COLLEGE STUDENTS' INTERPRETATION OF RESEARCH REPORTS ON GROUP DIFFERENCES: THE TALL-TALE EFFECT

THOMAS P. HOGAN
University of Scranton
Thomas.Hogan@Scranton.edu

BRIAN A. ZABOSKI
University of Scranton
ZaboskiBrian@gmail.com

TIFFANY R. PERRY
University of Scranton
TiffRPerry@gmail.com

ABSTRACT

How does the student untrained in advanced statistics interpret results of research that reports a group difference? In two studies, statistically untrained college students were presented with abstracts or professional associations' reports and asked for estimates of scores obtained by the original participants in the studies. These estimates were converted to inferred effect sizes and compared with the actual effect sizes. Inferred effect sizes substantially overestimated actual effect sizes for all reports, a phenomenon dubbed the tall-tale effect. The effect was obtained with a variety of reports and statistics. The tall-tale effect could be controlled somewhat with simple changes in wording. This finding suggests a program of research which would better calibrate inferences with those actually obtained in the research.

Keywords: *Statistics education research; Effect size; Research reporting; Interpreting research*

1. INTRODUCTION

How do people make sense out of the results of empirical research? The process typically begins with words, often framed as questions. For example, is there a difference between group A and group B on variable X? Or, is this therapy effective? Then come numbers: data get collected and analyzed. Finally, the numbers get translated into words. Yes, there is a difference between groups A and B. Yes, the therapy is effective. This words-numbers-words sequence is surely an oversimplification, but it also gets at the heart of the process. The "words to numbers" part of the sequence is a matter of research design. The "numbers to words" part of the sequence is a matter of inference and it presents a daunting challenge both for researchers and public consumers of the research.

Within the behavioral sciences, a key mechanism for drawing inferences from research data has been null hypothesis significance testing (NHST), which has dominated research reporting in the behavioral sciences for nearly a century. For the past 20 years,

researchers have proposed varied alternatives to and elaborations of reporting procedures to overcome obvious shortcomings of NHST (e.g., American Psychological Association, 2010; Kline, 2004; Odgaard & Fowler, 2010; Wilkinson and Task Force on Statistical Inference, 1999). Some of the proposed revisions present rather simple methods such as confidence intervals and Cohen's (1988) benchmarks for effect size, while others present more sophisticated developments of traditional procedures (e.g., Aguinis et al., 2010; Browne, 2010; Erceg-Hurn & Mirosevich, 2008; Kelley & Preacher, 2012; Killeen, 2005; McGraw & Wong, 1992; Tryon, 2001; Wagenmakers, 2007). The terms "practical significance" (Kirk, 1996) and "clinical significance" (Jacobson & Truax, 1991; Odgaard & Fowler, 2010) have become increasingly popular in these discussions. Hinting at the same down-to-earth application, McGraw and Wong (1992) even named their new statistic the "common language effect size statistic". All such efforts have aimed at improving interpretation and understanding of research results.

For the frequently encountered two-group comparison (e.g., treatment versus control, or male versus female), a measure of effect size (ES) provides an important adjunct to or even replacement (when accompanied by a confidence interval) for testing the null hypothesis of "no difference." The most common measure of ES takes the difference between means for the two groups divided by some index of standard deviation; hence, it is sometimes referred to as the standardized mean difference. Basically, the ES expresses the average difference in relation to within-group variability. (In various formulations, group variability may be defined by the control or base group standard deviation, the average of the two standard deviations, or the square root of the pooled variances; see Chapter 3 of Grissom & Kim, 2012.)

Emphasis on using measures of ES has risen rapidly as a mechanism for reporting and understanding results of empirical research (Grissom & Kim, 2012; Kline, 2004), especially in social/behavioral and biomedical sciences. For example, a measure of ES is nearly always reported in randomized clinical trials (RCTs), as illustrated by requirements in Cochrane reviews of biomedical treatments (see Higgins & Green, 2011, especially Chapter 11). Measures of effect size underlie two other emergent statistical procedures: statistical power analysis (Cohen, 1988) and meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009).

The interpretation of NHST, confidence intervals, and measures of effect size requires advanced statistical training possessed by a very limited audience. Even among highly trained professionals, however, knowledge of effect size measures is hardly universal. In a survey of APA Division 12 (Clinical Psychology) members, where 98.7% of respondents held a doctorate in psychology, self-rated knowledge of measures of effect size was only 3.74 on a scale of 1-5, where 3 was "moderate" and 13% of respondents rated themselves as only 1 or 2 (Berke, Rozell, Hogan, Norcross, & Karpiak, 2011). Browne (2010, p. 30) recounted that he "… once asked a well-published medical researcher what $p < 0.05$ meant to him. He said: 'It means that everyone on [treatment] X did better than everyone on Y'". Further exacerbating the problem of interpretation, Bakker and Wicherts (2011) found that for a random sample of articles in PsycINFO, only about 20% included any measure of effect size to accompany a χ^2, t, or F-test. Earlier, Kirk (1996) reported that use of effect size measures accompanying inferential statistics in four APA journals in 1995 ranged from a low of 12% to a high of 73%

An analogous issue arises primarily in a medical context where the numbers refer to probabilities or percentages (e.g., Bryant & Norman, 1980; O'Brien, 1989). For example, what does it mean when a medical doctor or a pharmaceutical company says "a *frequent* side effect is …" or "it is *likely* that …"? In this context, researchers have made considerable progress in mapping the number-word relationships (e.g., Brun & Teigen,

1988; Budescu & Wallsten, 1985; Budescu, Weinberg, & Wallsten, 1988; Mosteller & Youtz, 1990; Renooij & Witteman, 1999). In these applications, the numbers and their verbal counterparts necessarily fall in a restricted range, where the numbers define either probabilities (from 0–1) or percentages (from 0–100). O'Brien (1989) obtained probability ratings from general practitioners for 23 words/phrases ranging from "never" to "certain" and including such terms as "probable", "possible", and "moderate risk." Participants also rated the ambiguity of the terms. Bryant and Norman (1980) had physicians express the probability associated with 30 terms (e.g., "probable," "always," and "low probability"). Mosteller and Youtz (1990) had science writers state estimates of probabilities for 53 expressions (e.g., "very likely," "seldom," and "not very often"). Mosteller and Youtz also summarized 19 other studies using these probabilistic terms.

Three generalizations emerge from these studies of word-number correspondence. First, people do not interpret the terms with lexical precision. For example, in people's actual interpretation, "certain" does not mean a probability of 1.00; the term actually corresponds to (in people's translation of words to numbers) probabilities of .91, .95, and .98 in various studies (see Mosteller & Youtz, 1990, Table 1). Second, some terms carry such wide ranges of meaning for different individuals that use of the terms is inadvisable; for example, "significant chance" has a probability of .23 at the first quartile and a probability of .70 at the third quartile (see O'Brien, 1989, Table 1). Third, despite the first and second generalizations, it is possible to "scale" words to convey differences in quantitative connotations with reasonable accuracy and consensus. For example, one term (say, "very likely") is quite consistently rated as expressing a higher probability than another term (say, "often"), and both terms have relatively small ranges around their central ratings (see Mosteller & Youtz, 1990, Table 2). In a somewhat different take on the same matter within a medical context, Gigerenzer and colleagues (Gigerenzer, 2002; Gigerenzer & Edwards, 2003; Gigerenzer, Gassmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007) have articulated problems surrounding the interpretation (and often misinterpretation) of rates, ratios, and odds.

Expressing the range of meanings on a 0-100% scale simplifies the analysis of the words-to-numbers or numbers-to-words problem. In contrast, results of behavioral research studies usually find original expression in a quantitative form such as means, standard deviations, or correlation coefficients. Application of inferential statistics (e.g., *t* or *F* tests) adds reference to statistical significance. Results of such studies usually translate the quantitative material into a verbal form that refers to a group difference (e.g., this group did better than that group) with or without reference to statistical significance. Some full articles include measures of effect size; the rare summaries that do require advanced training for meaningful interpretation.

Persons in training, but lacking advanced statistical training, constitute an important audience for behavioral research findings. What conclusions do such persons reach based on behavioral research reports they are likely to encounter? The question posed here for behavioral research is analogous to the question of interpreting probabilities in a medical context. The present study extends the words-numbers-words relationships from medical-like probabilities to the context typically encountered in reports of behavioral research: inferences made by intelligent laypersons of results summarized in abstracts or news-like reports of research studies on group differences.

2. METHOD

Two studies were conducted that shared a general methodology in terms of stimulus material, type of participants, and method of responding. These shared characteristics are

described first, followed by details of each study. The following criteria guided selection of research reports (stimulus materials) for presentation to participants. First, the report had to present a two-group contrast in an experimental or quasi-experimental design. Second, the report had to use a dependent variable that could be easily understood by laypersons – for example, a score on a test or a rating scale. Third, the report had to describe a difference between the groups on the dependent variable; exact wording for the magnitude of the difference varied among the reports. Fourth, the original report had to provide the actual effect size or sufficient data to allow calculation of effect size for the group difference. Some reports were in the form of original abstracts as presented in journal articles; others were brief reports from publications of professional associations (American Psychological Association or Association for Psychological Science).

College students in introductory psychology courses served as respondents in both studies. The students represent typical consumers of behavioral research reports. The students have had some exposure to behavioral research methods but do not have advanced statistical training. They are intelligent individuals for whom reading and interpreting reports of behavioral research is a common requirement in their introductory psychology course. As described further under each study, they came from a great variety of major fields of study. Although they were students in a psychology course, few of them were psychology majors.

Most of the analyses presented below employed a measure of *inferred effect size*. Respondents first read the abstract or brief report of a study and then estimated the scores obtained by samples of cases in each of the two groups involved in the report (see sample report form presented later). For each respondent, the mean and standard deviation of the estimated scores for each group were calculated and these data yielded the effect size defined as $d = (M_1 - M_2)/((SD_1 + SD_2)/2)$, following Cepeda's (2008) recommendation on combining standard deviations. (Trial runs of the data using the square root of pooled variance resulted in no substantive differences from the procedure just described.) This version of *d* is referred to in the present paper as the inferred effect size, in contrast to the actual effect size in the report. We emphasize: respondents did not estimate effect sizes directly; they did not have the training to do so. Rather, they estimated scores or other measures on the dependent variable in the research and the authors calculated inferred effect sizes from these estimates. Participants wrote brief summaries of the studies in their own words before providing estimated scores. These summaries allowed for identification of cases where it appeared a participant did not really understand the study and, therefore, raised questions about the validity of the participant's estimated scores.

To illustrate what participants received in their packets, we offer an Appendix with full text of material for two of the reports and other instructions for the variety of tasks presented in the two studies, with further details given below for each study.

2.1. METHOD OF STUDY 1

Material Stimulus material for Study 1 consisted of the following abstracts or other brief reports of behavioral studies, each preceded by a brief descriptor used later. All four abstracts or news-like summaries can be found in the sources cited below.

1. Seeing Red: a 391-word news-like report from the APA *Monitor on Psychology* (Cynkar, 2007) for an article originally appearing in *Journal of Experimental Psychology: General* (Elliot, Maier, Moller, Friedman, & Meinhardt, 2007). The report stated that seeing the color red before taking a test depressed scores in comparison with individuals in the control group. The "scores" completed by participants for this study were number of anagrams (ranging from 0–15) completed in 5 minutes.

2. Full Access: the journal abstract (112 words) from Hove and Corcoran (2008) for an article in *Teaching of Psychology*. The abstract reported that giving students full, web-based access to course lectures improved students' grades in comparison with grades for students without such access, where grades were on a 300-point exam.

3. Glitzy Science: a 74-word report in the APA *Monitor on Psychology* (Novotney, 2009) for an article originally appearing in the *Journal of Experimental Psychology: Applied* (Mayer, Griffith, Jurkowitz, & Rothman, 2008). The report stated that using "glitzy" materials (cartoons, multimedia presentations, etc.) led to lower test scores on a transfer test. The report also referred to a retention test but did not find any difference on this variable. Participants provided estimates on both variables: retention on an 18-point test and transfer on a 13-point test.

4. Clean Scents: the abstract (83 words) for an article by Liljenquist, Zhong, and Galinsky (2010) appearing in the Association for Psychological Science (APS) *This Week in Psychological Science* for the full article appearing in *Psychological Science.* The report stated that being in a clean-smelling room leads to greater volunteering behavior, as reported on a 7-point Likert scale.

Participants Participants were 40 undergraduate students from a 5000-enrollment institution in the northeastern United States with institutional mean SAT scores (Critical Reading + Mathematics) of 1120 (69^{th} percentile on current national norms). Participants were predominantly female (32 female, 7 male, 1 unreported) and freshmen (80%), with 95% in the traditional 18-22 year age range (M = 20.25, SD = 6.87), from multiple sections of an introductory psychology course, and divided by categories of majors as follows: 30% allied health fields, 18% physical sciences and mathematics, 18% social sciences, 15% education, 8% humanities, and 12% undecided. Participants' self-reported grade point averages ranged from 2.45 – 3.94 (M = 3.16, SD = 0.41). Participants received partial credit for a research participation requirement in the introductory psychology course, volunteering to partake in this project as opposed to several others available to them.

Procedure Participants completed the tasks in four groups of 9-12 students each. They worked individually, without group consultation. The stimulus reports were presented in four counterbalanced orders in the four different groups. Participants first read an abstract or summary as described above, then wrote a summary in their own words, and finally entered the scores they thought were obtained by a random selection of 10 cases from each group on a separate sheet. As an illustration of the directions to participants, directions for estimating scores for the Clean Scents report read as follows:

> In this study, participants (some in a scented room, some in a non-scented room) completed ratings of their interest in volunteering for future Habitat efforts.
> They made their ratings on a 7-point scale from:
> 1 = LOW interest to 7 = HIGH interest.
> Let's say we take 10 students from each condition: scented room and non-scented room. Based on the report of results given in the abstract above, list what ratings you think they made about their interest in volunteering.
>
> | Students in Scented room | Students in Non-scented room |
> |---|---|
> | 1. _____ | 1. _____ |
>
> [*Each column extended to 10 members in the related group.*]

Similar directions were used with each of the four reports, customized for the particular dependent variables in the report. Instructions for writing the respondent's own summary were: "Now, in your own words, please write a brief, simple summary of what was done in this study and what the results were. You can refer back to the report, if you wish." Five blank lines followed these instructions. Examination of responses in this part of the task provided a type of internal validity check. For example, because all of the reports clearly said there was a difference between groups, if a participant's written summary said there was no difference or reversed the direction of the difference that would certainly raise a question about the validity of the participant's response. In fact, there were very few such questionable cases, as noted below for each dependent variable.

After stimulus reports were distributed and oral directions given about the nature of the task, participants were asked if they had any questions about how to proceed. Participants readily seemed to understand the nature of the task. The procedure advanced at a rate of approximately 10 minutes per report, the entire session lasting about 40 minutes, including completion of informed consent and demographic information forms. Each participant worked on his or her own, that is, without discussion with other group members.

Data Analysis Preliminary data screening resulted in the following reduction in numbers of cases used for final analyses. One participant from the original pool of 41 was eliminated from all analyses due to aberrant responses– the person did not appear to take the task seriously. One case omitted responses, apparently inadvertently, for the Full Access report, as did another case for the Glitzy Science retention variable. One case for the Clean Scents study, one for the Full Access study, two for the Glitzy Science retention variable and one for the Glitzy Science transfer variable did not seem to understand the studies (inferred from participants' written summaries, as judged by consensus of two authors after reviewing the written summaries), prompting deletion of these data. All missing or deleted cases combined constituted only 2.5% of the total data array and would have no effect on substantive conclusions. Thus, final numbers of cases for the analyses were as follows: 40 for Seeing Red, 39 for Clean Scents, 38 for Full Access and Glitzy Science transfer variable, and 37 for Glitzy Science retention variable. For estimated scores in the Clean Scents study, one participant assigned scores of 5 to everyone in one condition and scores of 3 to everyone in the other condition, resulting in standard deviations of zero. To allow calculation of inferred effect size for this case, the research team replaced the zero standard deviations with median standard deviations for all other participants on those variables.

From the estimated scores for 10 cases in each group, estimated *d* was computed for each participant on each report (the inferred effect size). From original full articles for stimulus materials, actual effect sizes were obtained (either as reported in the article or as calculated from data in the article) as follows. The Seeing Red study had three conditions: Red with $n = 19$, Green with $n = 27$, and Black with $n = 25$. The original report gave effect sizes as eta-squared for Red versus Green (.08) and Red versus Black (.06). For analyses presented below, these values were converted to *d*-values using formulas from Cohen (1988) and Wolf (1986) and then averaged. The Full Access original report provided a *d*-value, specifically identifying it as Cohen's *d*. Total *n* for the Full Access report was 204 (inferred from *df* given for the *t*-test) but numbers of cases in each subgroup were not given. The original articles for Glitzy Science and Clean Scents also gave *d*-values but did not state how they were calculated. Total *n* for the Clean Scents report was 99 (inferred from *df* given for the *t*-test) but numbers of cases in each subgroup were not given. As noted by Grissom and Kim (2012), it is not unusual for

reports to give a measure of effect size labeled as *d* without saying which of several different methods of calculation was actually used.

2.2. METHOD OF STUDY 2

Study 2 had four purposes. First, it attempted to replicate selected results from Study 1 with an independent sample. Second, it investigated whether similar results would be obtained from a media-type report and from a journal-based abstract of a research report. Third, it sought to show that overestimation of group differences would occur when data were in percentage form as well as in the form of means derived from estimated scores. Fourth, it attempted to show that the overestimation of group differences could be controlled to some extent by very simple changes in wording of an abstract.

Material

1. Seeing Red: The Seeing Red study, from Elliot et al. (2007), reported that using the color red on student ID numbers depressed test scores for those students. A randomly-selected half of the participants received the abstract from the journal in one version (R1), while the other half received a semi-popular audience report of the results appearing in the APA's *Monitor on Psychology* for the other version (R2), the latter having been used in Study 1. This contrast helped to answer the question of whether a semi-popular report (as used in Study 1) and an official abstract yield similar levels of inferred effect size. An effort to contrast the official journal abstract and the news-like report for the Glitzy Science report failed because the official abstract did not use the term "glitzy science," thus leading to confusion as to what was being compared, as revealed by participants' written summaries. The news-like (R2) version allowed for direct replication of this report used in Study 1. Estimated scores were given on the same 0–15 scale as in Study 1.

2. Violent Media: the journal abstract from Bushman and Anderson (2009) on desensitizing effects of violent media. The study had groups play either violent or nonviolent video games and then compared subjects' reactions on several measures. Participants provided estimates of scores on two of the variables: time taken (up to 180 seconds) to help a confederate victim and ratings (on a 1–10 scale) of the severity of a fight heard outside the lab; and for estimated percentages of participants who (a) helped the victim and (b) reported hearing the fight. See the Appendix for instructions to participants on recording responses for this task. The Violent Media study was intended to help answer the question of whether exaggerated inferred effect sizes arise with percentage data as well as with means of estimated scores.

3. Simulated Abstract: a 177-word abstract of a simulated report, created specifically for this study, on the effect of mental activity on short term memory (as measured on a 150-point test) in elderly subjects. One version (S1) concluded with a statement attributing "significantly higher" scores on a short-term memory test to those who completed a daily puzzle. The other version (S2) reported "slightly higher" scores. The crucial terms (significantly higher and slightly higher) appeared in a boldface font. The contrasting terms helped to determine whether simple word changes would impact people's inferred effect sizes. A simulated abstract was used so as not to tamper with the wording of an existing copyrighted abstract. However, the simulated abstract was prepared to be typical of a journal abstract.

Participants Participants included 88 young adults from the same institution as in Study 1, cooperating through their introductory psychology course. These were not the same students as in Study 1. Demographic data were as follows: 93% in the age range 18-

22 years (M = 18.70, SD = 1.48), 76% female (66 female, 21 male, 1 unreported), 69% freshman year, and with those categories of majors: 36% allied health fields, 18% undecided, 16% physical sciences and mathematics, 12% social sciences, 9% humanities, 8% education, and 1% business. Self-reported GPAs were not obtained in Study 2. Participants received partial credit for a research participation requirement.

Procedure As in Study 1, participants received a packet of material with a description of the above studies (an abstract or news-like summary), each followed by a page instructing participants to write a brief summary (same instructions as in Study 1). The next page or pages asked the participant to provide estimates of the original data obtained in the study. Alternate versions of reports (S1 and S2, R1 and R2) were arranged in random order for distribution to participants. There were 8 sessions with 5-19 individuals per session. The groups, tested in a classroom setting, moved through the reports in unison, allowing sufficient time for each person to complete a report before the group moved to the next report. As in Study 1, the groups completed the task at a rate of approximately 10 minutes per report, with the entire session lasting about 40 minutes, including completion of an informed consent form and a demographic information form. Again, participants worked individually, without group consultation.

Data Analysis Preliminary data screening proceeded for Study 2 much as it did for Study 1. Participants' self-prepared written summaries were scanned for aberrant responses and for misunderstanding of studies. These procedures resulted in the following reduction of cases used for final analyses. One participant from the original pool of 88 was eliminated from all analyses due to consistently aberrant responses. Several other cases presented anomalous responses to selected parts of the exercise, requiring elimination or adjustment of responses, including median substitutions for some responses. Final Ns for analyses reported below were as follows: for the Simulated Abstract, 46 for S1 ("significant difference") and 41 for S2 ("slight difference") for a total of 87; for Seeing Red, 43 for R1 (the official abstract), 42 for R2 (the news-like report) for a total of 85; for the Violent Media report, 87 for the time variable and both percentage variables and 86 for the severity variable. As in Study 1, the inferred effect size (d) was computed for each participant on each report and the actual effect sizes came from original articles, either as reported or as calculated from data in the article. The actual effect size for the Full Access report was described under Study 1. Effect sizes for the Violent Media study were given as d-values in the original article but without specifications of how they were calculated.

3. RESULTS

Results for both studies concentrate on comparisons of actual results as reported in original studies with inferences made by participants (in the form of inferred effect sizes) based on participants' reading of summaries of the studies. Additional analyses explore matters related to these comparisons.

3.1. RESULTS OF STUDY 1

Table 1 shows actual and inferred effect sizes for the four reports in Study 1. One report (Glitzy Science) included two dependent variables (transfer and retention); one had a significant difference (transfer), the other did not (retention); however, the report summary referred only to the transfer dependent variable. Inferred effect sizes hovered

around 2.0–2.5 regardless of the effect size in the original report. Typically, the inferred effect size was more than four times greater than the actual effect size.

Table 1. Actual and average inferred effect sizes for four reports in Study 1

Report	Actual *d*	Inferred *d* *M* (*SD*)
Seeing red	0.55	2.32 (1.92)
Full access	0.25	2.22 (1.77)
Glitzy science – transfer	0.80	2.03 (1.57)
Glitzy science – retention	0.05	1.50 (1.75)
Clean scents	0.47	2.05 (1.70)

Ranges of inferred effect sizes for individual participants also proved instructive. For example, for the Full Access report, only one participant generated an inferred effect size approximately the same as the actual effect size (0.22 versus 0.25). The next closest inferred effect size was nearly twice as large as the actual effect size and responses for fully 20% of participants yielded inferred effect sizes in excess of 3.0. Figure 1 shows the distribution of inferred effect sizes for the Full Access report and Figure 2 for the Clean Scents report. For the Seeing Red report, only three participants had inferred effect sizes less than the actual effect size and 18% had inferred effect sizes exceeding 3.0. Similar patterns in the range of inferred effect sizes occurred for the other reports and they yielded histograms similar to Figures 1 and 2, i.e., with noticeable positive skewness.

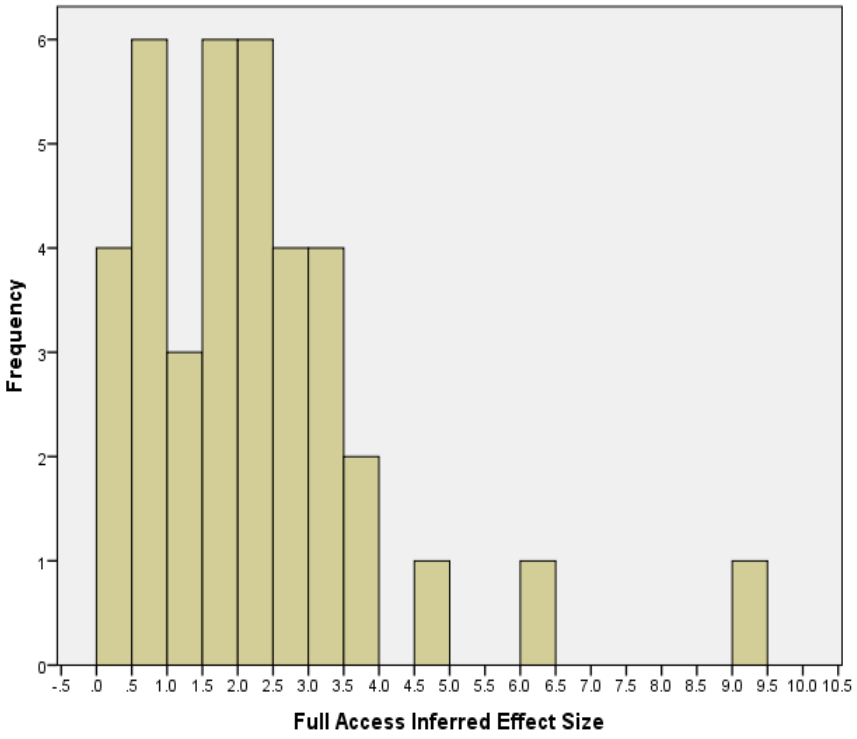

Figure 1. Histogram of inferred effect sizes for full access report.

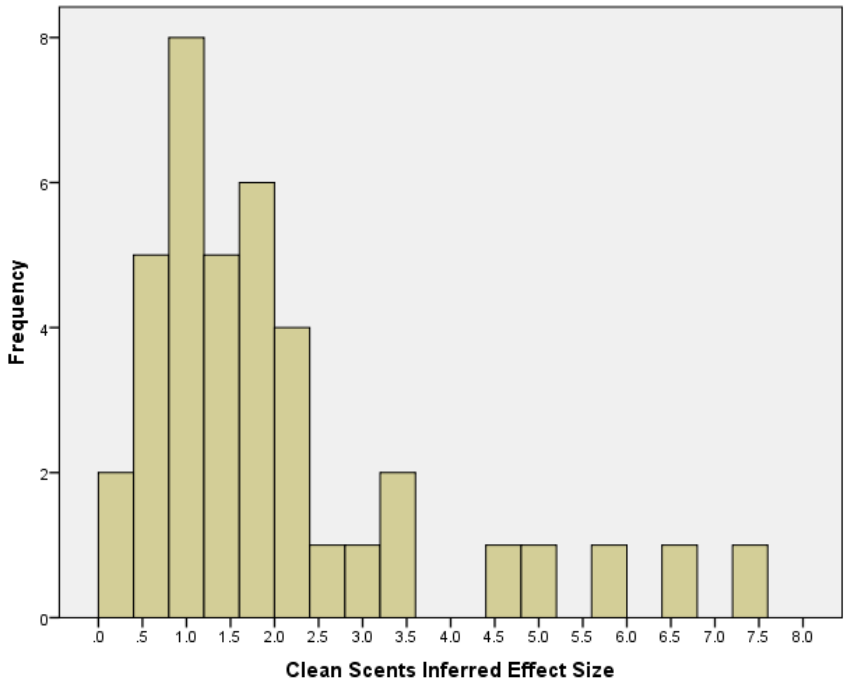

Figure 2. Histogram of inferred effect sizes for clean scents report.

The measures of effect size used here depend on both the difference in means (in the numerator) and on standard deviations (in the denominator). What is the interplay between these two sources of influence in participants' inferred effect sizes in comparison with data in the original articles? Table 2 shows the relevant comparisons for four of the variables. It presents means and standard deviations (SDs) taken from the original articles (all presented to one decimal place for consistency across reports) and means and SDs derived from estimated scores, that is, average means and SDs for individuals in Study 1. Note that participants were not asked to directly estimate means and SDs. They were asked to estimate scores and the authors calculated means and SDs from these estimated scores.

For all four variables in Table 2, on average, respondents substantially overestimated the difference in means. For example, for the Full Access report, the actual mean difference was approximately 10 points whereas the difference derived from estimated scores was approximately 39 points. For the two Glitzy Science variables, respondents estimated scores yielded SDs that reasonably approximated SDs in the original articles. Thus, all of the inflated inferred effect sizes (see Table 1) arise partly as a function of the difference in means. For the Full Access and Clean Scents variables, the inflation of inferred effect sizes is partly a function of difference in means and partly a function of somewhat reduced variability in estimated scores.

Table 2. Means and standard deviations: original reports and estimated scores

	In original Article		Derived from estimated scores	
Report/Group	*M*	*SD*	*M*	*SD*
Access				
Full access	237.5	30.1	264.7	23.6
No access	228.6	43.8	225.4	25.9
Glitzy science-transfer				
Glitzy	5.9	2.4	10.8	1.6
Not glitzy	4.1	1.9	7.8	1.9
Glitzy science-retention				
Glitzy	5.9	2.4	15.0	2.2
Not glitzy	5.8	1.8	12.2	2.4
Clean scents				
Scented	4.2	1.9	5.8	1.1
Not scented	3.3	2.0	4.0	1.3

Figures 3 and 4 depict some of the data in Table 2. The vertical bars in the figures show mean plus and minus one SD for the two groups in the original report and for participants in the current study. Figure 3 illustrates the situation for the Full Access study: overestimation of the difference in means combined with underestimation of differences in within-group variability. Figure 4 illustrates the situation for the Glitzy Science-Retention variable: overestimation of differences in means but reasonable approximation of within-group variability.

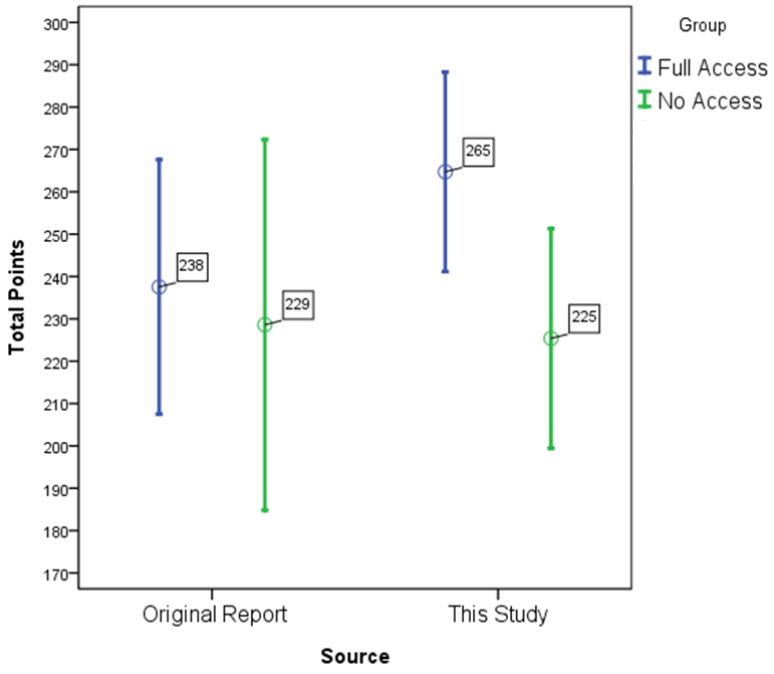

Figure 3. Mean and SD widths for original data and this study's data for full access.

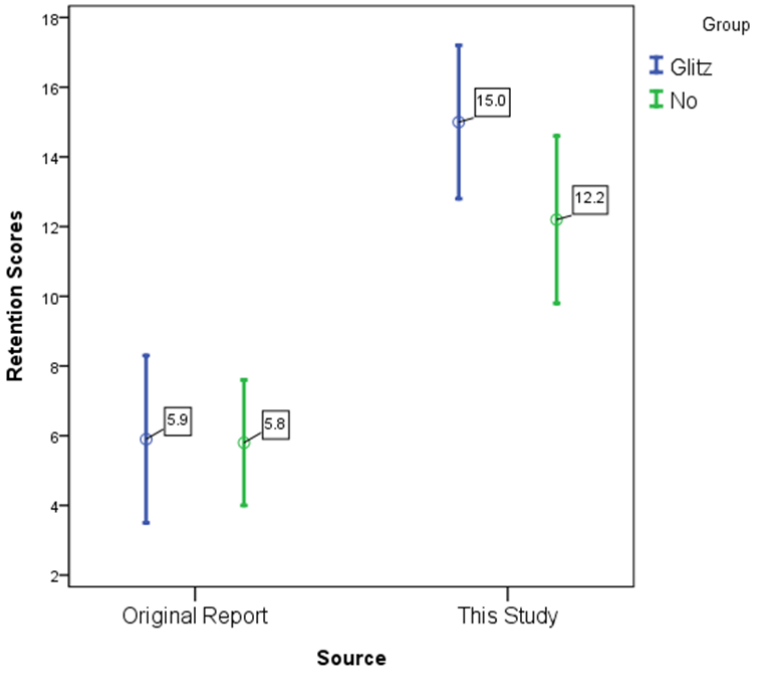

Figure 4. Mean and SD widths for original data and this study's data for glitzy science– retention.

3.2. RESULTS OF STUDY 2

Table 3 shows actual and inferred effect sizes for the reports in Study 2. Table 4 shows the actual percentages and estimated percentages for the two percentage-based variables in Study 2. For the Seeing Red reports used in Study 2, respondents (n = 42) reading the *Monitor* summary generated somewhat but not significantly greater inferred effect sizes than did respondents (n = 43) reading the official abstract ($t(81) = 1.45$, p = .150, d = 0.30). The average inferred effect size for Study 1 (n = 40) and Study 2 (n = 42) participants who read the *Monitor*'s version of the report differed little ($t(79) = .49$, p = .626, d = 0.10), showing good comparability of results for the two studies.

For the Violent Media article, the actual effect size (d = 0.61) for amount of time taken to help the victim was drastically less than the average inferred effect size (M = 3.53, SD = 3.48). The inferred effect size for differences in severity ratings (M = 2.92, SD = 1.44) also indicated a marked overestimation of the actual difference reported between the groups (d = 0.27). Inferred differences between violent and nonviolent conditions on the two percentage-type variables also greatly overestimated actual differences in the report (see Table 4). On one variable, the inferred difference was 31 percentage points (47–78%) and the actual difference was 4 percentage points (21–25%); on the other variable, the inferred difference was 33 percentage points (54–87%) and the actual difference was 4 percentage points (95–99%).

Table 3. Actual and average inferred effect sizes for reports in study 2

Report	Actual *d*	Inferred *d* *M* (*SD*)
Seeing red	0.55	
From abstract		2.01 (1.41)
From *monitor* [a]		2.50 (1.78)
Violent media		
Time to help	0.61	3.53 (3.48)
Severity of fight	0.27	2.92 (1.44)
Simulated abstract[b]		
"Significantly higher"		3.42 (2.14)
"Slightly higher"		1.99 (1.93)

[a] Compare *M* (*SD*) = 2.32 (1.92) from Table 1.
[b] The simulated abstract did not have an actual effect size.

Table 4. Actual and inferred group contrasts in the violent media report

	Actual		Inferred	
	Nonviolent	Violent	Nonviolent	Violent
% Likely to help victim	25%	21%	78%	47%
% Likely to hear fight	99%	95%	87%	54%

The Simulated Abstract that reported "slightly higher" scores produced a significantly lower inferred effect size than the abstract that reported "significantly higher" scores ($t(85) = 3.27$, $p = .002$, $d = 0.70$). This result demonstrated the possibility of controlling readers' impression of the data with very simple changes in wording– in this case, just a single word.

4. DISCUSSION

The discussion summarizes conclusions based on the separate studies and then develops broader generalizations that seem to emerge from the combination of the two studies. The discussion then offers some speculations about mental processes that might have given rise to the results and offers suggestions for follow-up research.

4.1. DISCUSSION OF STUDY 1

Results from Study 1 seem to support the following conclusions. First, on average, participants clearly overestimated the magnitude of group differences. This occurred for all four reports used in the study, including reports with a range of actual effect sizes and reports from journal abstracts as well as from news-like summaries. Second, the similarity of mean inferred effect sizes was remarkable. It seemed that no matter how the group difference was conveyed in the original summary, participants on average inferred a difference corresponding to an effect size of about 2. Third, participants (with the exception of one participant on just one variable) exhibited almost uncanny accuracy in

estimating the actual within-group variability as represented by their estimated scores for several variables.

4.2. DISCUSSION OF STUDY 2

The first goal of Study 2 was clearly achieved. Results from Study 1 were replicated with an independent sample, specifically using the Seeing Red news-like report. Inferred effect sizes based on this report were virtually identical for the Study 1 and Study 2 samples. In addition, Study 2 showed overestimation of effect sizes for two dependent variables in the Violent Media study, with inferred effect sizes again substantially overestimating actual effect sizes.

The second goal of Study 2 was successful in establishing that the overestimation of group differences occurred for both news-like reports and official journal abstracts. The news-like report, which might be expected to dramatize the group difference, gave a slightly greater, but statistically nonsignificant, overestimation. Although Study 1 and Study 2 used a mixture of news-like reports and official journal abstracts, this contrast for the Seeing Red report was the only instance allowing for a direct comparison of the two types of reports.

The third goal of Study 2 related to investigation of percentage-based reports of results. Percentage results in the Violent Media study clearly showed overestimation of group differences even more dramatically than differences based on means. The Violent Media report gave group differences for two variables where results were available in percentage form. In both cases, participants greatly overestimated the actual differences.

The fourth goal of Study 2 was accomplished. A very slight change in wording within a simulated abstract resulted in a noticeable difference in the effect size inferred from the report about the magnitude of a group difference. In this part of the study, there was no actual effect size since the abstract was just a simulation. Nevertheless, readers of the abstract could infer something about the magnitude of the difference underlying the report. And results showed that readers were sensitive to the change of a single particular word in the abstract.

4.3. GENERAL DISCUSSION

In general, participants' inferred effect sizes substantially overestimated the actual effect sizes in the studies. For all four reports in Study 1 (except for the variable on which no difference was reported), inferred effect sizes exceeded 2.0, a value well beyond Cohen's (1988) benchmark of 0.8 for a large effect size. The degree of overestimation ranged from a factor of 2.5 times the actual effect size to a factor of nearly 8 times as large as the actual effect size. Study 2 yielded overestimation by factors at least this large. That is, regardless of what the original data show, if a report says there was "a difference between groups," the intelligent layperson infers an effect size of about 2–3, a level that is virtually unheard of in actual behavioral studies, dubbed here as the *tall-tale effect*: What a layperson infers when reading about a group difference is a tall-tale in comparison with the actual magnitude of the difference. The tall-tale effect is potent and applies regardless of the magnitude of the difference (the actual effect size) reported in the original source.

In the Glitzy Science report in Study 1, two dependent variables were identified but the abstract referred to a difference between groups for only one of the dependent variables. Nevertheless, participants generalized the reported difference on the one dependent variable to the other dependent variable and with nearly the same degree of

inferred effect size. This result suggests that the tall-tale effect spreads from dependent variables which do show a difference to other variables which may not have a difference.

Results for the Violent Media report demonstrated that the tall-tale effect occurs for percentage-type variables, too. In fact, contrasts between actual and inferred results appeared even more dramatic than the differences in inferred effect sizes for variables based on means. The inferred differences for percentages were approximately eight times larger than the actual differences found in the study.

The contrast in results for the official abstract and the news-like report for the Seeing Red stimulus materials showed that the tall-tale effect emerges for both types of reports. The difference was statistically nonsignificant but tilted toward a stronger effect for the news-like summary. One might expect the news-like version to dramatize the group difference and, therefore, magnify the tall-tale effect. This point merits further investigation.

The results demonstrated that it is possible to control the tall-tale effect, at least to some degree, with changes in the ordinary language used in research reports. The change in one word in the simulated abstract (from "significantly" to "slightly") reduced the inferred effect size by approximately 40% (from 3.42 to 1.99). However, even the word "slightly" yielded a large inferred effect size. Nevertheless, the result suggests that the work described earlier on scaling words related to probabilities, in the context of medical applications, has the potential to improve laypersons' understanding of the real outcomes of behavioral research reports. It may be possible to develop a set of descriptive words that will calibrate laypersons' inferences to actual effect sizes. For example, it may be possible to establish a difference in inferred effect size for "very slightly" versus "slightly" versus "substantially" and so on. The goal is to calibrate the words used in abstracts and summaries with the actual magnitude of differences as represented in the detailed results of a study.

4.4. SUGGESTED HEURISTICS OPERATING IN THE TASK

It may be useful to apply the notion of heuristics, as developed by Tversky and Kahneman (1974), to how respondents seemed to think about reports of group differences as presented in abstracts and news-like summaries. Tversky and Kahneman noted that

> "… people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. In general, these heuristics are quite useful, but sometimes lead to severe and systematic errors" (p. 1124).

What principles and tendencies seem to operate in the task of interpreting reports of group differences as in the reports read by respondents? First, participants readily understood the nature of the task. After receiving very brief instructions, participants (128 in 12 sessions) raised no questions about how to proceed and only one case from each study had to be eliminated due to clearly aberrant responses. Second, participants clearly perceived the group differences conveyed in the abstracts. In over 1000 sets of responses, only a handful of cases revealed a fuzzy or incorrect understanding of the report as indicated by examination of respondents' written summaries. Third, participants almost always overestimated the magnitude of the differences between groups, often grossly so. This third point is the most important one in these two studies and the one successfully controlled, to some extent, by adjusting the wording in the simulated abstract. Furthermore, this third point relates most directly to the research on probabilistic word-number relationships in medical contexts. That research showed the possibility of selecting probabilistic words to maximize accurate, although less than perfect,

communication with persons who were well educated but not necessarily statistical experts. Further progress can be made along these lines to aid in the unending pursuit of better ways to communicate research findings. Fourth, respondents represented within-group variability in their estimated scores. For the two studies combined, out of 1072 instances where respondents gave estimated scores, in only two instances did a respondent assign the same score to all members of a group, as described earlier. (This respondent did represent variability in scores for all other reports to which she responded in Study 1.) Nothing in the directions for the exercises suggested that estimated scores had to vary within groups. Apparently, respondents "just knew" that scores would vary within groups. Fifth, for some variables, respondents closely approximated the within-group variability in the original reports. This result seems quite remarkable since nothing in the abstracts or summaries would have given a clue about the extent of within-group variability. For other variables, respondents underestimated within-group variability, but always represented considerable variability in their numerical estimates of scores.

These results relate to studies that directly examined persons' attempts to represent variability both between and within groups. Ben-Zvi (2004), noting that "the group comparison problem is one that students do not initially know how to approach" (p. 44), described several studies of students' approach to dealing with variation while the students were engaged in statistical training. Participants in the current studies were *not* engaged in statistical training. Thus, the results suggest a kind of primitive thinking about the group comparison problem. Gould (2004) emphasized the "make a picture" rule for data analysis. Results of the current studies suggest that participants seemed to create a type of picture based on reading the abstracts of studies. Upon reading that *Group A differed from Group B*, participants created a number-based picture that showed variation between groups corresponding to an effect size of 2-3 and, often, amazingly accurate representation of within-group variability.

The methodology in both Study 1 and Study 2 involved group administration of the exercises. An adaptation of the procedure for individual administration using the think-aloud technique might elucidate how individuals process the task. The two studies were not designed to provide detailed observation of respondents as they worked on the tasks. Nevertheless, investigators certainly developed impressions of how respondents proceeded. A typical case went like this: The respondent read the abstract, often going back over it and marking parts of the text. Next, the respondent proceeded to the sheet for recording the summary "in your own words" and completed that summary. Then the respondent proceeded to the page calling for estimated scores. There was a noticeable pause, with the respondent often staring straight ahead or at the ceiling. (This phase is where the think-aloud technique should be particularly useful.) Next the respondent started recording the estimated scores, usually all for one group, then for the other group. Finally, the respondent scanned the recorded scores, sometimes making a few adjustments. All these steps occurred in about 10 minutes for each report. This reconstructed scenario, and variations on it, needs to be validated with the individual administration and think-aloud technique.

4.5. LIMITATIONS AND FUTURE RESEARCH

The studies reported here have limitations. First, the method used to infer participants' understanding of the magnitude of group differences (estimating scores on the dependent variable and then calculating an inferred effect size) is a novel one. It provided a practical, specific way to operationalize a person's perception of the magnitude of group differences. However, there may be other ways to operationally

define participants' perception of the magnitude of differences, such as by direct estimation of means or by some type of adjustable graphic display. (The estimation of percentages for the Violent Media report in Study 2 did not suffer from this limitation because those participants estimated percentages directly.) As noted previously, it would be useful to interview participants as they worked through the task of inferring the magnitude of group differences. Second, by intent, all of the reports used in the two studies involved two-group contrasts. Reports referring to correlations between two variables were not examined. For example, what does the layperson, intelligent but lacking advanced statistical training, make out of a statement such as "SAT scores are correlated with GPA," or "scores on the Working Memory Test significantly predict job performance"? It seems likely that the tall-tale effect applies to reports of correlations as well, but that conjecture awaits future study. Third, further work on the tall-tale effect needs to be completed with a variety of groups. A useful follow-up study might involve graduate student TAs in statistics (as in Noll, 2011). The two studies reported here deliberately targeted college students, well-educated and with some exposure to the methods of behavioral science, but without advanced statistical training. Would the typical introductory statistics course, which usually includes coverage of the concept and methods of effect size measures, "cure" the problem? It seems unlikely that one course (perhaps one lecture on effect sizes) would cure the problem but the question needs treatment with such groups. Fourth, the effort to calibrate inferred effect size with changes of wording in abstracts was limited to just one contrast: "slightly" versus "significantly." Although that investigation showed promise, there is still much to explore for a full range of wording calibrations.

ACKNOWLEDGMENT

Earlier versions of the studies reported here appeared as posters at the Eastern Psychological Association meeting in Cambridge, MA, USA in 2011.

REFERENCES

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, *13*(3), 515–539.

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.).Washington, DC: Author.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. [Online: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3174372/pdf/13428_2011_Article_89.pdf]

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, *3*(2), 42–63. [Online: http://iase-web.org/documents/SERJ/SERJ3(2)_BenZvi.pdf]

Berke, D. M., Rozell, C. A., Hogan, T. P., Norcross, J. C., & Karpiak, C. P. (2011). What clinical psychologists know about evidence-based practice: Familiarity with online resources and research methods. *Journal of Clinical Psychology*, *67*(4), 329–339.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Browne, R. H. (2010). The *t*-test *p* value and its relationship to the effect size and P(X>Y). *The American Statistician*, *64*(1), 30–33.

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, *41*(3), 390–404.

Bryant, G. D., & Norman, G. R. (1980). Expressions of probability: Words and numbers. *New England Journal of Medicine*, *302*(7), 411.

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes*, *36*(3), 391–405.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(2), 281–294.

Bushman, B. J., & Anderson, C. A. (2009). Comfortably numb: Desensitizing effects of violent media on helping others. *Psychological Science, 20*(3), 273–277.

Cepeda, N. J. (2008). Effect size calculator.
[Online: http://cognitiveflexibility.org/effectsize/]

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed., first published in 1969). Hillsdale, NJ: Erlbaum.

Cynkar, A. (2007). Seeing red impairs test performance. *Monitor on Psychology*, *38*(5), 11.

Elliot, A. J., Maier, M. A., Moller, A. C., Friedman, R., & Meinhardt, J. (2007). Color and psychological functioning: The effect of red on performance attainment. *Journal of Experimental Psychology: General*, *136*(1), 154–168.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*(7), 591–601.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you.* New York: Simon & Schuster.

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: From innumeracy to insight. *British Medical Journal*, *327*(7417), 741–744.
[Online: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC200816/]

Gigerenzer, G., Gassmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53–96.

Gould, R. (2004). Variability: One statistician's view. *Statistics Education Research Journal*, *3*(2), 7–16.
[Online: http://iase-web.org/documents/SERJ/SERJ3(2)_Gould.pdf]

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York: Routledge.

Higgins, J. P. T., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. [Online: http://www.cochrane-handbook.org]

Hove, M. C., & Corcoran, K. J. (2008). If you post it, will they come? Lecture availability in introductory psychology. *Teaching of Psychology*, *35*(2), 91–95.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science: A Journal of the American Psychological Society*, *16*(5), 345–353.
[Online: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1473027/]

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746–759.

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

Liljenquist, K., Zhong, C., & Galinsky, A. D. (2010). The smell of virtue: Clean scents promote reciprocity and charity. *Psychological Science, 21*(3), 381–383. [Online: http://fpweb.fmarion.edu/wwattles/psy302/TheSmellofVirtue.pdf]

Mayer, R. E., Griffith, E., Jurkowitz, I. T. N., & Rothman, D. (2008). Increased interestingness of extraneous details in a multimedia science presentation leads to decreased learning. *Journal of Experimental Psychology: Applied*, *14*(4), 329–339.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*(2), 361–365.

Mosteller, F., & Youtz, C. (1990). Quantifying probabilistic expressions. *Statistical Science, 5*(1), 2–34. [Online: http://projecteuclid.org/download/pdf_1/euclid.ss/1177012242]

Noll, J. (2011). Graduate teaching assistants' statistical content knowledge of sampling. *Statistics Education Research Journal*, *10*(2), 48–74. [Online: http://iase-web.org/documents/SERJ/SERJ10(2)_Noll.pdf]

Novotney, A. (2009). In brief. *Monitor on Psychology*, *40*(1), 12. [Online: http://www.apa.org/monitor/2009/01/inbrief.aspx]

O'Brien, B. J. (1989). Words or numbers? The evaluation of probability expressions in general practice. *Journal of the Royal College of General Practitioners*, *39*(320), 98–100. [Online: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1711769/]

Odgaard, E. C., & Fowler, R. L. (2010). Confidence intervals for effect sizes: Compliance and clinical significance in the Journal of Consulting and Clinical Psychology. *Journal of Consulting and Clinical Psychology*, *78*(3), 287–297.

Renooij, S., & Witteman, C. (1999). Talking probabilities: Communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, *22*(3), 169–194.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*(4), 371–386.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

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

Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. [Online: https://www.apa.org/pubs/journals/releases/amp-54-8-594.pdf]

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Newbury Park, CA: Sage.

THOMAS P. HOGAN
Psychology Department
University of Scranton
800 Linden Street
Scranton, PA 18510-4596

APPENDIX – EXAMPLES OF MATERIAL USED IN THE STUDIES

NB: Material is presented here in condensed form to conserve space. Original material used with participants were double spaced, on separate pages (as indicated here), and sometimes in larger font. Each form provided spaces for estimated scores for ten students in each group.

Material for "Clean Scent" Study

(*Page 1*)
Here is the summary of a report issued by the American Psychological Society based on a study appearing in the journal *Psychological Science*, February 4, 2010.

The Smell of Virtue: Clean Scents Promote Reciprocity and Charity
Katie Liljenquist, Chen-Bo Zhong, and Adam D. Galinsky
Clean scents may promote good behavior: Volunteers who completed a task in a clean-smelling room (sprayed with citrus Windex) reported more interest in participating in and donating to a charity organization than volunteers who were in a regular-smelling room. In addition, volunteers playing a trust game in a clean-smelling room were likelier to return more money to a hypothetical partner than volunteers in a regular-smelling room. These findings indicate that clean scents not only motivate clean behavior, but may also encourage virtuous actions.

(*Page 2*)
Now, in your own words, please write a brief, simple summary of what was done in this study and what the results were. You can refer back to the report, if you wish.

(*Page 3*)
In this study, participants (some in a scented room, some in a non-scented room) completed ratings of their interest in volunteering for future Habitat efforts.
They made their ratings on a 7-point scale from:
1 = LOW interest to 7 = HIGH interest.

Let's say we take 10 students from each condition: scented room and non-scented room. Based on the report of results given in the abstract above, list what ratings you think they made about their interest in volunteering.

Students in Scented room	Students in Non-scented room
1. _____	1. _____

(*Each column extended to 10 members in each group*.)

Material for "Simulated Abstract"

(*Page 1*)
Here is a possible report from a psychological study on memory in the elderly entitled:

The Effect of Completing Puzzles on Short Term Memory in the Elderly

Abstract. It is well known that short-term memory (STM) deteriorates with age. Currently, research is directed at determining how to reduce these negative effects. This study aims to establish the effects of completing a daily puzzle on performance on a test of short term memory. The hypothesis stems from the idea that keeping an active mind will aid in retention of everyday events. In this study, 153 participants pooled from various residential nursing facilities located in the Boston area were asked to perform several short term memory tasks, such as digit span and simple word recall. The tasks were compiled to produce one STM score with a maximum of 150 points. Participants were randomly assigned to either complete a puzzle on a daily basis or keep their usual daily routine (control group). A different type of puzzle was used each day of the week, including crosswords, word searches, anagram squares, Sudoku, logic problems, and cryptograms. Participants who completed a puzzle daily scored **significantly higher** on the short term memory test than those who did not complete the puzzles.

(*Page 2*)
Now, in your own words, please write a brief, simple summary of what was done in this study and what the results were. You can refer back to the report, if you wish.

(*Page 3*)
In this study, participants were assigned to either the **puzzle condition** or the **no puzzle condition.** Then, they were asked to perform several short term memory (STM) tasks. One final STM score was produced.

Let's say we take 10 participants from each condition: puzzle or no puzzle. List what scores you think they received on the **150 point** STM tasks.

Puzzle Condition	No Puzzle Condition
1. _____	1. _____

(*Each column extended to 10 members in each group.)*

Instructions for making estimates for "Violent Media" study

(*Page 3*)

Let's take the first study reported in the article. In this study, participants either played a **violent or nonviolent video game** for 20 minutes. A loud fight was staged after gameplay. Groups were compared on the following measures:

1. **Helping Rates**: whether the participant left the room to help the victim.
2. **Time to Help:** the length of time (in seconds) from when the fight was over to when the participant left the room to help.
3. **Heard Fight:** whether the participant reported hearing the fight.
4. **Severity of Fight:** participant report of how serious the fight was on a 10-point scale.

Based on the report of results given in the abstract above, please answer the following questions:

1. What percentage of each group do you think helped the victim?

   Played Violent Game: _____ % Played Nonviolent Game: _____ %

2. What percentage of each group do you think reported hearing the fight?

   Played Violent Game: _____ % Played Nonviolent Game: _____ %

(*Page 4*)

Now, let's say we take 10 participants from each condition: violent game and nonviolent game. First, please list how long **(in seconds)** you think each participant took to help the 'victim'. (Use a **maximum** of three minutes, i.e. **180 seconds**.)

Violent Game Condition	Nonviolent Game Condition
1. _____	1. _____

(*Each column extended to 10 members in each group.*)

(*Page 5*)

Now, please estimate 10 participants' ratings of the severity of the fight. The ratings are on a 10-point scale with 1 being the least serious and 10 being the most serious.

Violent Game Condition	Nonviolent Game Condition
1. _____	1. _____

(*Each column extended to 10 members in each group.*)