient*	Absent	Absent	Absent
5. Mention of error bars in text	Optional	Absent	Absent	Absent
6. NHST for means	Optional	Necessary	Necessary	Absent
7. NHST for means that also appear in a figure	Optional	Optional	Absent	Absent

*By definition, *interval-inclusive* articles require at least one instance of Practice 2 or Practice 4.

(Practice 1 but not Practice 2). This type of article was partly consistent with Loftus's suggestions.

3. In an *NHST-only* article, NHST was used (Practice 6) but no other practices that we examined were in place. NHST-only articles were not consistent with Loftus's recommendations.

(4) A *noninferential* article reported sample means but did not use any of Practices 2–7 (NHST, CIs, or error bars).

PHASE 1

Articles Published in *Memory & Cognition*, 1990–2000

In Phase 1, we examined reporting practices in 696 empirical articles published in *Memory & Cognition* from 1990 to 2000. Loftus was editor from 1994 to 1997. *Pre-Loftus* articles were published in 1990, 1991, and 1992—before Loftus's (1993a) policy was published. Articles published in 1993 were not included, to avoid any contamination from early responses to Loftus's editorial. Margaret Jean Intons-Peterson preceded Loftus as editor, having commenced in 1990. In her incoming editorial, Intons-Peterson (1990) mentioned the use of descriptive statistics, including variability estimates. However, her emphasis was on traditional inferential statistics.

The *Loftus* articles are all the empirical articles accepted for publication by Loftus. They are articles published between 1994 and 1997, except those 1994 articles handled by Intons-Peterson, plus 1998 articles handled by Loftus. *Post-Loftus* articles were those published in 1998 (excluding those handled by Loftus), 1999, and all but Issue 6 of 2000. All of these were accepted for publication by Morton Ann Gernsbacher. Gernsbacher's (1998) editorial made no comment on statistical procedures or on Loftus's recommendations.

Table 2 shows the number of articles examined in each year. In any year, fewer than 6% of the articles published did not meet our inclusion criteria.

Coding Reliability

In Phase 1, coding reliability was established for all phases. The eight authors of the present work were the coders. A random sample of 41 (11%) articles from 1992 to 1998 was recoded by a second coder. There were 13 items of information to be recorded for each article (seven practices, plus details of error bars and CIs). All 9 coding discrepancies (out of a possible 533) arose from missed items. Four were missed items in the original coding, meaning that 99% of the originally coded details were confirmed on recoding. Misses were distributed across items and coders. We accepted this level of coding accuracy as adequate.

Results

The results are presented in Figures 1–3 and summarized in Table 3. Figures 1–3 show the results year by year, and Table 3 presents percentages for the pre-Loftus, Loftus, and post-Loftus periods.

Table 2
Phase 1: Number of Empirical Articles Coded in Each Year

Period	Year	Number of Articles
Pre-Loftus	1990	59
	1991	52
	1992	64
Total		175
Loftus	1994	36
	1995	58
	1996	62
	1997	72
	1998*	65
Total		293
Post-Loftus	1998	23
	1999	92
	2000†	113
Total		228
Overall total		696

*Includes one article from 1999 accepted by Loftus. †Includes articles from all but the last issue of 2000.

We present figures without error bars because, arguably, we have a complete description of reporting practices in *Memory & Cognition* over the years that we examined in Phase 1; we report for each year complete data rather than a sample. However, in the rightmost column of Table 3 we provide the maximum *SE* for a proportion given the total number of articles under consideration, for each row. These *SEs* refer to a single sample estimate of a proportion; the corresponding *SE* for a difference in proportions would be about 1.4 (actually $\sqrt{2}$; Loftus & Masson, 1994) times as large. The differences we discuss are large relative to these *SEs*.

Types of articles. Figure 1 shows the percentages of articles classified as one of three of the four types defined above. At most, 5% of articles were noninferential; these are not shown in Figure 1 (see Table 3). Just over half of the pre-Loftus articles were NHST-only (Figure 1 and Table 3). In the Loftus years, 29% to 37% of articles were NHST-only; post-Loftus (in 2000), this proportion rose to a high of 55%. The percentage of NHST-with-figure articles dropped from 37% pre-Loftus to 15% in Loftus's final year (1998). Post-Loftus, the percentage rose to an average of 22% (Table 3).

An average of only 8% of pre-Loftus articles were interval inclusive. The percentage rose dramatically during Loftus's time, peaking at 51% (1997 and 1998), but then dropped post-Loftus to 25% by 2000 (Figure 1; Table 3).

Responses to Loftus's recommendations. Loftus recommended using figures with error bars in lieu of NHST. Figure 2 shows use of error bars (Practice 2) and CIs (Practice 4 and Practice 2 when error bars are CIs). In both cases, there were strong increases under Loftus and a marked reduction post-Loftus. However, under Loftus fewer than half the articles showed error bars on figures, and at most one third included CIs (Figure 2, Table 3).

We identified articles that followed Loftus's recommendations closely in that they used figures with bars

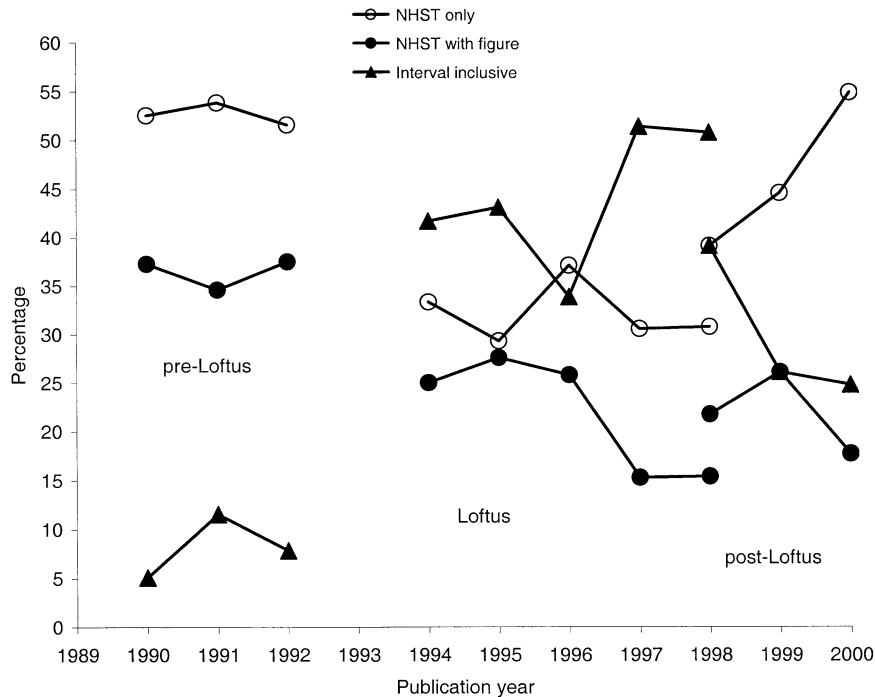


Figure 1. Percentages of three types (see text) of empirical articles published in *Memory & Cognition* in the pre-Loftus, Loftus, and post-Loftus periods (Phase 1).

without any reports of NHST (Practice 3 and absence of Practice 6). We call these articles *full Loftus*. A maximum of 6% of articles in 1997 were full Loftus. Only one of the pre- or post-Loftus articles was full Loftus.

Error bars. Inclusion of error bars (Practice 2) peaked at 47% in 1997 (Figure 2). When error bars were labeled, they were identified as *SE* or *CI* bars, except for one article with *SD* bars. Figure 3 shows the percentage of articles that included *SE* and *CI* bars. During Loftus's time, there was a partial shift from use of *SE* bars to use of *CI* bars. Pre-Loftus, few bars were provided (Figure 2), but in 40% to 50% of articles with bars (see Figure 3) the bars were unclear, meaning that at least one figure showed bars without any statement of what they represented. The peak percentage of Loftus articles with bars in which the bars were unclear was 38% in 1995, but it dropped to 7% in 1998. By 2000, the percentage of articles with bars in which the bars were unclear was 29%. When *CI*s were published (Practice 4), the 95% level of confidence was used in all but two unspecified cases.

Error bars were mentioned in the text (Practice 5) in a maximum of 21% of articles that included bars, in 1997. (Any reference at all to bars beyond a bald statement that a figure showed error bars counted as a "mention.") They were never mentioned in articles from 1990 to 1992. Overall, only 6% of the articles accepted by Loftus both included and mentioned error bars. Post-Loftus, the figures were 9% (1998), 4% (1999), and 3% (2000).

Discussion

Loftus's call for reform had some success. The profile of article types accepted by Loftus differs markedly from the profile of those accepted by his editorial predecessor. Under Loftus, 32% of articles were NHST-only, in comparison with 53% in 1990–1992. The percentage of articles including error bars accepted by Loftus was almost six times the pre-Loftus percentage (41% vs. 7%). Figure 2 shows a strongly increasing proportion of articles using *CI*s over Loftus's period as editor. Publication of full Loftus articles and articles with discussion of error bars reached a low peak in 1997. Labeling of error bars was best under Loftus in 1998.

Clearly, Loftus's success was limited. Fully 32% of the articles he accepted were NHST-only and fewer than half (41%) reported any bars. Very few authors followed Loftus's guidelines fully by relying on figures to the exclusion of NHST. Very few authors discussed the error bars shown in their figures.

In 1997, a maximum of 21% of articles with bars mentioned bars in text. In total, only 26 (4%) of the 696 articles we examined both included and mentioned bars, and few of those mentions supported substantive interpretation of the data. We provide some examples, but, to preserve authors' anonymity, we do not cite the references, which are available on request.

Some authors simply described what the bars were, as in "a 95% confidence interval." Others used *CI*s to do

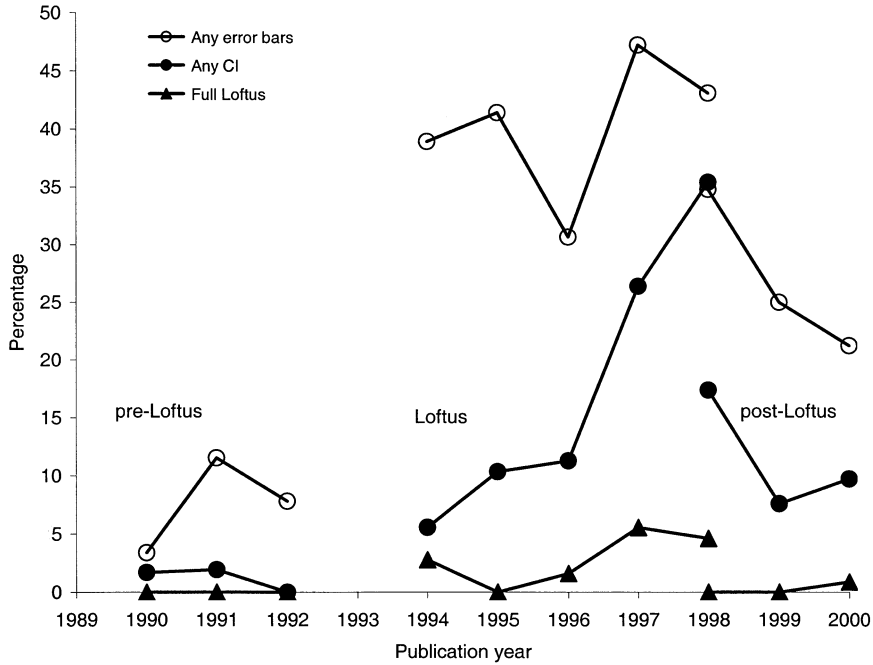


Figure 2. Percentages of three additional types (see text) of empirical articles published in *Memory & Cognition* (Phase 1).

traditional NHST, as in “the . . . score . . . was significantly greater than zero (as indicated by the 95% CI . . .).” Some of these authors relied largely on the traditional language of NHST: “Thus, confidence intervals . . . were calculated in order to assess the magnitude of the differences between conditions They revealed

that Conditions 1 and 2 (5.5% vs. 10.5% errors, respectively) differed significantly, indicating an . . . effect.”

In four articles, the authors referred to overlap of the bars: “The estimates . . . are quite different and . . . [the] confidence intervals do not overlap . . .” “The confidence intervals for parameter *c* . . . overlap almost completely,

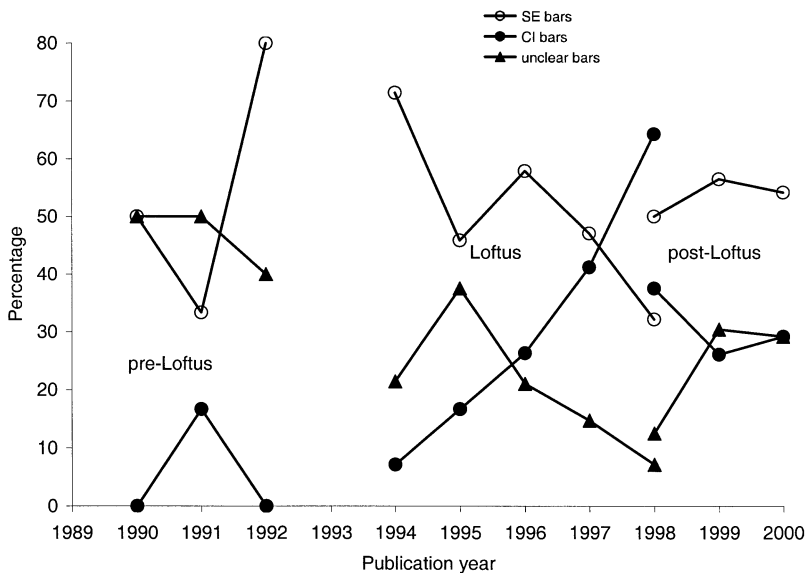


Figure 3. Percentages of three types of error bars in empirical articles published in *Memory & Cognition* that included a figure with error bars (Phase 1).

Table 3
Phases 1–3: Percentages of Articles of Various Types

Phase and Period	Main Types of Articles*				Additional Types of Articles†				Maximum <i>SE</i>
	Interval Inclusive	NHST With Figure	NHST Only	Noninferential	Full Loftus‡	Any CI§	Any Bar**	No.†	
Phase 1: <i>M&C</i> , 1990–2000									
Pre-Loftus (1990–1992)	8	37	53	3	0	1	7	175	3.8
Loftus (1994–1998)	45	21	32	2	3	19	41	293	2.9
Post-Loftus (1998–2000)	27	22	49	3	<1	10	24	228	3.3
Phase 2: Loftus's <i>M&C</i> authors									
Before <i>M&C</i>	11	33	54	3	0	2	11	95	5.1
<i>M&C</i> (Loftus)	43	19	35	3	1	19	41	95	5.1
After <i>M&C</i>	22	37	37	4	1	7	21	95	5.1
Phase 3: Other authors, recent years	22	36	40	2	1	5	22	119	4.6

Note—*M&C*, *Memory & Cognition*. *In each row, the four main types of articles sum to 100% except for rounding errors. †Number of articles for each row (maximum *SE*, in percentage points, for each row). ‡*Full Loftus* refers to an article using figures with bars (Practice 2) without any reports of NHST (absence of Practice 7). §CI as bars in figure, or as numerical values in text or table. **Any error bars in figure.

showing that these parameter estimates differ very little between conditions” (Cumming & Finch, 2003, discuss interpretation of overlapping bars).

The discussion of error bars could hardly be described as substantive. In the third example above, a reason for calculating CIs is given: to assess the size of differences. However, no authors discussed CI width or the precision of estimates. Such discussion could provide a useful basis for describing patterns in results and to suggest plausible interpretations of the results in terms of the original measurement scale. Authors could conceivably have used figures, error bars, and CIs more generally to evaluate their results, but we found virtually no evidence in the published articles that authors were relying on error bars to support their conclusions in the way that Loftus proposed. Rather, figures with bars and CIs simply appeared along with traditional NHST analyses.

Lack of support from the editorial team does not appear to have been a factor in Loftus's limited success. Loftus's associate editors were cognizant of his views on data analysis and presentation and generally supported and followed his guidelines; reviewers were aware of Loftus's philosophy and did not deviate substantially from it (G. R. Loftus, personal communication, June 29, 2000).

Loftus routinely asked authors who submitted an NHST-only article to provide figures and *SEs*. In some cases, he himself provided the *SEs* or the appropriate graphical displays. He (G. R. Loftus, personal communication, April 6, 2000) reported that

I must have . . . requested approximately 300 confidence intervals . . . , and I probably computed about 100 . . . for . . . investigators who had no idea of how to compute confidence intervals based on information provided by their ANOVA packages. . . . Many people still insisted on . . . enormous tables rather than graphing . . . (I also . . . provided graphs from tables for numerous investigators).

However, full support and systematic follow-up was not possible for all of the authors who did not respond to direct requests for changes to data analysis and presentation.

Loftus noted several additional difficulties that authors had with his requests: “Many people seemed to confuse standard errors with standard deviations. . . . Many people seemed to exhibit deep anxiety at the prospect of abandoning their *p* values” (G. R. Loftus, personal communication, April 6, 2000). Authors appeared to have difficulty producing appropriate summary tables, visual displays, and *SEs*. Popular statistics packages used by psychologists might not have readily produced the appropriate information (G. R. Loftus, personal communication, April 6, 2000). Loftus's comments suggest, however, that we cannot assume that all the analyses in the articles we examined were the work of the articles' authors—the analysis might be owed to Loftus or his associate editors. Loftus's impression that authors relied essentially on traditional NHST is also consistent with our earlier conclusion.

Typical experiments reported in *Memory & Cognition* involve several factors, including some repeated measures. Such complex designs usually require calculation of several different *SEs* for assessment of the various factors. Loftus's experience suggested that the calculations required were not familiar to authors, and while he was editor “there was also quite a tempest about how to compute confidence intervals in repeated-measures designs” (G. R. Loftus, personal communication, April 6, 2000). In response, Loftus and Masson (1994) published an explanatory article to assist authors who needed to calculate CIs for within-subjects designs.

Loftus's experience suggests that authors need more than simple encouragement to change their statistical reporting practices. He suggested that lack of knowledge was a problem for authors. In Phase 4, we will present data from a survey of the authors who published under Loftus. We hold further discussion of possible difficulties of reform until after we have reported Phase 4 results.

Table 3 and Figure 1 show that the profile of post-Loftus articles differs from that of the Loftus articles in two principal ways: NHST-only articles increased from 32% to 49%, whereas interval-inclusive articles dropped from

45% to 27%. In accordance with this, a smaller percentage of articles with error bars was accepted by Gernsbacher than by Loftus.

For the period from 1998 to 2000, the percentage of NHST-only articles returned to pre-Loftus levels (around 50%). However, for interval-inclusive articles—articles including error bars and/or reporting CIs—the percentages post-Loftus (27%) were more than three times those for pre-Loftus articles (8%).

In essence, any legacy to *Memory & Cognition* from Loftus appears to be a somewhat greater use of error bars (and CIs) in comparison with that of the pre-Loftus period. As we noted earlier, Loftus's successor, Morton Ann Gernsbacher, made no explicit attempt to continue Loftus's reforms, nor has Colin MacLeod subsequently. Recent authors might not be aware of the proposals Loftus made, although it would be reasonable to expect that they might have been influenced by the style of articles published in *Memory & Cognition*—including those accepted by Loftus.

The conclusions we have drawn from Phase 1 are based on comparisons of articles accepted by three different editors. Differences in the profiles of articles might arise from different editorial styles and preferences, differences among the authors, or broader changes over time. We cannot be sure what proportions of the differences among the three time periods were attributable to Loftus's reform efforts. Loftus's influence on subsequent authors contributing to *Memory & Cognition* might not be expected to be strong, given that his policies were not explicitly continued. Presumably, Loftus would have had most direct influence on the authors whose manuscripts he accepted.

In Phase 2, we sought to control for possible author differences by examining articles published in other journals by the lead authors who had published in *Memory & Cognition* under Loftus. In Phase 1, we made an across-authors comparison in *Memory & Cognition* of the pre-Loftus, Loftus, and post-Loftus periods. In Phase 2, we made a similar, within-authors comparison in psychology journals generally. We looked in particular for evidence of continued reform practices in articles published by Loftus authors subsequent to acceptance of their *Memory & Cognition* articles.

PHASE 2

Authors Publishing in *Memory & Cognition* Under Loftus

In Phase 2, we examined the reporting practices of 260 unique lead authors who had published empirical articles in *Memory & Cognition* under Loftus. For each author, we conducted a PsycINFO database search to identify other empirical articles for which he or she was the lead author. A *before* article was one published by the author in any year preceding the year of his or her publication in *Memory & Cognition*. An *after* article was one published in any year after publication in *Memory &*

Cognition. Before and after articles published in *Memory & Cognition* were excluded. We selected the article published closest in time to the author's *Memory & Cognition* publication, subject to local availability.

In all, 170 before articles and 119 after articles published in 67 different journals covering a wide variety of areas of psychology were analyzed using the procedure for Phase 1. For 95 authors, we found both before and after articles. We present data from the matched sets of before articles, Loftus *Memory & Cognition* articles, and after articles by this group of 95 authors. The results closely resemble a comparison of all before articles, Loftus *Memory & Cognition* articles, and all after articles.

Results

The results are presented in Table 3, which includes the maximum *SE* for a single sample estimate of a proportion, as is relevant for comparisons between Phases 1 and 2. For matched sample comparisons within Phase 2, the *SE* of differences in proportions will typically be smaller than the *SE* values shown.

As in Phase 1, each article was classified in one of four types. The Phase 2 rows in Table 3 show, for the 95 authors, the percentage of each type for the before articles, the Loftus *Memory & Cognition* articles, and the after articles. The results are similar to the Phase 1 results: For example, in Phase 2 the proportions of articles using practices closely related to Loftus's recommendations (full Loftus, any CI, any bars) were very similar to those in Phase 1. Phase 2 results for Loftus *Memory & Cognition* articles are very similar to those for Phase 1 Loftus articles, as was expected, because the Phase 2 authors are a subset of the Phase 1 authors. It is notable that the Phase 2 results for before articles in a variety of journals are remarkably similar to those for the Phase 1 pre-Loftus articles in *Memory & Cognition*. However, there were somewhat fewer NHST-only after articles (37%) in Phase 2 than NHST-only post-Loftus articles (49%) in Phase 1, and more NHST-with-figure after articles (37%) in Phase 2 than NHST-with-figure post-Loftus articles (22%) in Phase 1. To put it another way, in Phase 1 the main increase from pre- to post-Loftus was in the percentage of interval-inclusive articles; the main decrease was in NHST-with-figure articles. In Phase 2, the major increase from before to after was also found in interval-inclusive articles; however, the decrease was found in NHST-only articles.

Consistency of article types. Comparison of the types of Phase 2 articles published before, during, and after Loftus illustrates some changes in reporting practices. The matched Phase 2 data allow also an examination of change and consistency in the data presentation practices of individual authors. We report the numbers of authors who changed their reporting practices, the ways in which they did so, and the numbers of authors who retained their style. This allows a further characterization of responses to Loftus's reform.

Table 4
Phase 2: Cross-Classification of Frequencies of Main Article Types for Before, Loftus, and After Articles

Before Article	After Article	Loftus Article			Row Total
		Interval Inclusive	NHST With Figure	NHST Only	
Interval inclusive	Interval inclusive	9*			9†
	NHST with figure	1			1
	NHST only				
	Noninferential				
NHST with figure	Interval inclusive	6			6
	NHST with figure	5	6	2	15
	NHST only	5	2	1	8
	Noninferential	1	1		2
NHST only	Interval inclusive	2	2	1	5
	NHST with figure	3	4	10	18
	NHST only	8	1	18	27
	Noninferential		1		1
Noninferential	Interval inclusive	1			1
	NHST with figure		1		1
	NHST only				0
	Noninferential			1	1
Consistent articles, Loftus and after		18	11	19	0

Note—Each of the 95 Phase 2 authors is counted once in the body of the table and once in the Row Total column, according to that author's pattern of article types, for the periods before, Loftus, and after. *Frequencies in bold are numbers of articles of consistent type for the periods before, Loftus, and after. †Frequencies in bold in the Row Total column are numbers of articles of consistent type for the periods before and after Loftus.

Table 4 shows frequencies for the cross-classification of types of before, *Memory & Cognition*, and after articles for Phase 2. Bold figures within the body of the table represent frequencies for consistent types across all three articles. The row totals (in the rightmost column) show the cross-classification of before and after articles, with consistent types of before and after practices shown in bold. The bottom row shows the number of consistent types of Loftus and after articles.

First, overall, 35% of lead authors published three articles of the same type. Over half of these (18 of 33) were NHST only. Otherwise, the data are best described in terms of the type of the before article. Nearly all the lead authors who published an interval-inclusive before article (9 of 10; see Table 4) published interval-inclusive *Memory & Cognition* and after articles.

Second, consider authors publishing an NHST-with-figure before article. These authors used a variety of practices. One fifth (19%) were consistent across all three articles, and 55% moved in a reform direction in that their Loftus articles were interval inclusive, but only 19% went on to publish an interval-inclusive after article later on.

Third, consider authors who published NHST-only before articles. Over one third (35%) published three NHST-only articles (Table 4); 41% published more reformed articles under Loftus, and 45% more reformed after articles. However, only 22% were more reformed in both their *Memory & Cognition* and after articles.

Discussion

In Phase 2, the profiles of before and after articles differed: There were more interval-inclusive (22% vs. 11%)

and fewer NHST-only (37% vs. 54%) after articles. This is broadly similar to the Phase 1 finding of more interval-inclusive (27% vs. 8%) and fewer NHST-with-figure (22% vs. 37%) articles post-Loftus in comparison with pre-Loftus. Overall, around one third of the Phase 2 lead authors kept a consistent style across their before, Loftus, and after articles. Consistency was greatest for authors "already converted" to Loftus's view—that is, those publishing interval-inclusive articles. The examination of consistency of statistical reporting style fails to show strong or lasting influences from Loftus's reform attempts.

The conclusions from Phase 2 are essentially the same as those from Phase 1. Loftus's success as editor resided in increasing the reporting of error bars and reducing the proportion of NHST-only articles. This success was limited, since very few articles complied fully with the spirit of Loftus's proposals; in the vast majority, NHST was still reported and used as the basis for interpretation. Discussion of error bars in the text provided virtually no evidence that authors had moved away from traditional practices; error bars were mentioned in the text in at most 6% of articles (under Loftus). In both phases, there was a general shift in a reform direction from before Loftus to Loftus, followed by a substantial but not total falling back after Loftus.

However, before the somewhat reformed profiles of after versus before articles and post-Loftus versus pre-Loftus articles can be attributed to authors' experience of publishing under Loftus, we need to consider whether there may have been broader changes over the relevant years. In Phase 3, we examined current statistical practice in contemporary psychology journals.

PHASE 3 Other Authors

In Phase 3, we compared contemporary reporting practices of authors who had not published in *Memory & Cognition* under Loftus with those of authors who had. To obtain a sample for comparison with the Phase 2 after articles, we examined the first empirical article following each after article in the same journal. Articles were rejected if the lead author had published in *Memory & Cognition* under Loftus, and the next suitable article was selected. The Phase 3 sample comprised 119 articles published between 1995 and 2000 in 38 different journals across experimental psychology.

The results for Phase 3, presented in Table 3, are strikingly similar to the results for the after articles described in Phase 2. Fully 95% of the articles included NHST. Only 5% of the Phase 3 authors included CIs in their article, only 1 of the 119 articles mentioned the error bars in the text, and only one article was full Loftus. Our picture of recent statistical practice in published experimental psychological research is that only around 22% of articles are interval inclusive, about 40% of articles rely on NHST without interval estimation or visual displays of data, and another 36% include conventional figures without error bars. Few contemporary authors use the reformed practices we studied, and the great majority report traditional NHST and rely on it as the primary means of drawing inferences from data.

PHASE 4 Survey of Authors Who Published in *Memory & Cognition* Under Loftus

Acceptance of Loftus's proposals was less than enthusiastic (Phase 1), perhaps because authors had difficulty in understanding and implementing his recommendations. In Phase 4, we sought opinions of contact authors of articles accepted by Loftus in *Memory & Cognition*. (We used the first contact author. In most cases, this was the lead author.) We asked the authors about their data presentation philosophies and practices, and their views of and experiences with Loftus's proposals. We aimed to identify any problems the authors may have had in under-

standing and implementing Loftus's recommendations. The authors were invited to provide further explanations of their yes–no responses, and a final open-ended question was asked about their opinions of the recommendations and any problems they might have experienced in meeting them. Seven features of the open comments were coded independently by two of the authors of this article.

The authors were surveyed using contact e-mail addresses published in *Memory & Cognition* from 1994 to 1998. Surveys were sent to three groups of authors, classified according to the type of the most recent article each author had published under Loftus. The few contact authors publishing noninferential articles were not included.

Table 5 shows that e-mail addresses were published for 83% of the contact authors but that 33% of the e-mails sent to those addresses were returned undelivered. (Note that Table 5 reports data for authors, whereas Tables 1 and 2 report data for articles.) The response rate was 42% (59 replies to 142 e-mails assumed delivered). Overall, 86% of the respondents chose to make additional commentary.

Results and Discussion

The results should be interpreted cautiously: The sample sizes are small, and the *SEs* relatively large (Table 5). Furthermore, the characteristics of the authors who chose to respond to the survey are unknown, apart from the type of article they published under Loftus. Bearing these caveats in mind, some large differences between NHST-only and interval-inclusive authors can be noted, and some group differences correspond to those that might be expected given the type of article published.

Table 6 shows that the respondents reported almost universal use of hypothesis testing, and there was strong agreement that hypothesis testing provided a scientific criterion for evaluating the results of their studies (84%) and was necessary for understanding these results (71%). In contrast, only around 30% agreed that *SEs* or CIs were more informative than hypothesis tests (Table 6). Most respondents thus disagreed with the predominant reform view that in many situations CIs are more informative than hypothesis tests because they provide information about the set of population parameter values that is consistent with the observed data.

Table 5
Phase 4: Data for Loftus Authors, by Type of Articles They Published in *Memory & Cognition*

Data	Types of Articles		
	Interval Inclusive	NHST With Figure	NHST Only
Number of authors in the Loftus period	117	56	80
Number with published e-mail addresses	90	52	69
Number of e-mails assumed received	58	38	46
Number of replies	22	14	23
Effective response rate	38%	37%	50%
Number of authors giving open comments	22	10	19
Percent respondents giving open comments	100%	71%	83%
Maximum <i>SE</i> for proportion of replies received	11%	13%	10%

Table 6
Phase 4: Percentages of Authors Giving Various Survey Responses, by Type of Article Published Under Loftus

Response	Types of Article ^a			Overall
	Interval Inclusive	NHST With Figure	NHST Only	
Usual data presentation practices for published articles				
Authors stated that they usually include:				
Figures	100	100	83	93
SEs	86	71	74	78
Hypothesis tests	95	100	100	98
CIs	19	14	14	16
Loftus's policy				
Authors stated that while preparing their <i>M&C</i> articles				
they were aware of Loftus's policy	86	79	61	75
they had followed Loftus's recommendations	68	21	0	32
Loftus's guidelines were relevant to their studies	86	50	30	56
they had difficulty calculating the statistics Loftus suggested	15	14	19	16
they had difficulty preparing the figures	14	14	19	15
figures made the findings easy to see	85	79	56	74
figures without error bars were clearer than those with error bars	24	50	27	32
hypothesis testing was necessary to understand the results	60	77	78	71
hypothesis testing provided a scientific criterion for the results	75	77	96	84
SEs or CIs were unnecessary	9	21	33	20
SEs or CIs were more informative than hypothesis tests	50	23	6	29

Note—Sample sizes vary slightly from question to question due to some nonresponses. *M&C*, *Memory & Cognition*; *SE*, standard error; *CI*, confidence interval.

Reported practices and published articles. Almost all the respondents (93%) claimed that they usually publish figures (Table 6). By contrast, we found 37%–49% of recent articles (Table 3: Phase 1 post-Loftus, Phase 2 after, and Phase 3) included no figures (NHST-only). Our Phase 4 sample may have self-selected in favor of authors more supportive of reform, although 23 respondents had published NHST-only Loftus articles. Our respondents may have given an inaccurate account of their customary practices.

Group differences in authors' stated analytic philosophies and views about the relevance of Loftus's guidelines reflected the statistical presentation in their *Memory & Cognition* articles only to some extent. For example, just one of the NHST-only authors agreed that CIs or SEs provided more information than hypothesis testing. On the other hand, 56% of NHST-only authors agreed that "figures made the findings easy to see," although they had included no figures in their own Loftus articles. These authors may have produced figures that were subsequently not published, or, despite the specific wording of the questions, they may have been responding about their practices more generally. These overall patterns of findings lead us to offer the speculative interpretation that NHST is so deeply entrenched in psychology that authors fail to realize how often their final data presentations and interpretations reduce to NHST. The entrenched NHST mindset may explain at least some of the patterns of response we observed and may constitute one of the major obstacles to statistical reform.

Difficulties with Loftus's recommendations. Practical problems that authors mentioned included (1) knowing

how to calculate SEs or CIs, (2) obtaining appropriate statistics from computer packages, (3) using standard statistical packages to produce figures with appropriate error bars, (4) representing error bars for more complex experimental designs with interactions, (5) achieving reasonable resolution in figures so that values could be read, and (6) interpreting figures with poor and inconsistent labeling of bars. A number of comments related to obtaining appropriate SEs (or CIs) in complex designs, including those with one or more repeated measures. Some of the comments suggested that some of the authors did not understand that SEs and CIs can provide a basis for making statistical inferences. However, the proportion of authors reporting difficulties with Loftus's recommendations was relatively low (15% and 16% for SEs and CIs, respectively).

Fully 57% of the respondents who commented openly made statements supporting Loftus's policy, and some of these were quite strong. Some authors agreed on the need to move away from routine reliance on NHST; they argued that replication and theoretical corroboration were important. Authors also had a number of concerns about the policy, including how readers would interpret error bars and effect sizes, how accurately figures could be read, and readers' poor understanding of the debate that had motivated Loftus. Others stated that there was a general expectation to see NHST, that it was traditional and standard, and that change would be difficult; the need for a statistical method that would allow inferences about higher order interactions remained. Some argued that there was an essential equivalence between hypothesis testing procedures and CIs. One author suggested

that Loftus's approach left readers to carry out an "intuitive" hypothesis test—a task that would be difficult if findings were not clear-cut.

The nature of the practical problems mentioned and concerns about Loftus's policy are consistent with Loftus's perceptions. However, Loftus's impression suggests that a much larger proportion of authors might have had such problems and concerns than we found in the authors' self-reports. The low level of reported difficulties may reflect the generous practical assistance that Loftus and his associates gave to authors, and perhaps some self-selection by more confident or competent authors to respond to our survey.

GENERAL DISCUSSION AND CONCLUSIONS

Our main conclusion, based on the results of Phases 1 and 2, is that Loftus's efforts led to fairly substantial increases in the use of figures with bars in *Memory & Cognition* during his editorship, but that afterward fewer such figures were published, although the rate remained higher than it was pre-Loftus. At no stage, however, were CIs or other error bars often used to support the interpretation of data, and NHST remained throughout the greatly dominant approach to inference. A substantial proportion of psychologists will need to make major changes to their statistical practices if the latest recommendations of the APA are to be implemented successfully.

We investigated whether the changes we observed were particular to *Memory & Cognition* (Phase 1) or to Loftus authors (Phase 2), or were general across psychology publications (Phase 3). In discussing the Phase 2 results, we noted the close agreement between the pre-Loftus articles in *Memory & Cognition* and the before articles in a broad range of psychology journals. None of the pre-Loftus articles could have been influenced by Loftus's (1993a) policy and, although 55% of the before articles were published in 1994 or later, none appeared in *Memory & Cognition*, so Loftus's influence was unlikely to be a factor. The Phase 1 pre-Loftus articles and the Phase 2 before articles thus provide an estimate of practices across experimental psychology in the early to mid-1990s. Phase 3 gives the general picture for the late 1990s and 2000. Between these two periods, there was an increase in interval-inclusive articles from 9% to 22% and a decrease in

NHST-only articles from 53% to 40%. Therefore, in terms of the article types we analyzed, the general practices of psychologists have shown a modest shift in a reform direction over the last decade: Somewhat more articles now include figures with bars or CIs in some form.

We now consider our results for Phases 1, 2, and 3 together. The profile of after articles of Loftus authors (Phase 2) closely matches that of psychology articles generally (Phase 3). In a comparison between Phase 1 and Phase 3, the profile of post-Loftus articles in *Memory & Cognition* can scarcely be said to be reformed: There are slightly *more* NHST-only articles post-Loftus (Phase 1, 49%) than in psychology journals generally (Phase 3, 40%). We conclude that the change from pre- to post-Loftus (Phase 1) and that from before to after articles (Phase 2) are both similar to the change in psychology generally over the same period.

Loftus's efforts may have contributed to this general change, and we have no way of assessing directly the extent of any such contribution. However, it would be reasonable to expect any persisting influence of Loftus's policies to be shown most strongly in later volumes of *Memory & Cognition* (Phase 1) and/or in the subsequent practices of the authors whose works he published in that journal (Phase 2). In both cases, however, the changes are very similar to those described in Phase 3 for psychology generally, and so we are unable to identify any additional effect on authors exposed to Loftus's policy. We conclude that Loftus's influence was most evident in the articles published in *Memory & Cognition* under his editorship. There is little evidence that Loftus's efforts had persisting or more general influence.

Loftus's experiment should not, however, be dismissed as merely a well-intentioned failure because it appears not to have substantially changed publishing practices beyond its immediate domain. Reformers have undertaken very few empirical investigations of how reform might be achieved in practice. The Loftus experiment stands nearly alone as a large-scale, sustained reform attempt in psychology. Loftus published advocacy and exhortation (e.g., Loftus, 1993b, 1996), presented a strong and reasoned editorial policy (Loftus, 1993a), and also provided practical assistance—for example, through the techniques described by Loftus and Masson (1994) and the assistance given to numerous authors who had difficulty with *SEs* and CIs. This important case study pro-

Table 7
Number of Citations of Selected Loftus Publications, by Year of Citation

Publication	1994	1995	1996	1997	1998	1999	2000
Loftus (1993a)	0	5	5	6	4	2	2
Loftus (1993b)	1	7	4	5	6	6	4
Loftus & Masson (1994)		5	15	30	42	41	31
Loftus (1996)				4	6	7	7

Note—Data are from the annual publications of the *Social Sciences Citation Index* (e.g., Institute for Scientific Information, 2000).

vides a salutary lesson for reform: Change was limited and difficulties great, despite large and sustained efforts by Loftus and his editorial team.

The Loftus experiment may have some continuing influence. It caused many hundreds of researchers to encounter reform issues; perhaps they will be more readily influenced by future reform efforts. Citation counts (see Table 7) show that publications central to the experiment continue to be noticed, especially Loftus and Masson (1994), which gave an exposition of techniques. Although the results of Loftus's efforts seem limited, Loftus may have been ahead of his time: Statistical reform in psychology has recently been receiving more attention than before.

Our study suggests that a number of strategies are required if recent policy recommendations (TFSI; APA, 2001) are to be successfully implemented. For example, improved software, new textbooks, and new statistics curricula may be needed (Friedrich, 2000) to help authors produce and understand the "new" statistics and visual displays. Examination of figures with bars published in *Memory & Cognition* under Loftus reveals a wide array of representations—some ingenious, but some visually and conceptually dreadful. There are few conventions and little guidance for authors wishing to report CIs or to show error bars in figures depicting results from the complex designs that typify experimental psychology or, more generally, wishing to use CIs to inform their interpretations of data. Loftus and Masson's (1994) article responded to this need during Loftus's reform (see also Estes, 1997), and articles by Cumming and Finch (2001, 2003), Loftus (2002), and Masson and Loftus (2003), as well as a special issue of the *Canadian Journal of Experimental Psychology* (Masson, 2003), also address these issues.

Referring to Table 3, pooling of results for pre-Loftus and before *M&C* articles suggests that early in the 1900s about 9% of articles in psychology journals were interval-inclusive. Results for after *M&C* and Phase 3 suggest that toward the end of the decade about 22% were of this type. Loftus (personal communication, April 6, 2000) has also suggested that practices may be changing: "As we enter the new millennium, . . . I seem to see many more confidence intervals appearing in journal articles and in talks (e.g., at Psychonomics). Also, there seem to be fewer unwieldy tables, fewer p values, and more graphs." A more recent comment is from Peter Dixon (personal communication, May 16, 2003):

As editor of *Canadian Journal of Experimental Psychology*, I insist that all graphs have appropriate error bars. Although many papers are submitted with barless graphs, I have encountered little resistance to suggestions that error bars be added, and many authors have been willing to add graphs where none were included before. I don't recall that being nearly as true during my brief stint as one of Loftus's associate editors.

Our findings suggest, however, that any change in statistical practices has been small. Our definition of interval-inclusive articles did not require exclusion of NHST. In-

deed, in all our analyses we found very few empirical articles that did not include NHST. Our analysis of the textual references to error bars provided little evidence that authors had incorporated information from the bars into the statistical arguments that they made; inclusion of bars and/or CIs appeared to make little difference to the NHST practices routinely used. NHST remains by far the dominant method for drawing inferences from data. Our results suggest that more than editorial initiative is needed if these entrenched practices are to be changed.

REFERENCES

- AMERICAN PSYCHOLOGICAL ASSOCIATION (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- BAKAN, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423-437.
- CARVER, R. (1978). The case against significance testing. *Harvard Educational Review*, *48*, 378-399.
- CARVER, R. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, *61*, 287-292.
- COHEN, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.
- COHEN, J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, *49*, 997-1003.
- CUMMING, G., & FINCH, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational & Psychological Measurement*, *61*, 530-572.
- CUMMING, G., & FINCH, S. (2003). *Inference by eye: Confidence intervals, and how to read pictures of data*. Manuscript submitted for publication.
- ESTES, W. K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review*, *4*, 330-341.
- FINCH, S., CUMMING, G., & THOMASON, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational & Psychological Measurement*, *61*, 181-210.
- FINCH, S., THOMASON, N., & CUMMING, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory & Psychology*, *12*, 825-853.
- FRIEDRICH, J. (2000). The road to reform: Of editors and educators. *American Psychologist*, *55*, 961-962.
- GERNSBACHER, M. A. (1998). Editorial. *Memory & Cognition*, *26*, 1.
- HAMMOND, G. [R.] (1996). The objections to null hypothesis testing as a means of analysing psychological data. *Australian Journal of Psychology*, *48*, 104-106.
- HARLOW, L. L., MULAİK, S. A., & STEIGER, J. H. (EDS.) (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- INSTITUTE FOR SCIENTIFIC INFORMATION (2000). *Social sciences citation index*. Philadelphia: Author.
- INTONS-PETERSON, M. J. (1990). Editorial. *Memory & Cognition*, *18*, 1-2.
- KIRK, R. E. (1996). Practical significance: A concept whose time has come. *Educational & Psychological Measurement*, *56*, 746-759.
- LOFTUS, G. R. (1993a). Editorial comment. *Memory & Cognition*, *21*, 1-3.
- LOFTUS, G. R. (1993b). A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers*, *25*, 250-256.
- LOFTUS, G. R. (1996). Why psychology will never be a real science until we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161-171.
- LOFTUS, G. R. (2002). Analysis, interpretation, and visual presentation of experimental data. In J. Wixted & H. Pashler (Eds.), *Stevens' handbook of experimental psychology: Vol. 4. Methodology in experimental psychology* (3rd ed., pp. 339-390). New York: Wiley.
- LOFTUS, G. R., & MASSON, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476-490.

- LYKKEN, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*(3, Pt. 1), 151-159.
- MASSON, M. (2003). Introduction to the special issue on alternative methods of data interpretation. *Canadian Journal of Experimental Psychology*, *57*, 139.
- MASSON, M., & LOFTUS, G. R. (2003). Using confidence intervals for graphically based interpretation. *Canadian Journal of Experimental Psychology*, *57*, 203-220.
- MEEHL, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting & Clinical Psychology*, *46*, 806-843.
- NICKERSON, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-301.
- OAKES, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, U.K.: Wiley.
- SCHMIDT, F. L. (1992). What do data really mean? Research findings, meta-analysis and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173-1181.
- SEDLMEIER, P., & GIGERENZER, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309-316.
- TUKEY, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83-91.
- WILKINSON, L., & THE TASK FORCE ON STATISTICAL INFERENCE, APA BOARD OF SCIENTIFIC AFFAIRS (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

(Manuscript received August 5, 2003;
revision accepted for publication January 8, 2004.)