By the Numbers: Toward More Precise Numerical Summaries*

Gaurav Sood† Andrew Guess‡

February 24, 2017

Unlike the natural sciences, there are few true zeros in the social sciences (p. 960, Gelman, 2011). All sorts of variables are often weakly related to each other. And however lightly social scientists, exogenous events, or other actors intervene, effects of those interventions are rarely precisely zero. When there are few true zeros, categorical statements, like "two variables are significantly related" or "the intervention had a significant effect," convey limited information—about sample size and luck (and manufactured luck). Yet these kinds of statements are the norm in abstracts of the top political science journal, the *American Political Science Review* (*APSR*). As we show later, only 10% of the empirical articles in recent volumes of *APSR* have abstracts with precise quantitative statements. The comparable number for the *American Economic Review* (*AER*) is 35%.

Informal inspection also suggests that coarse descriptions are common in other sections of social science articles, aside from being exceedingly frequent in the social science vernacular. (We would like to quantify both.) Studies are often summarized as "the study shows this broad phenomenon exists." For instance, Pasek, Sood and Krosnick (2015) write, "Moreover, much past research shows that guessing is often biased." Surprisingly, and more problematically, comparison of results often enough also takes the same form. For instance, a study on motivated learning contextualizes the results as follows: "Our finding is consistent with other research showing that

*We thank Justin Esarey, Andrew Gelman, Don Green, Kabir Khanna, Brendan Nyhan, and Daniel Stone for useful comments. The data and the scripts behind the analysis presented here can be downloaded at: http://github.com/soodoku/quant-discipline/

†Gaurav is a data scientist. He can be reached at: gsood07@gmail.com

‡Andrew is a Postdoctoral Fellow at the Social Media and Political Participation (SMaPP) Lab, New York University. He can be reached at: guess@nyu.edu

even when people agree on factual information, they often still interpret the information in a motivated manner" (from Khanna and Sood 2015). The phrase "finding is consistent with" yields nearly 300,000 results on Google Scholar. And informal inspection suggests that "consistent" covers a surprisingly large range of effect sizes and measures.

None of this is to say that the decision to summarize imprecisely is made without deliberation. Undoubtedly, some resort to coarse summaries because they are not confident about their theories, measures, or models. Others likely use coarse summaries because they think coarse summaries are a more effective way to communicate. In fact, there is some empirical support for the latter thesis. A survey of undergraduate and graduate students found that 77% of the students thought that people preferred to receive verbal expressions of uncertainty over numerical expressions in their everyday lives (Wallsten et al., 1993). (However, the paper also notes that 85% of the students felt comfortable switching to another mode if they thought the other mode suited their needs better.)

Our hunch, however, is that the most common form of coarse summaries in scientific communication—categorical statements around statistical significance—arise as a natural consequence of scientists thinking in terms of the Null Hypothesis Statistical Testing (NHST) framework, which in turn is likely underpinned by a Popperian understanding of science (Gelman and Shalizi, 2013). For instance, hypotheses written in the form of coarse categorical statements around statistical significance, such as, "X will be significantly associated with Y," are exceedingly frequent. These kinds of hypotheses reflect an understanding of science in which scientific progress comes from falsification rather than improvements in measurement.

Whatever the root cause, the use of coarse summaries likely leads to serious problems. First, coarse summarizations risk misinterpretation. Partly because the mapping between verbal phrases and numerical ranges varies between communicators and recipients (Capriotti and Waldrup, 2011), recipients map the same verbal expression to very different numbers (Beyth-Marom, 1982; Bocklisch et al., 2010; Brun and Teigen, 1988; Simpson, 1944, 1963). For instance, Beyth-Marom (1982) elicited numerical mappings of 30 verbal expressions on a

100-point scale and found that the average inter-quartile range of the numerical mapping for a phrase was 14.4. Analogous numbers—standard deviations of the numerical mappings—from Brun and Teigen (1988), based on 27 phrases, and Bocklisch et al. (2010), based on 13 phrases, were 14.2% (translated from a 0–6 scale) and 11.15%, respectively. Relatedly, mapping numerical ranges to verbal phrases in a way that minimizes misclassification error still yields an error rate of nearly 28% (Elsaesser and Henrion, 2013) (see also Bocklisch et al. 2010).

Not only are the mappings variable but the variation is also systematic. Numerical mappings of verbal phrases vary systematically as a function of the phrases used and the characteristics of the recipient. Prominently, numerical mappings of verbal phrases about infrequent events (e.g., "seldom," "rarely," "uncommon") tend to be much less reliable (Wallsten, Fillenbaum and Cox, 1986). Interpretation of vague verbal summaries is also subject to cognitive errors. The vaguer a statement, the greater the opportunity to fill in the missing detail. And it is plausible, neigh likely, that people do not resist the opportunity to impute, using common heuristics, such as overweighting accessible information, interpreting evidence in a way that is congenial to their prior beliefs, etc., to impute missing details (Nickerson, 1998; Tomz and Van Houweling, 2009; Brun and Teigen, 1988; Wright, Gaskell and O'Muircheartaigh, 1994).

Use of such cognitive shortcuts is liable to lead to systematic biases in inferences. For instance, in the extreme, confirmation bias implies that people will read an uninformative vague statement as evidence that their priors are correct. It follows that on reading such a statement, people will walk away with yet greater certainty about their priors. For instance, a person who initially believes that a law allowing concealed carry would increase gun crime may optimistically conclude after reading a study summary reporting a positive effect ("the study shows that laws allowing concealed carry increase gun crime") that allowing concealed carry increases gun crime by 20%. The same person reading a summary reporting the opposite effect ("the study shows that laws allowing concealed carry decrease gun crime") may optimistically conclude that the decline is real but of a much lower magnitude.

Finally, coarse summaries may lead to erroneous inferences because of how people use

language ordinarily. For instance, a person with flat priors about selective exposure may reasonably interpret a vague summary ("people engage in selective exposure") as implying that most people read news stories from sources that they think are aligned with their party. A more precise numerical statement of the sort that gives the proportion of news stories consumed from ideologically congenial sources would preempt the risk of such misinterpretation.

Besides misinterpretations of topical effect sizes, coarse summarizations also risk conveying misleading ways of thinking about science—as falsification or simple directional claims rather than as a constant effort to obtain less biased and more precise estimates of actual quantities of interest. Presenting more precise estimates may instill in readers a better appreciation of the point that Donald Green made in an interview in the aftermath of the LaCour scandal: "That's what makes the study interesting. Everybody knows that there's some degree of truth in these propositions, and the reason you do an experiment is you want to measure the quantity."[1]

Making more precise numerical statements may also improve how we understand the results of studies. And over the longer term, by making us think more carefully about our priors, precise numerical summaries may improve how we think about science and interpret scientific results. For instance, presenting precise numerical summaries in abstracts may help us more quickly filter studies in which results appear "too big."[2]

With this preface, we proceed to examine the frequency of vague judgments in social science abstracts.

[1] "An Interview With Donald Green, the Co-Author of the Faked Gay-Marriage Study." Jesse Singal. *New York Magazine*. Published on May 21, 2015. http://nymag.com/scienceofus/2015/05/co-author-of-the-faked-study-speaks-out.html. Green was giving the interview in the aftermath of revelations that Green's co-author had fabricated the data in a highly-publicized study on the persuasive effects of canvassing on attitudes toward gays (Broockman, Kalla and Aronow, 2015; McNutt, 2015).

[2] Pushing at an open door: When can personal stories change minds on gay rights? Andrew Gelman. Monkey Cage. The Washington Post. Published on December 19, 2014.

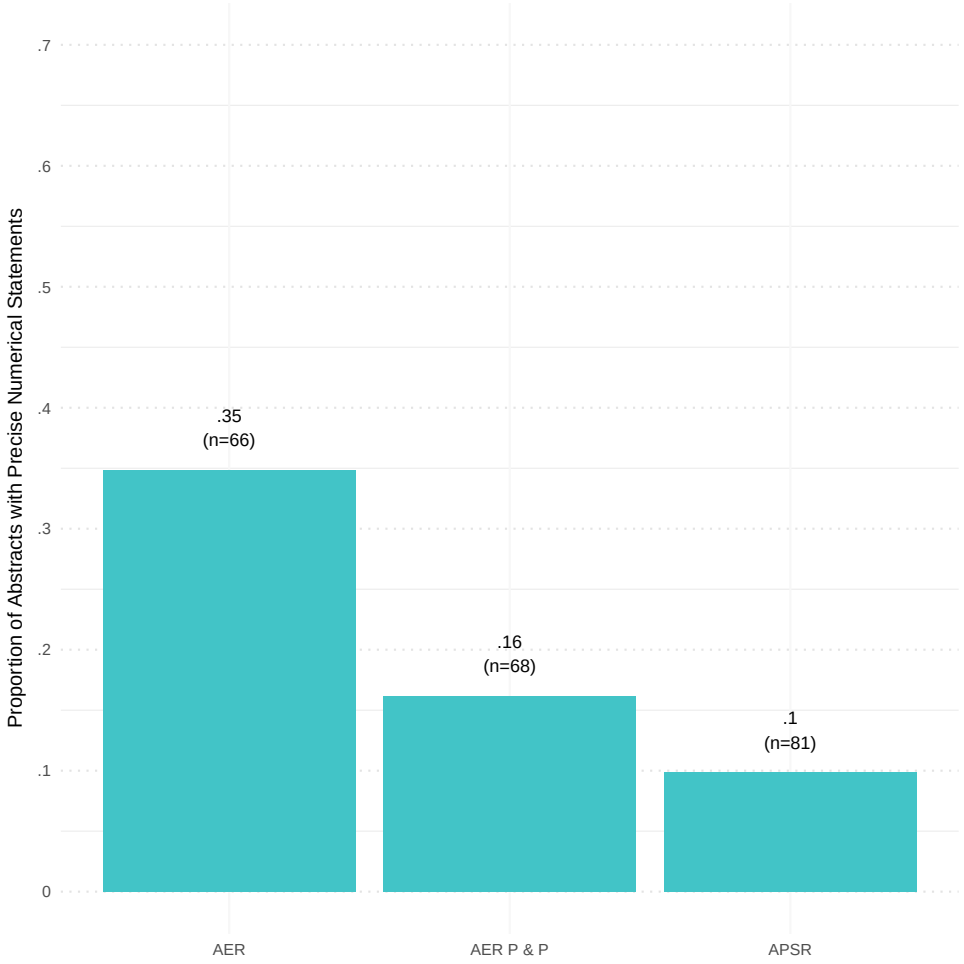
How Common Are Precise Numerical Summaries in Abstracts?

To assess how common coarse summaries are vis-à-vis more precise numerical summaries of results, we coded 310 abstracts—117 *APSR*, 100 *AER*, and 93 *AER Papers & Proceedings* (*AER P & P*). The *AER* and *AER P & P* samples span June 2014–June 2016 (Vol. 104, 6 through Vol. 105, 6), while the sample of *APSR* abstracts spans February 2013–May 2015 (Vol. 107, 1 through Vol. 109, 2). Given we are only interested in articles in which there are empirical results that can be summarized precisely, we subset on empirical papers. This leaves us with 66 *AER*, 68 *AER P & P*, and 81 *APSR* abstracts.

What we mean by precise numerical summaries of results deserves careful attention. Precision is on a continuum, with summaries ranging from very imprecise to very precise. But for clarity and convenience, our coding scheme captures only one end of the scale. We code summaries of results that take the following form as precise: "A% change in X caused a B% change in Y" or "the intervention caused B% change in Y." For instance, we code the following statements as precise: "The average proportion of 'no' votes is about 40% higher for applicants from (the former) Yugoslavia and Turkey," "I find that a one-percentage-point increase in the personal vote received by a gubernatorial candidate increases the vote share of their party's secretary of state and attorney general candidates by 0.1 to 0.2 percentage points." The complementary set includes statements like: "increasing numbers of armed military troops are associated with reduced battlefield deaths," "We find support for these arguments using original data from Uganda," etc.

Only about 10% of the empirical articles in recent volumes of *APSR* have abstracts with precise quantitative statements, similar to the percentage for *AER P & P*. The comparable number for *AER* is 35% (see Figure 1). None of the numbers are appealing, but the numbers for *APSR* stand out. The frequency of coarse summaries of empirical results in abstracts is, however, an imperfect indicator of the dominance of NHST inspired reasoning. The disparity between *APSR* and *AER* likely also stems from a lack of widely understood measures in political science. For instance, in

Figure 1: Proportion of Precise Numerical Statements in Abstracts of Empirical Papers in the *APSR*, *AER*, and the *AER P & P*



Economics, variables like unemployment, inflation, GDP, etc. are widely understood and studied. In political science, only a few variables like turnout come close to being widely understood.

In all, the data shed much needed, but still weak, light on the issue. It is our hope, however, that this note will stimulate discussion about social scientific writing, and increase efforts to address (what we contend is) the root cause of a particularly common coarse description—categorical statements around statistical significance.

References

Beyth-Marom, Ruth. 1982. "How probable is probable? A numerical translation of verbal probability expressions." *Journal of forecasting* 1(3):257–269.

Bocklisch, Franziska, Steffen F Bocklisch, Martin RK Baumann, Agnes Scholz and Josef F Krems. 2010. The role of vagueness in the numerical translation of verbal probabilities: A fuzzy approach. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. pp. 1974–1979.

Broockman, David, Joshua Kalla and Peter Aronow. 2015. "Irregularities in LaCour (2014)." *Work. pap., Stanford Univ. http://stanford. edu/ dbroock/broockman_kalla_aronow_lg_irregularities. pdf* .

Brun, Wibecke and Karl Halvor Teigen. 1988. "Verbal probabilities: ambiguous, context-dependent, or both?" *Organizational Behavior and Human Decision Processes* 41(3):390–404.

Capriotti, Kim and Bobby E Waldrup. 2011. "Miscommunication of uncertainties in financial statements: a study of preparers and users." *Journal of Business & Economics Research (JBER)* 3(1).

Elsaesser, Christopher and Max Henrion. 2013. "How Much More Probable is" Much More Probable"? Verbal Expressions for Probability Updates." *arXiv preprint arXiv:1304.1501* .

Gelman, Andrew. 2011. "Causality and Statistical Learning1." *American Journal of Sociology* 117(3):955–966.

Gelman, Andrew and Cosma Rohilla Shalizi. 2013. "Philosophy and the practice of Bayesian statistics." *British Journal of Mathematical and Statistical Psychology* 66(1):8–38.

Khanna, Kabir and Gaurav Sood. 2015. "Motivated Learning or Motivated Responding? Using Incentives to Distinguish Between the Two Processes." *Typescript* .

McNutt, Marcia. 2015. "Editorial retraction." *Science* p. aaa6638.

Nickerson, Raymond S. 1998. "Confirmation bias: A ubiquitous phenomenon in many guises." *Review of general psychology* 2(2):175.

Pasek, Josh, Gaurav Sood and Job Krosnick. 2015. "Misinformed About the Affordable Care Act? Leveraging Certainty to Assess the Prevalence of Misinformation." *Journal of Communication* .

Simpson, Ray H. 1944. "The specific meanings of certain terms indicating differing degrees of frequency.".

Simpson, Ray H. 1963. "Stability in meanings for quantitative terms: A comparison over 20 years." *Quarterly Journal of Speech* 49(2):146–151.

Tomz, Michael and Robert P Van Houweling. 2009. "The electoral implications of candidate ambiguity." *American Political Science Review* 103(01):83–98.

Wallsten, Thomas S, David V Budescu, Rami Zwick and Steven M Kemp. 1993. "Preferences and reasons for communicating probabilistic information in verbal or numerical terms." *Bulletin of the Psychonomic Society* 31(2):135–138.

Wallsten, Thomas S, Samuel Fillenbaum and James A Cox. 1986. "Base rate effects on the interpretations of probability and frequency expressions." *Journal of Memory and Language* 25(5):571–587.

Wright, Daniel B, George D Gaskell and Colm A O'Muircheartaigh. 1994. "How much is 'quite a bit'? Mapping between numerical values and vague quantifiers." *Applied Cognitive Psychology* 8(5):479–496.