Journal of Experimental Political Science **1** (2014) 81–98
doi:10.1017/xps.2014.11

Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee

Alan Gerber*, Kevin Arceneaux†, Cheryl Boudreau‡, Conor Dowling§, Sunshine Hillygus¶, Thomas Palfrey‖, Daniel R. Biggers* and David J. Hendry*

Abstract

The Standards Committee of the Experimental Research Section of the American Political Science Association has produced reporting guidelines that aim to increase the clarity of experimental research reports. This paper describes the Committee's rationale for the guidelines it developed and includes our Recommended Reporting Standards for Experiments (Laboratory, Field, Survey). It begins with a content analysis of current reporting practices in published experimental research. Although researchers report most important aspects of their experimental designs and data, we find substantial omissions that could undermine the clarity of research practices and the ability of researchers to assess the validity of study conclusions. With the need for reporting guidelines established, the report describes the process the Committee used to develop the guidelines, the feedback received during the comment period, and the rationale for the final version of the guidelines.

Keywords: Reporting guidelines, experiment, survey, laboratory, field, transparency, replication.

INTRODUCTION

The Standards Committee of the Experimental Research Section was charged with preparing a set of reporting guidelines for experimental research in political science. The committee defined its task as compiling a set of guidelines sufficient to enable the reader or reviewer to follow what the researcher had done and to assess the validity of the conclusions the researcher had drawn. Although the guidelines do request the reporting of some basic statistics, they do not attempt to weigh in on statistical

*Yale University, New Haven, CT, USA; email: alan.gerber@yale.edu (corresponding author)
†Temple University, Philadelphia, PA, USA
‡University of California, Davis, Davis, CA, USA
§University of Mississippi, University, MS, USA
¶Duke University, Durham, NC, USA
‖Caltech, Pasadena, CA, USA

controversies. Rather, they aim for something more modest but nevertheless crucial: to ensure that scholars clearly describe what it is they did at each step in their research and clearly report what their data show. In this report, we discuss the rationale for reporting guidelines and the process used to formulate the specific guidelines we endorse. The guidelines themselves are included as Appendix 1.

IS THERE A NEED FOR REPORTING STANDARDS?

Some basic information is essential for evaluation and communication of a study's research design and findings. Reporting guidelines can serve as a tool to assist scholars in their efforts to write papers that do not inadvertently leave out important details that would undermine the value of their studies' contributions. In this spirit, the Standards Committee's guidelines request that authors report a basic and common-sense list of items. However, if these guidelines are essentially a codification of existing practice, they would be innocuous but of little use. We investigated the possibility that reporting is already exemplary by surveying the reporting in a sample of articles that use experimental methods selected from the discipline's leading journals. We find that reporting is generally good, but even in work at the apex of the academic pyramid there are significant reporting gaps. We did not investigate all journals or include unpublished manuscripts, but it is our impression that the level of reporting omissions in these other outlets is likely to be similar or worse.

To get a rough idea of the extent to which political science experimental research already follows the reporting standards we recommend, we had two coders independently evaluate a random sample of published research. To form the population from which we sampled, we began by first identifying every article that employed some sort of experimental manipulation over the past nine years (2005–2013) in 16 highly regarded general interest and field journals.[1] Articles were included if they used any of the following words in their abstracts: "experiment," "randomly assigned," "treatment," "control," "treatment group," or "control group." Articles were then excluded if it was clear from the context of the abstract that the study was not referring to an experiment in the sense of a researcher-controlled manipulation.[2] This search yielded 490 articles in total. Each article's experiment(s) was then classified as field, survey, or laboratory based on its

[1]These journals are the *American Journal of Political Science*, *American Political Science Review*, *American Politics Research*, *British Journal of Political Science*, *Comparative Political Studies*, *International Organization*, *International Studies Quarterly*, *Journal of Conflict Resolution*, *Journal of Politics*, *Political Analysis*, *Political Behavior*, *Political Psychology*, *Political Research Quarterly*, *Public Opinion Quarterly*, *Quarterly Journal of Political Science*, and *World Politics*. The intention was to assess whether there were reporting deficiencies in the work published in a collection of journals considered among the most prestigious and influential in the discipline.

[2]For example, abstracts describing "natural experiments" or "quasi-experiments," or simple offhand uses of the search terms (e.g., referring to the implementation of a new policy as a "policy experiment") were eliminated. We can be reasonably confident that false identification of studies as experiments was not a problem in that our final sample used for analysis did not include any non-experimental studies.

abstract (and text when necessary).[3] Field experiments were defined as the random assignment of subjects in a naturalistic setting (e.g., voters) to treatment and control conditions. These studies proved relatively easy to classify. The distinction between survey and lab experiments was more ambiguous, as a number of experiments that are conducted in a laboratory setting follow identical procedures to survey experiments with the exception of the location in which the survey was conducted (i.e., surveys conducted in a lab setting that might have without substantial alteration been administered face-to-face, over the phone, or via the Internet). Because of this similarity, we defined survey experiments broadly to include not only those administered via a traditional survey mode, but also any experiment carried out in a laboratory that could have been conducted in an identical fashion through one of the common survey formats.[4] As such, "laboratory experiments" were limited to those studies conducted in a laboratory environment that involved viewing video, any interaction of subjects with other participants, multiple waves (e.g., a pre-test survey two weeks before the experimental manipulation of interest), and/or a cognitive distractor task that could not have been completed through an alternative surveying method (e.g., math problems calculated by hand).

From the collection of the 490 experimental articles we identified, we drew a random sample of 60 for coding. We employed stratified sampling based on two factors in addition to laboratory, field, or survey. To see whether reporting standards improved as experiments became more common in the discipline, we classified articles as "early" (2005–2009) or "late" (2010–2013). A second distinction was "general interest" versus "field" journals, which permits the determination of whether reporting differs between what are sometimes considered to be the top three general interest journals (*American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics*) and those that tend to be more specialized in nature (the other 13 journals examined). This results in $2 \times 2 \times 3$ distinct cells and is well-suited for comparison across time period, journal type, and experiment type.[5]

Starting with the full proposed reporting guidelines (see Appendix 1), we developed a coding scheme to measure the completeness of reporting. We boiled the guidelines down to 19 reporting items that we took as representative of the information that should be reported in any experimental research report. Table 1 presents a brief description of how these 19 items were translated into coding standards. The items range from things that seem quite important, such as a complete description of the treatments, to items that would be useful to know and are indications of the level of care in reporting exercised by the authors, such as reporting the response rates to any surveys the researchers administer. For each article, adherence to each of the 19 reporting items was coded on a three-point scale

[3]Five articles were coded as employing multiple types of experiments and were excluded from the analyses.

[4]Experiments that test the utility of different survey modes (e.g., face-to-face versus over the Internet) were classified as survey experiments as well.

[5]See Table A1 in the supplementary appendix for counts from each cell.

Table 1
Reporting Guidelines Used in Coding Analysis

Subjects and Context
Did the author(s) report who was eligible to participate in the study?
Did the author(s) report dates defining the periods of recruitment and when the experiments were conducted? Did the authors report the dates of any repeated measurements as part of a follow-up?
If a survey: Did the author(s) identify the survey firm used and describe how it recruits respondents if the survey firm is not well known?
If a survey: Did the author(s) provide the response rate and how it was calculated?
Allocation Method
Did the author(s) report whether random assignment was used?
If random assignment was used, did the author(s) report the unit of randomization (individuals, groups, households, etc.)?
Did the author(s) provide a table (in the text or an appendix) showing baseline means and standard deviations for demographic characteristics and other pretreatment measures (if collected) by experimental group?
Treatments
Did the author(s) report what treatment was given to each of the treatment groups and what was given to the control group?
Did the author(s) make the complete treatment materials available? If treatments are scripts, did the author(s) make the exact scripts available? If the treatments are question wordings, did the author(s) make exact variations in question wordings available? If the treatments are mailings, did the author(s) make sample mailings available?
Measurement
Did the author(s) report how the outcome variables are measured and coded?
If there is an index used, did the author(s) report exactly how it was constructed?
Did the author(s) report how all other variables included in the statistical models are measured and coded?
CONSORT Diagram Information
Did the author(s) report the number of subjects initially assessed for eligibility for the study?
Did the author(s) report exclusions prior to random assignment and the reasons for the exclusions?
Did the author(s) report the number of subjects assigned to each experimental group?
Did the author(s) report the proportion of each group that received its allocated intervention and the reasons why subjects did not receive the intended intervention?
Did the author(s) report the number of subjects in each group that dropped out or for other reasons does not have outcome data?
Did the author(s) report the number of subjects in each group that is included in the statistical analysis, and the reasons for any exclusions?
Statistical Analyses
Did the author(s) report sample means and standard deviations, and Ns for the outcome variables using intent-to-treat (ITT) analysis (i.e., means and standard deviations for the entire collection of subjects assigned to a group, whether a treatment is successfully delivered or not)?

(0–2) that denoted the extent and quality of the reporting vis-à-vis the ability to adequately evaluate the research design and analyses. A score of "0" is awarded when no pertinent information is provided, a "1" corresponds to some information but important omissions, and a "2" signifies either trivial or no omissions. Given the meaning assigned to each value, the coding scheme is somewhat lenient in awarding a "1" (as opposed to a "0") but rather stringent with the assignment of a "2." Thus, a "1" corresponds to a wide range of missing information (in terms of severity)

while "0" and "2" are more narrowly defined as completely missing and (almost) completely reported, respectively.[6]

Results

Table 2 shows the percentage of articles with missing information (i.e., those coded as a "0" or "1") for the 19 reporting items, both for the overall sample (column 8) and broken down by various sampling strata. The results are based on coding the sample of 60 articles. Five articles were drawn at random from each of the 12 cells (experiment type × journal type × time period); two articles were assigned to coder 1, two to coder 2, and the fifth to both coders to provide an estimation of the consistency between their efforts. This design thus yields 72 observations on the reporting standards of 60 experimental articles (48 independently analyzed and 12 coded by both coders). In the results presented in Table 2, a jointly coded article is considered to have missing information only if both coders come to this conclusion.[7]

Although most articles in the discipline's top general interest and field journals report most items, the columns reveal substantial missing information in reporting across all types of experiments, in both general and specialized journals, and in both the early and late periods. Before we discuss the prevalence of missing information, it is useful to understand what sort of events triggered such a coding. The final column of Table 2 describes, for each item, an example of incomplete reporting found in the sample. Omissions regarding the subject selection that triggered a coding less than "2" included leaving out an explanation of how subjects were recruited, omitting details about the subject characteristics (such as year or major for college students), or excluding details about the timing of subject recruitment and conduct of the experiment (such as what year or time of year the experiment took place). Omissions regarding subject characteristics included failure to report any sample means for some variables used in the analysis and failure to report baseline covariate levels or balance across experimental groups. Omissions regarding the treatments involved failure to provide scripts or images for all the treatments. Omissions regarding the

[6]Standards deemed as inapplicable to the particular study (e.g., there was no survey and so a response rate was not reported) were not given a value.

[7]The coders worked first on the jointly assigned twelve articles, independently determining the degree to which each met the prescribed reporting standards. After completing this task, their results were compared (and are presented in Table B1 of Appendix 2). We use the average value for the jointly coded articles (as opposed to treating them as two separate observations), though treating the coding as two distinct observations does not change our reported results (see Table B2 of Appendix 2). Column 1 of Table B1, which reports the proportion of cases that matched exactly, reveals a high level of agreement between the two efforts. Of the 19 standards coded, 16 shared the same score at least 75% of the time. The congruence further improves when we combine those coded as "0" or "1," creating a complete- or partial-omission versus no-omission distinction. Based on this comparison, column 3 shows that the proportion of matches is less than 75% for only one of the standards. For those items on which column 1 reports an agreement rate of less than three-quarters, the coders again met with one of the authors to discuss issues that arose in the coding process and further develop clearly defined guidelines under which to code the remaining articles. Coders then evaluated the remaining articles. Further details of this coding process are presented in Appendix 2C, and a full coding report is presented in Appendix 2D.

Table 2
Percent of Articles Receiving Any Downgrade by Category

Coding category	Experiment type			Journal type		Time period		Overall	Example
	Survey	Lab	Field	General	Specialized	Early	Late		
Did the author(s) report who was eligible to participate in the study?	5.0 20	50.0 20	10.0 20	16.7 30	26.7 30	20.0 30	23.3 30	21.7 60	Undergrads from a specified institution with no info on how recruited, age, year in college, etc.
Did the author(s) report dates defining the periods of recruitment and when the experiments were conducted?	20.0 20	50.0 20	10.0 20	33.3 30	20.0 30	33.3 30	20.0 30	26.7 60	No dates reported and no general sense of time period (e.g., season, year, semester, etc.)
(If a survey:) Did the author(s) identify the survey firm used and describe how it recruits respondents?	7.1 14	0.0 1	100.0 3	30.0 10	12.5 8	33.3 9	11.1 9	22.2 18	Survey firm not reported; numbers called from a directory but no info on how it was created
(If a survey:) Did the author(s) provide the response rate and how it was calculated?	100.0 14	100.0 1	33.3 3	100.0 10	75.0 8	77.8 9	100.0 9	88.9 18	Internet survey that fails to report response rate or number contacted to get final sample size
Did the author(s) report whether random assignment was used?	0.0 20	0.0 20	0.0 20	0.0 30	0.0 30	0.0 30	0.0 30	0.0 60	N/A
If random assignment was used, did the author(s) report the unit of randomization?	5.0 20	0.0 19	0.0 20	0.0 29	3.3 30	0.0 29	3.3 30	1.7 59	State in different places that individual and household are unit of randomization
Did the author(s) report baseline means and standard deviations for pretreatment measures by experimental group?	95.0 20	85.0 20	75.0 20	86.7 30	83.3 30	90.0 30	80.0 30	85.0 60	Fail to do so when clear (from later analyses) that such info was collected and/or exists

Table 2
(continued)

Coding category	Experiment type			Journal type		Time period		Overall	Example
	Survey	Lab	Field	General	Specialized	Early	Late		
Did the author(s) report what treatment was given to each of the treatment and control groups?	0.0 20	0.0 20	0.0 20	0.0 30	0.0 30	0.0 30	0.0 30	0.0 60	N/A
Did the author(s) make complete treatment materials available?	10.0 20	30.0 20	35.0 20	23.3 30	26.7 30	33.3 30	16.7 30	25.0 60	Fail to make lab instructions or all scripts/survey vignettes available (only some provided)
Did the author(s) report how the outcome variables are measured and coded?	10.0 20	5.0 20	0.0 20	6.7 30	3.3 30	6.7 30	3.3 30	5.0 60	Report survey response outcome as range over certain values but do not report value labels
(If an index was used:) Did the author(s) report exactly how it was constructed?	0.0 4	0.0 1	--- 0	0.0 5	--- 0	0.0 2	0.0 3	0.0 5	N/A
Did the author(s) report how all other variables included in the statistical models are measured and coded?	18.8 16	17.7 17	10.5 19	20.0 25	11.1 27	16.0 25	14.8 27	15.4 52	Report for some or most but not all variables
Did the author(s) report the number of subjects initially assessed for eligibility for the study?	60.0 20	65.0 20	40.0 20	56.7 30	53.3 30	50.0 30	60.0 30	55.0 60	Survey experiment that does not report number initially contacted to get final sample size
Did the author(s) report exclusions prior to random assignment and the reasons for the exclusions?	10.0 20	15.0 20	15.0 20	20.0 30	6.7 30	16.7 30	10.0 30	13.3 60	Fail to report reasons for exclusion (could be multiple) and exact numbers excluded

Table 2
(continued)

Coding category	Experiment type			Journal type		Time period			Example
	Survey	Lab	Field	General	Specialized	Early	Late	Overall	
Did the author(s) report the number of subjects assigned to each experimental group?	65.0 20	30.0 20	10.0 20	43.3 30	26.7 30	30.0 30	40.0 30	35.0 60	No exact numbers (e.g., "roughly equal" or "between X and Y assigned to each group")
Did the author(s) report the proportion of each group that received its allocated intervention and the reasons why some did not?	15.0 20	0.0 20	20.0 20	13.3 30	10.0 30	16.7 30	6.7 30	11.7 60	Treatment administered at multiple sessions; some do not attend all but number not reported
Did the author(s) report the number of subjects in each group that does not have outcome data?	85.0 20	50.0 20	40.0 20	53.3 30	63.3 30	56.7 30	60.0 30	58.3 60	Overall number or general estimate reported but not by group
Did the author(s) report the number of subjects in each group that is included in the statistical analysis, and the reasons for any exclusions?	80.0 20	65.0 20	60.0 20	60.0 30	76.7 30	63.3 30	73.3 30	68.3 60	Only overall number reported and no reasons provided for exclusions that occur at this point
Did the author(s) report sample means, standard deviations, and Ns for the outcome variables using intent-to-treat analysis?	75.0 20	80.0 20	70.0 20	73.3 30	76.7 30	73.3 30	76.7 30	75.0 60	Only report regression models with covariates (not regression with treatment indicators only)

Note: Cell entries for each coding category are the percentage of articles that received a score of zero or one with frequencies presented below. Frequencies differ when categories were coded as not applicable for certain articles.

outcome variable involved failure to report the exact question or response options used to code the outcome. Omissions regarding attrition included failure to report attrition by experimental group or to provide any explanation for why some subjects are missing from the statistical analysis calculating treatment effects.

Column 8 of the data in Table 2 lists the rates of missing information in reporting for the overall sample. There are several main findings. First, it appears that the use of random assignment and some description of the assignment process were almost universally reported. Second, authors frequently do not provide basic data tables, such as pretreatment means for important covariates or mean outcomes by experimental group following treatment. Third, although a basic description of the treatments is always provided, important details are frequently missing, and complete materials are not provided in the paper or a linked appendix a quarter of the time. Fourth, it is often difficult to learn the response rates for surveys (when used), the number of subjects initially assessed for eligibility, the number of subjects assigned to each experimental group, the number of subjects in each group that is missing outcome data, and the reasons why subjects that were assigned to treatment groups are missing from the statistical analysis.

Table 2 also presents the breakdown of reporting omissions by type of experiment, journal type, and time period. Although there is some interesting variation across type of experiment, we restrict our comments to journal type and year of publication. First, the reporting patterns for the general interest and the specialized journals (columns 4 and 5 of Table 2) are quite similar. A priori, it might have been expected that the general interest journals in our sample would adhere to higher reporting standards because of stricter standards imposed by peer review. Alternatively, we may have expected some of the specialized field journals to adhere to higher reporting standards because certain fields have a longer history of using experimental methods. Neither of these propositions, however, is borne out by the data. The largest difference in coding scores between general and specialized journals is 25 percentage points in the case of reporting the response rates from surveys, with an average difference of about 8.6 points across coding categories.[8] Higher rates of missing information occur on 10 standards in general journals and six standards in specialized journals (with rates being the same on the remaining two standards). As such, we cannot conclude that experimental studies published in either general or specialized journals adhere to stricter reporting standards.

Second, and somewhat surprisingly, reporting appears similar for the 2005–2009 and 2010–2013 periods (columns 6 and 7 of Table 2). For only four of the 19 standards does the difference in the percentage of articles with missing information exceed 10 points. The largest disparity between the early and late periods is about 22 percentage points for the two measures concerning the use of private survey firms (not surprising, given the much smaller sample sizes for which these categories were relevant), with an average difference of less than eight points across coding

[8]This average calculation excludes the category of whether the author(s) reports how an index was constructed, as there were no data in our sample of specialized journals for this category.

Table 3
Percent of Articles with Omissions

	Experiment type			Journal type		Time period		
	Survey	Lab	Field	General	Specialized	Early	Late	Overall
Percent with three or more substantial omissions (major or complete)	90.0 20	90.0 20	85.0 20	90.0 30	86.7 30	93.3 30	83.3 30	88.3 60
Percent with six or more substantial omissions (major or complete)	50.0 20	40.0 20	15.0 20	36.7 30	33.3 30	30.0 30	40.0 30	35.0 60
Percent with three or more complete omissions (standard coded as "0")	55.0 20	30.0 20	10.0 20	36.7 30	26.7 30	30.0 30	33.3 30	31.7 60

Note: Cell entries are percentage of articles with omissions, with frequencies presented below. Major omissions are standards coded as a "1," while complete omissions are standards coded as a "0." For jointly coded articles, the average score is rounded up if necessary (i.e., an article's standard assigned a "1" by one coder and a "2" by another (for an average score of "1.5") is treated as a "2" and thus not considered a substantial omission, while an article's standard assigned a "0" by one coder and a "1" by another (for an average score of "0.5") is treated as a "1" and thus a major but not complete omission). Only one article (a lab experiment in a general journal from the late time period) exhibited six or more complete omissions.

categories. Once again, we find that there is no general sense in which studies in the early or late time periods are adhering to more stringent reporting standards. In eight of the 19 coding categories, a greater percentage of studies from the late period have important omissions, while the same is true for eight of the 19 categories for studies from the early period (with scores in the remaining three categories tied).

Table 3 considers the sample by article rather than by reporting category. It appears that the omissions reported in Table 2 are not all confined to a minority of papers that report very little. Rather, the substantial majority of papers leave out useful information. Approximately one-third of all articles have three or more items coded 0, signifying complete or nearly complete omission of a reporting item. We see that there is very little difference across journal type (general versus field) or time period (early versus late), but substantial difference across experiment type (more omissions in survey and lab than field experiments). Overall, 88.3% of the articles had three or more items and 35% had six or more items with more than trivial reporting omissions (i.e., coded 0 or 1).

To summarize the findings, most items are reported most of the time. However, there are substantial omissions, and these omissions occur in all types of experimental research, in both general interest and field journals, and reporting quality does not appear to be improving over time.

THE COMMITTEE AND PROCESS

To assist authors and readers, the Standards Committee composed reporting guidelines. The process used to draft the reporting guidelines was designed to

represent the major experimental methodologies and to permit both interaction among the committee members and also feedback from the section membership. The committee consisted of six members: Alan Gerber (Chair, Yale University), Kevin Arceneaux (Temple University), Cheryl Boudreau (University of California, Davis), Conor Dowling (University of Mississippi), Sunshine Hillygus (Duke University), and Thomas Palfrey (Caltech). The committee held three conference calls and exchanged numerous e-mails from early 2011 to early 2012, after which they submitted the proposed guidelines to the section president. The guidelines were also posted on the section website and section members were informed of this, told there would be a comment period, and asked for comments. Following the comment period, the committee reviewed whether, in light of the comments, any changes to the proposed reporting standards were needed. After this review, the guidelines were finalized. The version of the guidelines that emerged from this process appears at the end of this report (Appendix 1).

CREATION OF THE INITIAL GUIDELINES

During the first meeting, the committee divided up the territory according to committee expertise. Arceneaux created a list of proposed reporting guidelines for field experiments; Boudreau and Palfrey for laboratory experiments; Hillygus for survey experiments. Prior to the second meeting, Dowling and Gerber assembled the various lists, along with reporting standards from three other fields—(1) medicine (CONSORT[9]), (2) economics (e.g., *Econometrica*[10], Palfrey and Porter (1991); *American Economic Review*[11]; *Journal of Political Economy*[12]), and (3) psychology (American Psychological Association, JARS (2008))—for the committee to discuss.

During the second meeting, the committee reviewed the various reporting standards (both our internally created documents and those from the other three fields) and, because it was felt there was sufficient overlap across types of experiments, we decided to craft a single set of reporting standards. After this second meeting, Boudreau consolidated the work of the committee members into a single document, which then went to each committee member for additional changes and comments. During the third meeting, the committee discussed this newly created document (any and all changes, whether additions or subtractions, were considered). Based on discussions held during this meeting, Dowling and Gerber made minor changes to the document and circulated the revised document to the committee for comments. The document, "Recommended Reporting Standards for Experiments (Laboratory, Field, Survey)," was unanimously approved by the committee and then posted to the section for comment.

[9]http://www.consort-statement.org/
[10]http://www.econometricsociety.org/submissioninstructions.asp#replication
[11]http://www.aeaweb.org/aer/data.php
[12]http://www.press.uchicago.edu/journals/jpe/datapolicy.html?journal=jpe

COMMENT PERIOD AND REVISIONS

During the comment period we received a comment that took issue with one guideline. Professor Diana Mutz (University of Pennsylvania) questioned a provision that was listed under a subsection of Part C, which recommended, "If random assignment used, provide evidence of random assignment." A bullet point under this heading recommended, "If demographic or other pretreatment variables were collected, [provide] a table (in text or appendix) showing baseline means and standard deviations for demographic characteristics and other pretreatment measures by experimental group."

Mutz objected that examination of group means, contrary to the implication of this guideline, is not a method of determining whether the experiment used random assignment. In particular, as detailed in a paper forwarded along with her comment (Mutz and Pemantle, 2013), she objects to formal balance testing in which researchers evaluate the integrity of the randomization procedure by inspecting whether there are statistically significant differences in the means of pretreatment variables across treatment groups. First, she argues that, apart from the rare blunder in implementation, if the experiment did employ random assignment then this was, in fact, the procedure employed. The hypothesis that the groups have been randomly assigned is not an open question to be adjudicated by statistical analysis, since differences in covariate balance can arise by chance in any given random sample. Mutz does agree that formal tests or their rough ocular equivalents may be useful to detect errors in the implementation of randomization, but contends that these errors are uncommon. Of more concern to Mutz is the possibility that formal balance tests may lead researchers to make inadvisable modeling decisions, such as including covariates that may reduce the precision of the treatment effect estimates. For more detailed discussion of these issues, we recommend that readers consult Professor Mutz's paper.

We considered Mutz's arguments and although the cautions raised are well taken, we decided that on balance the value of the information contained in a table of pretreatment means and standard deviations outweighs the dangers Mutz highlights. First, and sufficient for our recommendation that researchers be asked to report covariate sample statistics, observing unexpectedly large differences in group means or variances can signal problems in data collection or analysis. Detectable imbalances can be produced in several ways (other than chance). These include, but are not limited to, mistakes in the randomization coding, failure to account for blocking or other nuances in the experimental design, mismatch between the level of assignment and the level of statistical analysis (e.g., subjects randomized as clusters but analyzed as individual units), or sample attrition.

Although the usefulness of the sample statistics for purposes of error detection is sufficient for their inclusion in the guidelines and the basis of our judgment, we add in passing that, for some readers, there may be other uses of summary statistics for covariates for each of the experimental groups. For instance, if there is imbalance,

whether statistically significant or not, in a pretreatment variable that is thought by a reader to be highly predictive of the outcome, and this variable is not satisfactorily controlled for, the reader may want to use the baseline sample statistics to informally adjust the reported treatment effect estimates to account for this difference. Finally, we note that the guidelines do not counsel any particular modeling response to the table of covariate means that we ask the researcher to provide. As stated earlier in this report, our guiding principle is to provide the reader and the reviewer the information they need to evaluate what the researcher has done and to update their beliefs about treatment effects accordingly. As such, the guidelines try to "break ties" in the direction of asking authors to show a table that might be useless, rather than omit information that could be of use to some reasonable share of researchers.

To avoid any confusion about our intentions, we decided to amend the guidelines in light of Professor Mutz's concerns by including a brief preface to this reporting item to elaborate a bit on the rationale for the table of sample means. We now write that:

"If random assignment used, to help detect errors such as problems in the procedure used for random assignment or failure to properly account for blocking, if demographic or other pretreatment variables were collected, provide a table (in text or appendix) showing baseline means and standard deviations for demographic characteristics and other pretreatment measures (if collected) by experimental group."

CONCLUSION

Any set of reporting guidelines is a snapshot that reflects a discipline's evolving norms and level of knowledge. Some issues that are ignored here may be considered significant in the coming years and some information that is requested by the guidelines may be considered insignificant. Although these guidelines draw on the experience of other disciplines, some features may be deemed by our scholarly community to be burdensome, pointless, or ill-considered. We encourage researchers to share their views and any relevant experiences they have with the guidelines that we have produced with committee members. We plan to gather these comments and, if needed, revise the guidelines after a year or so.

SUPPLEMENTARY MATERIAL

To view supplementary material for this paper, please visit http://dx.doi.org/10.1017/xps.2014.11.

REFERENCES

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (JARS). (2008). Reporting standards for research in psychology:

why do we need them? What might they be? *American Psychologist*, 63(9), 839–851.
Mutz, D. C., and Pemantle, R. (2013). The perils of randomization checks in the analysis of experiments. Typescript. University of Pennsylvania.
Palfrey, T., and Porter, R. (1991) Guidelines for submission of manuscripts on experimental economics. *Econometrica*, 59(4), 1197–1198.

APPENDIX 1: RECOMMENDED REPORTING STANDARDS FOR EXPERIMENTS (LABORATORY, FIELD, SURVEY)

This appendix describes recommended minimum reporting standards for experimental research. These are minimum standards and cannot anticipate all the particular things that are worth reporting in each study. Further, the reporting recommendations generally leave the question of how statistical analysis should be conducted and reported to the researcher, though there are several basic features of the data that we include as recommended minimum reporting standards. This appendix is taken verbatim from a document that was put together by the Standards Committee of the Experimental Research section of APSA. (The committee members were Alan Gerber (Chair), Kevin Arceneaux, Cheryl Boudreau, Conor Dowling, Sunshine Hillygus, and Tom Palfrey.) We view this as a tool for researchers who wish to communicate their work more effectively and a checklist that may be helpful to the researcher who wants to prepare a study that can be easily understood and evaluated. The standards are similar to the CONSORT reporting standards, which have been embraced by medical researchers conducting experimental research and are now the minimum reporting requirements for several major medical journals.

A. Hypotheses

- State specific objectives or hypotheses.
  - What question(s) was (were) the experiment designed to address?
  - What are the specific hypotheses to be tested?

B. Subjects and Context

- Report eligibility and exclusion criteria for participants.
  - Why was this subject pool selected? Who was eligible to participate in the study? What would result in the exclusion of a participant? Were any aspects of recruitment changed (such as the exclusion criteria) after recruitment began?
- Report procedures used to recruit and select participants.
- How were participants contacted for recruitment? Were incentives offered?
  - If there is a survey: Identify the survey firm used and describe how they recruit respondents.

- Report recruitment dates defining the periods of recruitment and when the experiments were conducted.
  - Also list dates of any repeated measurements as part of a follow-up.
- Describe settings and locations where the data were collected.
  - In the field, lab, classroom, or some other specialized setting?
  - Other relevant specifics of the population: e.g., large public university vs. small private university; geographic location; etc.
- If there is a survey: Provide response rate and how it was calculated.

C. Allocation Method

- Report details of the procedure used to generate the assignment sequence (e.g., randomization procedures).
- If random assignment used, report details of procedure (e.g., any restrictions, blocking).
  - Note the unit of randomization (individuals, groups, households, etc.). Pay careful attention to report clustered random assignment if subjects were assigned at some level other than the individual subject.
- If random assignment used, to help detect errors such as problems in the procedure used for random assignment or failure to properly account for blocking, provide a table (in text or appendix) showing baseline means and standard deviations for demographic characteristics and other pretreatment measures (if collected) by experimental group.
  - If blocking was used, and group assignment proportions were not equal across blocks, provide a table for each of the blocks. If there are too many blocks for this to be practical, combine blocks to present weighted averages of covariates using inverse probability weighting.
- Describe blinding.
  - Were participants, those administering the interventions, and those assessing the outcomes unaware of condition assignments?
  - If blinding took place, include a statement regarding how it was accomplished and how the success of blinding was evaluated.

D. Treatments

- Provide a detailed description of the interventions in each treatment condition, as well as a description of the control group.
  - Descriptions should be sufficient to allow precise replication: Summary or paraphrasing of experimental instructions in the article text; verbatim instructions and/or other treatment materials provided in an appendix.
- State how and when manipulations or interventions were administered.
  - Method of delivery: Pen-and-paper vs. computer or Internet vs. face-to-face communications vs. over the telephone.

- If computerized, the software should be described and cited. (If possible, programs should be included in an appendix so as to be available for purposes of replication.)
- For lab experiments (and other experiments, when relevant):
  - Report the number of repetitions of the experimental task and the group rotation protocol. Report the ordering of treatments for within-subject designs. Any piggybacking of other protocols should be reported. Report any use of experienced subjects or subjects used in more than one session or treatment.
  - Report time span: How long did each experiment last? How many sessions were subjects expected to attend? If there were multiple sessions, how much time passed between them?
  - Report total number of sessions conducted and number of subjects used in each session.
  - Report whether deception was used.
  - Report treatment fidelity: Evidence on whether the treatment was delivered as intended.
    - Report any instructional anomalies or inaccuracies.
    - Were subjects given quizzes on the experimental instructions?
    - Were there practice rounds? If so, how many and what were the results?
    - Did subjects complete a post-experiment debriefing, interview, or questionnaire? If so, is there evidence that subjects understood the instructions and treatments?
    - Did the experimental team observe aspects of the intervention?
    - Provide descriptions of manipulation checks, if any.
  - Were incentives given? If so, what were they and how were they administered?

E. Results

1. Outcome Measures and Covariates

- Provide precise definitions of all primary and secondary measures and covariates.
  - For indices, provide exact description of how they are formed. For survey items, provide exact question wording in an appendix. Provide a copy of the complete survey questionnaire (in an online appendix if it is long).
- Clearly state which of the outcomes and subgroup analyses were specified prior to the experiment and which were the result of exploratory analysis.

2. CONSORT Participant Flow Diagram

- Complete CONSORT Participant Flow Diagram

- An example of a CONSORT flow diagram can be found at http://www.consort-statement.org. The flow diagram records the initial number of subjects deemed eligible for the experiment and all losses of subjects during the course of the experiment. The flow chart follows the subjects from initial recruitment to the sample used in the main analyses, providing readers clear information on the amount of attrition and exclusions. The chart also reports the portion of each treatment group that received the allocated intervention and if not, why this was not accomplished. Naturally, in the event that there is zero or very trivial non-compliance with group assignment or zero or very trivial attrition, researchers may decide it is more convenient to report the information that would otherwise be shown in the CONSORT diagram in the text and omit the diagram.

Note that the CONSORT flow chart entries include:

- Number of subjects initially assessed for eligibility for the study.
- Exclusions prior to random assignment and reasons for the exclusions.
- Number of subjects initially assigned to each experimental group.
- The proportion of each group that received its allocated intervention and the reasons why subjects did not receive the intended intervention.
- The number of subjects in each group that dropped out or for other reasons do not have outcome data.
- The number of subjects in each group that are included in the statistical analysis, and the reasons for any exclusions.

3. Statistical Analysis

- Researchers will conduct statistical analysis and report their results in the manner they deem appropriate. We recommend that this reporting include the following:
  - Report sample means and standard deviations for the outcome variables using intent-to-treat (ITT) analysis (means for the entire collection of subjects assigned to a group, whether the treatment is successfully delivered or not).
    - If the experiment uses block randomization with unequal assignment rates, present ITT analysis by block or present overall means using inverse probability weighting.
  - Note whether the level of analysis differs from level of randomization and estimate appropriate standard errors.
  - If there is attrition, discuss reasons for attrition and examine whether attrition is related to pretreatment variables.
  - Report other missing data (not outcome variables):
    - Frequency or percentages of missing data by group.
    - Methods for addressing missing data (e.g., listwise deletion, imputation methods).

    - For each primary and secondary outcome and for each subgroup, provide summary of the number of cases deleted from each analysis and rationale for dropping the cases.
  - For survey experiments: Describe in detail any weighting procedures that are used.

F. Other Information

- Provide additional information about the experiment.
  - Was the experiment reviewed and approved by an IRB?
  - If the experimental protocol was registered, where and how can the filing be accessed?
  - What was the source of funding? What was the role of the funders in the analysis of the experiment?
  - Were there any restrictions or arrangements regarding what findings could be published? Are there any funding sources where conflict of interest might be an issue?
  - If a replication data set is available, provide the URL.