

Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods

Diana C. Mutz* and Robin Pemantle†

Abstract

In this essay, we more closely examine three aspects of the Reporting Guidelines for this journal, as described by Gerber and colleagues (2014, *Journal of Experimental Political Science* 1(1): 81–98) in the inaugural issue of the *Journal of Experimental Political Science*. These include manipulation checks and when the reporting of response rates is appropriate. The third, most critical, issue concerns the committee's recommendations for detecting errors in randomization. This is an area where there is evidence of widespread confusion about experimental methods throughout our major journals. Given that a goal of the *Journal of Experimental Political Science* is promoting best practices and a better understanding of experimental methods across the discipline, we recommend changes to the Standards that will allow the journal to play a leading role in correcting these misunderstandings.

Keywords: Randomization check, manipulation check, response rates, standards.

Establishing reporting guidelines for studies of any kind is an important step in the direction of improving research. The Standards Committee is to be commended for taking on this difficult and time-consuming task for experimental designs (see Gerber et al. 2014). We have no doubt that this is a positive development, something good for science as a whole, as well as for our particular discipline.

Nonetheless, in the spirit of making something that is already quite good even better, we would like to highlight some of the problems with the standards as currently written. This is not to detract from the committee's accomplishment, but is offered in the spirit of constructive suggestions for revision.

We discuss three aspects of the recommendations. The first concerns manipulation checks, a practice of great importance in experimental methodology that is not addressed by the Standards. The second is a more minor point concerning the

*Political Science and Communication, University of Pennsylvania, Philadelphia, PA USA; email: mutz@sas.upenn.edu

†Department of Mathematics, University of Pennsylvania, Philadelphia, PA USA

28 reporting of response rates, and how studies become classified as surveys under their
29 suggested framework. The third issue concerns the recommendations for detecting
30 errors in randomization. Our critique and suggestions for improvement on this front
31 require substantial background, so we save this more involved issue for last.

32 **MANIPULATION CHECKS**

33 First, we recommend that manipulation checks be added to the JEPS checklist of
34 desirable components of experiments. As is the case with many other items on the
35 checklist, this requirement will not be relevant to every experiment, but it will be
36 applicable to a large number of them and, most importantly, it will improve what
37 can be learned from their results.

38 Manipulation checks establish that the treatment has had an effect on the
39 theoretically relevant causal construct. In other words, manipulation checks are
40 “a way of ensuring that an experiment actually has been conducted (i.e., that
41 the IV has been effectively manipulated)” (Sansone et al. 2008). The majority
42 of experiments in political science do not report manipulation checks, despite
43 their prominence in other social science disciplines. Many social science disciplines
44 have deemed them basic enough to be required in all but a limited number of
45 cases. As a sociology volume on experimentation argues, “It is an essential part
46 of an experiment to include manipulation checks. . . . It is equally important
47 to report the results of these checks” (Foschi 2007: 129). The *Handbook of*
48 *Methods in Social Psychology* similarly advises, “Indeed, many editors of social
49 psychology journals require these (manipulation checks) to be conducted as
50 a matter of principle before accepting research for publication.” While some
51 kinds of experiments within political science do not include variable experimental
52 treatments at all (e.g., game theoretic experiments), a majority do involve one or
53 more randomly assigned treatments intended to induce variation in the causal
54 variable.

55 In some cases, manipulation checks are unnecessary. For example, if a persuasion
56 experiment manipulates the length of a message in order to evaluate whether long
57 messages tend to be more persuasive than short ones, and one message has twice
58 the number of words as another, then length has been manipulated, and it need
59 not be the case that subjects recognize or remember the length of the argument
60 to which they were exposed. Given that the independent variable construct and
61 its operationalization are completely identical, a manipulation check would be
62 unnecessary under these conditions.

63 The problem with assuming that the independent variable is identical to its
64 operationalization is that this is frequently not the case. Nonetheless, in political
65 science the experimental treatment is usually just assumed to have successfully
66 altered the independent variable, and the results are interpreted as such. For example,
67 when an experimental treatment suggesting “many people believe that . . . trade

68 can lead to lower prices for consumers,” did not lead to more support for trade,
69 the author concluded that it would not be worthwhile to convince people that trade
70 lowers the costs of consumer goods in order to increase support for trade (Hiscox
71 2006: 756). Without knowing whether subjects actually believed this treatment, null
72 effects cannot be distinguished from weak or ineffective manipulations. Likewise,
73 when political scientists look for causal effects from policy threat, the salience of
74 national identity, incivility, or innumerable other treatments, the causal construct is
75 not identical to the operationalization of the treatment, so without a manipulation
76 check, there is no reason to assume that the causal construct was successfully
77 manipulated. Moreover, researchers have a tendency to underestimate the strength
78 of treatment that is required to produce a change in the independent variable.
79 As a result, an experiment may not actually test its hypothesis of interest. As the
80 *Handbook* further notes, “For experiments to have the best chance of succeeding
81 (i.e., for the IV to have an effect on the DV) the researcher needs to ensure that the
82 manipulation of the IV is as strong as possible. Indeed, if there were a first rule of
83 experimentation, this might be it.”

84 Our recommendation in favor of systematically encouraging manipulation checks
85 goes beyond consistency with the experimental traditions established in other
86 disciplines. It also stems from our belief that (1) consistently effective treatments are
87 a highly unrealistic assumption, and that (2) the absence of manipulation checks
88 frequently impedes the accumulation of scientific knowledge from experimental
89 studies in political science. We begin by delineating the conditions under which
90 manipulation checks seem essential in political science experiments and then
91 illustrate how their absence impedes scientific knowledge.

92 Independent variables in experimental studies may involve latent constructs that
93 are manipulated only indirectly, as described above, or direct treatments in which the
94 treatment and the operationalization are one and the same. Manipulation checks
95 are essential to ensure construct validity when treatments are indirect manipulations
96 of other constructs (Cozby 2009; Perdue and Summers 1986). Without verifying
97 successful manipulation of the independent variable in the studies, even
98 outcome effects consistent with the original hypothesis become difficult to
99 interpret.

100 The reason that manipulation checks have not been emphasized in experimental
101 political science may stem from the nature of early studies in this field, which tended
102 to examine tangible rather than latent constructs as independent variables. Does
103 a baby care booklet sent from one’s congressional representative improve attitudes
104 toward the elected official? (Cover and Brumberg 1982). Can information on voter
105 registration improve turnout? (Gosnell 1942). So long as the operationalization
106 of the treatment and the independent variable are one and the same, there was no
107 need for a manipulation check.

108 But as experiments in political science have become far more ambitious, frequently
109 using indirect strategies to manipulate latent independent variables, a smaller

110 proportion of independent variables meet these criteria, as Newsted and colleagues
111 (1997: 236) suggest.

112 Even in cases in which manipulations appear obvious, they may not be so. For
113 example, some early research in information presentation used treatments that
114 confounded the information form (e.g., table or graph) with other factors, such as
115 color, making interpretations difficult. Manipulations checks can help to uncover
116 such problems and should be as much a part of the development of a measurement
117 strategy in an experiment as the dependent variables.

118 Unfortunately, even when the operationalization of a given independent variable
119 is well-known, widely established and frequently used, there is still no guarantee
120 that one has successfully manipulated the independent variable in any given study.
121 For example, a widely used cognitive load manipulation appears to be responsible
122 for highly inconsistent results in studies of how cognitive load affects charitable
123 donations. In Kessler and Meier's (2014) careful replications of laboratory studies
124 using the same subject pool, setting and experimental protocol, they discovered
125 that the explanation for contradictory findings was that the manipulation varied
126 in efficacy due to session order effects when multiple experiments were executed
127 within a single hour-long session. The treatments produced the intended variance
128 in the independent variable only when subjects were already somewhat fatigued.
129 Thus, even with a well-established treatment, manipulation checks were essential
130 to the correct interpretation of the experimental findings. In cases such as these,
131 manipulation checks clearly contribute to researchers' ability to differentiate among
132 competing interpretations.

133 Encouraging manipulation checks is particularly important in an era when survey
134 experiments have enjoyed increased popularity. When survey experiments have
135 participants respond from remote, unobservable locations, there is no way to know
136 for certain if subjects were even exposed to the treatment, let alone whether they were
137 affected in the way the investigator intended. Particularly with large heterogeneous
138 population samples who may not pay as much attention to treatments administered
139 online as they would in a lab, treatments can easily fail.

140 Simple exposure to a treatment obviously does not guarantee its effectiveness. The
141 question is not whether "a treatment was successfully delivered," as indicated in the
142 current guidelines, but instead whether the treatment manipulated the independent
143 variable as intended. Subjects may doubt the veracity of information they are
144 given, or they may not find the treatment as threatening, anxiety-inducing, or as
145 counter-attitudinal (or whatever the treatment happens to be) as the investigator
146 intended.

147 Within political science there already has been some recognition of the problem
148 of inattentive respondents. For example, Berinsky et al. (2014) suggest that
149 studies should include post-treatment questions about details of the stimulus
150 in order to assess respondents' levels of awareness and attention to stimuli.
151 However, in most studies, use of such "screeners" does not address the same

152 question as a manipulation check. Rather than address whether a subject was
153 *exposed* to the treatment, manipulation checks are designed to assess whether
154 the treatment successfully induced variance in the independent variable. For
155 this reason, even laboratory experiments in which exposure is assured require
156 manipulation checks. While use of screeners seems reasonable, they are not a
157 substitute for manipulation checks. Although there are some studies for which
158 exposure to a stimulus is, in fact, the independent variable construct, most studies
159 use treatments to induce a change in a latent construct, a change that may,
160 or may not, have been accomplished among those who correctly answered a
161 screener.

162 Surprisingly, even some studies using treatments that seem obvious rather
163 than latent, such as the race of a person in a photograph, demonstrate
164 that such treatments can easily fail. For example, respondents often disagree
165 about the race of a person in a picture they are shown (Saperstein and
166 Penner 2012). For this reason, whatever the particular manipulation was
167 *intended* to convey should be verified before any meaningful conclusions can be
168 drawn.

169 Finally, although we have made the positive case for including manipulation
170 checks as part of the checklist in studies using latent independent variables, it is
171 worth considering whether there is any potential harm that should be considered in
172 encouraging them. We know of no one who has argued that they are harmful to the
173 integrity of a study so long as they are asked after the dependent variable is assessed.
174 Scholars across the social sciences concur that so long as manipulation checks are
175 included after measurement of the dependent variable, there is no potential harm in
176 including them. The only cost is in the respondent time spent on the manipulation
177 check assessment.

178 But there is substantial danger if one chooses to omit a manipulation check. The
179 risk is primarily of Type II error. The reader of a study without a manipulation
180 check has no idea if a null finding is a result of an ineffective or insufficiently
181 powerful treatment, or due to a theory that was simply incorrect. This is an
182 extremely important distinction for purposes of advancing scientific knowledge.
183 A recent article in *Science* demonstrates why this is problematic. Franco and
184 colleagues (2014) use the TESS study database as a source of information on
185 the file drawer problem, that is, the extent to which findings that do not achieve
186 $p < 0.05$ are less likely to see the light of publication. In order to estimate
187 how likely null findings were to be published, they analyzed all TESS studies
188 tracking whether the anticipated effect on the dependent variable was found
189 or not found, and classified as null findings those that did not produce effects
190 on the dependent variable. However, in many, and perhaps even most of these
191 latter cases, the independent variable was not verified as having been successfully
192 manipulated. Thus, the lack of significant findings was not informative with respect
193 to the theories under investigation or with respect to their anticipated effect
194 sizes.

195 We would not recommend including manipulation checks as part of the JEPS
196 checklist if they were informative on only rare occasions. But weak or ineffective
197 manipulations are an exceedingly common problem. A study that finds no significant
198 effects on the dependent variable and does not include a manipulation check for a
199 latent variable is not at all informative; on the other hand, we can learn a great deal
200 from an identical study with identical results that includes a manipulation check
201 documenting a successful treatment. In the latter case, the theory is clearly in need
202 of revision. Ignoring manipulation checks thus impedes the growth of scientific
203 knowledge.

204 Particularly given that JEPS has publicly stated its intent to publish
205 null results (a decision that we wholeheartedly endorse), it is essential to
206 encourage manipulation checks whenever possible. Otherwise, a null result is
207 not informative. More specifically, this practice inflates the possibility of Type
208 II error and leads researchers to prematurely abandon what may be viable
209 hypotheses. Wouldn't some other researcher recognize this potential problem
210 and attempt a replication? There are already strong disincentives to replicate
211 even significant experimental findings; the idea that researchers will pursue
212 replications of null results (but this time including a manipulation check) seems
213 improbable.

214 **REQUIREMENT OF RESPONSE RATES FOR SURVEYS**

215 A second issue concerns convenience samples. As currently written, the reporting
216 standards confuse mode of data collection with the type of sample used (probability
217 samples versus convenience samples). For purposes of applying these standards, the
218 committee defines a survey as any study that uses survey data collection methods
219 or that *could* conceivably have been executed as a survey, even if it was actually
220 executed in a laboratory.¹

221 For studies that qualify as surveys by virtue of their data collection method, the
222 Reporting Standards state, "If there is a survey: Provide response rate and how it
223 was calculated." The problem with applying this requirement to all studies that use
224 survey data collection methods is that many survey experiments in political science
225 use Mechanical Turk, Polimetrix or another opt-in data platform. There is nothing
226 meaningful about a response rate when utilizing a convenience sample. Even if
227 such a figure could be calculated, it would have no bearing on the quality of the
228 study. When all subjects opt in to a study, this concept is meaningless. Given that
229 the CONSORT diagram that the Standards Committee recommends (see Moher
230 et al. 2010; Schulz et al. 2010) already codifies the practice of indicating people who

¹An exception to this is that experiments that use video in a lab are classified as lab experiments even when they use survey methods to collect data (Gerber et al., 2014: 83). Given that videos are now also administered online as experimental treatments within surveys, this distinction is confusing.

231 drop out after a study has begun, attrition has already been covered in the other
232 requirements.

233 If there are no claims to representativeness being made by the authors, we see no
234 reason to require response rates. As many survey researchers have demonstrated,
235 the representativeness of a sample is not a straightforward function of the response
236 rate. If the authors are making claims about accurately representing some larger
237 population, then it would make sense to ask for a demographic comparison of
238 their sample to the population in question. But if the sample is being treated as
239 a convenience sample for purposes of an experiment, and not as a representative
240 one, then it is not informative to require response rates based on the means of data
241 collection used either in a lab or in an opt-in survey.

242 **RANDOMIZATION CHECKS/BALANCE TESTING**

243 Finally, our chief concern with the Standards has to do with the recommendation
244 on “Allocation Method” which addresses randomization procedure and the
245 distribution of pre-treatment measures. As the third point under Section C states,

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

250 This point contains a *directive*, “Provide a table . . .” as well as a *justification*, “to
251 help detect errors . . .” While we laud the goal of detecting errors, we find both the
252 directive and its connection to the justification problematic.

253 In order to understand our criticism of this recommendation, we begin by
254 clarifying some ambiguous uses of terms. We next discuss the role of randomization
255 in experimental design. Finally, we discuss the proper roles, if any, for balance
256 testing/randomization checks. Our discussion of the experimental method may
257 seem circuitous, but it is necessary because the mistaken pairing of the directive
258 and the justification produces potentially harmful consequences that are difficult to
259 grasp without understanding the possible motives behind such a recommendation.
260 Historically the adoption of randomization checks came first, while attempts at
261 justification have been more of an afterthought. Only by understanding the common
262 interpretation of this practice can one make sense of what is of value in this regard
263 and what is not.

264 **Terminology**

265 Throughout the recommendations and the accompanying report, four terms are
266 used to describe the “other variables” that are neither independent nor dependent
267 measures, but are the subject of the directive described above: pre-treatment
268 measures, covariates, demographics, and control variables. Whether or not the

269 authors of the Standards Report had precise meanings in mind for each of these,
270 both the report and our discussion of it will benefit from making these definitions
271 explicit.

272 For purposes of our discussion, the term “pre-treatment measure” is the most
273 self-evident, and we will assume it refers to any measure in the data set that is
274 assessed before the treatment occurs and thus could not have been affected by the
275 treatment. The term “covariate,” on the other hand, is typically used for that subset
276 of pre-treatment measures that are incorporated in the statistical model used to
277 test experimental hypotheses. It is not clear from the report whether “covariate” is
278 meant in this manner or is meant as a synonym for “pre-treatment measure.” This
279 confusion exists throughout political science (see, e.g., Arceneaux and Kolodny
280 2009: 760).

281 Importantly, this distinction becomes blurred if the model is not pre-specified; as
282 in the Standards Committee’s recommendations, we endorse the pre-specification
283 of models and will use the term “covariate” only for measures that researchers had
284 planned to include in the model in advance. As outlined in many sources (e.g.,
285 Franklin 1991), the purpose of a covariate is to predict variance in the dependent
286 variable that is clearly not attributable to the treatment. For this reason a covariate
287 must be a pretreatment variable, although not all pretreatment variables must be
288 included as covariates. Covariates need to be selected in advance based on what one
289 knows about the major predictors of the dependent variable in the experiment. Their
290 purpose is to increase the efficiency of the analysis model by eliminating nuisance
291 variance.

292 The term “demographic” is usually reserved for that subset of pre-treatment
293 measures which describe characteristics of the sort found on census data: age,
294 education, race, income, gender and the like. If they are to be used in the analysis
295 of an experiment, they should be pre-treatment measures as well. However, there is
296 no reason to include such measures as covariates unless one has reason to believe
297 they are strong predictors of the dependent variable. For most outcomes in political
298 science experiments, demographics are only weakly predictive at best. The purpose
299 of a covariate is to increase the power of the experiment by reducing variance. As
300 argued in Mutz and Pemantle (2011), the gain from adjusting by a weak predictor of
301 the dependent variable does not overcome the cost in transparency and robustness.
302 Adding an extremely weak predictor can even reduce power due to the loss of a
303 degree of freedom.

304 There are other legitimate reasons to include a measure as a covariate. One
305 is a suspected interaction. At times, demographics are included as hypothesized
306 moderators of the treatment effect. For example, if one has reason to believe that
307 the treatment will be greater among the poorly educated, then the moderator and
308 its interaction with treatment are included in the analysis model.

309 The fourth term, “control variable,” is borrowed from the observational data
310 analysis paradigm. Control variables are variables that are included in statistical

311 models in order to eliminate potentially spurious relationships between the
312 independent and dependent variables that might otherwise be thought to be
313 causal. Given that potentially spurious relationships between the independent
314 and dependent variables are eliminated by randomization in experimental
315 designs, we find the frequent use of this term out of place in experimental
316 research.

317 Observational studies often use demographics as standard control variables.
318 However, there is a tendency for political scientists to use the terms demographics
319 and control variables interchangeably, regardless of whether demographics are a
320 likely source of confounding or spurious association. The term “control variable”
321 is rampant in published experiments in political science (for just a few examples, see
322 Hutchings et al. 2004; Ladd 2010; Michelbach et al. 2003: 29; Valentino et al. 2002),
323 even when there is no evidence of differential attrition or any need to “control for”
324 other variables.

325 Notably, covariates serve a very different purpose from control variables and
326 should be selected based on different criteria. Covariates are included in the
327 statistical model for an experiment because of their anticipated relationship
328 with the *dependent variable*, to increase model efficiency. Control variables are
329 included in observational analyses because of their anticipated relationship with the
330 *independent variables*, to prevent spurious relationships between the independent
331 and dependent variables. Choosing covariates solely due to their correlation with the
332 independent variable is problematic in experiments, as we discuss at greater length
333 below.

334 Because these four terms are used more or less interchangeably in the Standards
335 Report, the recommendation is unclear as to which and how many variables the
336 committee would like to see broken down by condition in a table. Is it the ones
337 included in the original (pre-specified) model? This makes little sense because
338 those variables are already included in the model. At other times it appears they
339 are concerned specifically with those variables not included in the model, such as
340 demographics or other available pre-treatment measures which the experimenter
341 had no reason to include. But if these variables are not central to the outcome of
342 interest, it is unclear why balance on those variables is important.

343 As discussed further below, and as illustrated by many examples from political
344 science journals, the recommendation in favor of displaying all pretreatment means
345 and standard errors by experimental condition is more likely to promote confusion
346 than clarity. Indeed the number of pretreatment variables used in balance tests has
347 reached numbers as high as fifteen or more, and many more pre-treatment measures
348 are often available including variables such as household internet access, party
349 identification, age, education, race, gender, response option order, household size,
350 household income, marital status, urbanicity, home ownership, employment status,
351 if the respondent was head of the household, children in household and region,
352 to cite one example (see, e.g., Malhotra and Popp 2012). As in this example, it is

353 unclear why a particular set of pretreatment means is selected for comparison and
354 how and why one should compare them.

355 **The Role of Randomization**

356 Fisher (1935) introduced randomization as a way to make treatment and control
357 groups *stochastically* equal, meaning they are equal on average. If a researcher wants
358 to make experimental groups as equal as possible on a *specific set* of dimensions,
359 he or she would not use simple random assignment. Random assignment produces
360 random deviations of relative size inversely proportional to the square root of
361 the sample size, whereas a matched block design produces almost no deviation at
362 all. In other words, randomization is not meant as a mindless way to implement
363 blocking on known variables. The benefit of randomization is that it distributes
364 all *unknown* quantities, as well as the known quantities, in a (stochastically) equal
365 manner. This is where random assignment derives its title as the “gold standard” for
366 causal inference: because unknown factors as well as known ones are stochastically
367 equalized, possible confounding is ruled out by design.

368 The flip side of the bargain is that confounding is ruled out only stochastically.
369 The precise inference that can be drawn is that the observed data must be caused
370 by the treatment unless an event occurred which has probability less than p , where
371 p is usually equal to 0.05. Historically, this is where the confusion begins to seep in:
372 what is the nature of the “exceptional” event of probability less than 0.05, where a
373 possible Type I error occurs?

374 The strong (but mistaken) intuition of many researchers is that one should be
375 able to examine the data and see whether the randomization was unlucky. If it
376 were possible to do this, analyses of experimental data would look very different:
377 the rooting out of unlucky draws would be built into the analysis, in a manner
378 specified in advance, and accompanied by precise confidence statements. There are,
379 in fact, rejection sampling schemes that accomplish this. The downside of rejection
380 sampling schemes is that one cannot treat the randomization that one chooses to
381 keep as if it were the first and only one; instead, complex statistical adjustments
382 must be made (Morgan and Rubin 2012). Notably, what is accomplished by doing
383 so is a reduction in variance and a consequent increase in the statistical power of
384 the experiment, not a reduction in the probability of a Type I error. The important
385 point here is that balance is not necessary for valid inference in experiments. As Senn
386 (2013: 1442) explains, “It is not necessary for groups to be balanced. In fact, the
387 probability calculation applied to a clinical trial automatically *makes an allowance*
388 *for the fact that groups will almost certainly be unbalanced*, and if one knew that they
389 were balanced, then the calculation that is usually performed would not be correct”
390 (emphasis in original).

391 With experimental data, judicious choice of covariates can greatly increase the
392 power of the analysis, but this is a separate issue from confidence in the result. If
393 one wants more confidence, he or she should use a smaller p -value. If a researcher

394 uses a p -value of 0.05, then he or she will have to put up with a one in twenty chance
395 that the result is mistaken. No amount of balance testing or digging into the data
396 will eliminate or lower this uncertainty.

397 **The Role of Covariates**

398 Once a covariate is included in an analysis, the estimate of the treatment effect
399 will be adjusted for this variable. Thus, there are as many potential estimates of
400 treatment effects as there are sets of covariates that could be selected from among
401 all available pre-treatment measures. Normatively, the model (including the precise
402 set of covariates) is selected on the basis of theoretical considerations before the
403 analysis is run, in which case there is one actual estimate of treatment effect. If
404 the model is not pre-specified, the confidence statement surrounding the estimate is
405 invalidated. For this reason, the second point under Section E in the Report—which
406 asks researchers to be explicit about pre-specification of the model—is essential.

407 The most important observation to make about the many potential estimates
408 of treatment effects is that the probability of an error is equally likely with any
409 of the potential estimates. This is not to say that it does not matter which model
410 is chosen. A better choice will reduce variance, increase efficiency, and lead to
411 smaller confidence intervals. But it will not reduce the chance of Type I error.
412 Likewise, the inclusion of other variables in the model will not increase robustness.
413 Instead, the inclusion of covariates requires meeting additional assumptions that
414 are not otherwise required. In particular, the relationship between the dependent
415 variable and each of the covariates must be linear, the regression coefficient for each
416 covariate should be the same within each treatment condition, and the treatments
417 cannot affect the covariates, which is why they must be assessed pretreatment.

418 Including a large number of covariates in an analysis simply because they are
419 demographics, or because they are available in the pretest is clearly inadvisable.
420 With experimental data, “Rudimentary data analysis replaces scores of regressions,
421 freeing the researcher from the scientific and moral hazards of data mining” (Green
422 and Gerber 2002: 810–11). But the problem goes beyond the risks of data mining.
423 Many experimental studies suggest that findings are more “robust” if they survive
424 models that include additional covariates (e.g., Harbridge et al. 2014: 333; Sances
425 2012: 9). In reality, adding covariates simply because they are available reduces the
426 robustness of the model (introducing an assumption of independent linear effects
427 that do not interact with treatments), reduces transparency, and is unlikely to add
428 any power.

429 **What Purpose can Randomization Checks Serve?**

430 It is crucial to any experiment that its random assignment be correctly accomplished.
431 How might one detect errors in this regard? The first line of defense is a sufficiently
432 detailed description of the randomization mechanism. Was it the RAND() function
433 in Excel, a physical device such as a die, spinner, jar of balls, deck of cards, was

434 it a printed random number table, or some other device? Was it pre-generated or
435 generated as needed? If it was a blocked design, how was the blocking implemented?
436 The randomization process is mentioned in Section C, and while we endorse this
437 recommendation, it does not go far enough. The brief text in this recommendation
438 and its sub-paragraph on hierarchical sampling do not cover enough bases to
439 effectively prevent randomization errors. Because randomization is a *process* rather
440 than an outcome, we think a more thorough description of the process is in order as
441 recommended in Section 8a of the CONSORT (2010) checklist (Moher et al. 2010;
442 Schulz et al. 2010).

443 The Report somewhat mischaracterizes our argument in saying we agree “that
444 formal tests or their rough ocular equivalents may be useful to detect errors in the
445 implementation of randomization.” The important points are (1) that such tests are
446 not *necessary* in order to detect randomization problems; and (2) that they are not,
447 in and of themselves, sufficient evidence of a randomization problem. Due to the
448 rarity of randomization failure, we believe that the impulse to check for balance is
449 probably spurred by something other than skepticism over the functionality of the
450 random assignment process.

451 The terms “balance test” and “randomization check” are typically used
452 interchangeably to indicate a table of the distribution of pre-treatment measures
453 across treatment groups, often along with a statistical statement concerning the
454 likelihood of the extremity of the distribution having been produced by chance
455 alone. Such a statistic can be reported for each variable or a joint test can be
456 reported as a single omnibus statistic for the joint distribution of all test variables.
457 If one tests for differences in each variable individually, a large number of such tests
458 obviously increases the chance of finding significance. If one uses a joint test, it will
459 take into account the number of variables, but it will still matter a great deal which
460 particular variables are chosen for inclusion in the omnibus test. A standard example
461 of including such a check reads, “Randomization check shows that demographics
462 and political predispositions do not jointly predict treatment assignment ($X^2_{[24]} =$
463 18.48, $p = 0.779$)” (Arceneaux 2012: 275).

464 The report does not provide guidance as to what these balance variables should be,
465 except to refer to them as “pretreatment” or “demographic” variables. In examples
466 such as the one above, the exact variables are not mentioned. Given the many
467 different outcome variables that are examined in political science experiments, it is
468 unclear why demographics, in particular, are deemed particularly important when
469 other variables may be more pertinent to the outcome under study.

470 To reiterate, the Standards Committee calls for tables of unspecified pre-treatment
471 measures across treatment groups “to help detect errors” in randomization. Most
472 importantly, it is not clear how such tables accomplish this task. The distribution of
473 pre-treatment measures across conditions provides evidence of errors only if a faulty
474 randomization device was used; in other words, we are testing the null hypothesis,
475 which is the assumption that the randomization mechanism worked. If we reject

476 the null hypothesis and conclude that the randomization device was faulty, then
477 the study can no longer be considered an experiment nor be published as one. In
478 practice, however, when imbalance is identified, this is seldom the course of action
479 that is taken as we describe further below.

480 **Other Uses of Balance Testing**

481 The Standards Report (Gerber et al. 2014: 92) suggests that there are additional
482 reasons to require balance tests.

483 Detectable imbalances can be produced in several ways (other than chance). They include,
484 but are not limited to, mistakes in the randomization coding, failure to account for blocking
485 or other nuances in the experimental design, mismatch between the level of assignment
486 and the level of statistical analysis (e.g., subjects randomized as clusters but analyzed as
487 individual units), or sample attrition.

488 It is worth considering these additional rationales individually. Mistakes in coding
489 variables do indeed occur with regularity, but why should they be more likely to
490 occur with randomization variables than with the coding of other variables? Failure
491 to account for blocking is already addressed elsewhere in the requirements where
492 authors are required to describe whether and how their sample was blocked, as well
493 as how they accomplished the random assignment process. Likewise, the description
494 already must include mention of the unit of analysis that was randomized, so if the
495 authors then analyze the data at a different unit of analysis, this will be evident.

496 The one scenario in which balance testing does make sense is when
497 experimental studies take place over time, thus raising the possibility of differential
498 sample attrition due to treatment. Sample attrition does not indicate a broken
499 randomization mechanism, and it is already covered in the CONSORT diagram.
500 Assuming a control condition is present, it sets an expectation for acceptable
501 attrition levels. And if there is differential attrition across experimental conditions,
502 then it makes perfect sense to conduct balance tests on pretreatment variables
503 among post-test participants. If the post-attrition distribution of pre-treatment
504 measures across treatment groups is distinguishable from the random pre-treatment
505 distribution, then the experiment is clearly confounded.

506 For various reasons, we believe that error detection and differential attrition are
507 not the primary reasons that balance testing is popular. Instead, as described above,
508 we believe part of its appeal stems from researchers' strong intuition that they
509 can unearth the unlucky draw. Further, the Report of the Standards Committee
510 explicitly says that error detection is not the only reason for doing randomization
511 checks. As stated on page 5 of the standards document,

512 There may be other uses of summary statistics for covariates for each of the experimental
513 groups. For instance, if there is imbalance, whether statistically significant or not, in a
514 pretreatment variable that is thought by a reader to be highly predictive of the outcome,
515 and this variable is not satisfactorily controlled for, the reader may want to use the baseline

516 sample statistics to informally adjust the reported treatment effect estimates to account for
517 this difference.

518 There are several problems with this statement, which we address in order of
519 appearance. First, the phrase “statistically significant or not” is meaningless in
520 the context of adjustment. The only thing one can test statistically is the null
521 hypothesis, which is the assumption that the randomization mechanism worked. If
522 one is estimating treatment effects, then one is already assuming that the mechanism
523 worked, so there is no question of significance. This point has been made many times
524 in the literature in political science (Imai et al. 2008) as well as in other disciplines
525 (Boers 2011; Senn 1994).

526 Further, it is important to think through the consequences of this requirement
527 for reviewers as well as authors. This statement implies that it is acceptable and
528 even appropriate for a reviewer to (either subjectively or based on a prescribed test)
529 perceive an imbalance in the table, assert that it is a variable that might be related to
530 the dependent variable, and therefore insist that something be done to address the
531 situation.

532 Regardless of whether there is a statistical test, what happens next? Is it incumbent
533 upon the author to somehow “prove” that randomization was done appropriately?
534 How can this possibly be accomplished? And if we conclude from an author’s
535 inability to produce such evidence that randomization was not done correctly, then
536 what? If balance tests/tables are truly being used to ascertain whether random
537 assignment was done correctly, then the only logical response to concluding that it
538 was not done correctly is to throw out the study altogether, or possibly analyze it as
539 purely observational data.

540 Random assignment was either done correctly or it was not; there is no middle
541 ground. This does not appear to be widely understood. As an experimental study in
542 the *American Journal of Political Science* explained, “To test the robustness of our
543 randomization scheme, we tested for any differences among the other observables
544 on which we did not block.. ..” (Butler and Broockman, 2011: 467). *Results* can
545 certainly be more or less robust, but random assignment is either done correctly or
546 it is not; there are no varying degrees of randomization.

547 **Fixing a Broken Mechanism?**

548 The assumption that one can “fix” a broken random assignment by the virtue
549 of adding a covariate is commonplace throughout our top journals. For example,
550 in a *Public Opinion Quarterly* article we are told that, “Partisanship is included
551 in the analysis because of imbalances in the distribution of this variable across
552 the conditions.” (Hutchings et al. 2004: 521). Likewise, an article in the *American*
553 *Journal of Political Science* assures us that “Every relevant variable is randomly
554 distributed across conditions with the exception of education in Study 1. When we

555 included education in our basic models, the results were substantially the same as
556 those we report in the text” (Berinsky and Mendelberg 2005: 862).

557 There is no logic to including a “control” variable to correct for lack of true
558 random assignment on just one or a few characteristics, a point that does not seem
559 to be widely understood by political scientists. For example, Barabas and colleagues
560 (2011: 21) assert that “we observed non-random treatment assignment (i.e., $p <$
561 0.10 differences between the treatment and control groups on partisanship, age,
562 education, and race) which necessitates the use of statistical controls later in the
563 paper.” Of course, by “non-random,” the authors probably did not mean that their
564 randomization mechanism was faulty; therefore, they continue to treat the study as
565 an experiment, not as an observational study resulting from a failed randomization
566 mechanism.

567 Adding a variable to the statistical model for an experimental analysis because
568 it failed a randomization check is an inferior model choice (see Imai et al. 2008;
569 Mutz and Pemantle 2011). It is a misnomer to say that it “controls” for the lack of
570 balance and there is no defensible reason to accept this as a “fix” for a broken random
571 assignment mechanism, if that is indeed what we are looking for by providing such
572 tables.

573 We suspect that instead of a failure to randomize, what many authors and
574 reviewers actually have in mind is the unlucky chance that experimental conditions
575 are unbalanced on some variable of potential interest. Of course, if it is a strong
576 predictor of the dependent variable, a pre-treatment measure of that variable should
577 have been used for blocking purposes or as a planned covariate in the model to
578 increase model efficiency regardless of balance; this is the only appropriate purpose
579 of covariates.

580 But more importantly, using a third variable to try to “correct” a model for
581 imbalance ignores the fact that the alpha value used to test experimental hypotheses
582 *already takes into account* that cells will be uneven on some characteristics due
583 to chance. The committee report states that “we . . . do not counsel any particular
584 modeling response to the table of covariate means that we ask the researcher to
585 provide.” However, given that the only example provided of what one might do with
586 this information is to adjust the treatment effects by including covariates, this seems
587 somewhat misleading. As they elaborate, “Our guiding principle is to provide the
588 reader and the reviewer the information they need to evaluate what the researcher
589 has done and to update their beliefs about the treatment effects accordingly.” But
590 exactly how should the reviewer or reader “update” his or her beliefs about the
591 effects of treatment based on such a table?

592 If such a table truly serves as evidence (or lack thereof) that proper random
593 assignment was accomplished, then such tables will greatly affect a study’s chances
594 of publication. By requiring such information, an editor automatically suggests to
595 readers and authors that it is both informative and relevant because it is worth
596 valuable journal space. If it is to be required, it seems incumbent upon the editors to

597 inform authors as to how they will interpret such information. Will they conclude
598 that random assignment was done incorrectly on this basis and thus automatically
599 reject it from an experimental journal? Will they compel authors to provide evidence
600 that random assignment was done correctly, and if so, what would be considered
601 compelling evidence?

602 And are they required to present evidence of balance even on demographic
603 variables that bear no relation to the outcome variables simply because they
604 are widely used as control variables in observational analyses or on all pre-
605 treatment measures because they happen to be in the study? We maintain that such
606 practices have no scientific or statistical basis and serve only to promote further
607 methodological confusion.

608 The report does not distinguish between pre-treatment measures available to the
609 researcher but not chosen for inclusion in the model, and those chosen in advance
610 for inclusion in the model. If, as is common with survey data, there are dozens of
611 available pre-treatment measures, then is balance supposed to be reported for all
612 of them? If so, why? As Thye (2007: 70) has noted, “Not all the factors that make
613 experimental groups different from control groups are relevant to the dependent
614 variable; therefore, not all factors must necessarily be equated. Many differences
615 simply do not matter.” To advocate such a practice is to encourage mindless
616 statistical models, which should not be promoted by any journal. It encourages
617 a misunderstanding of what randomization does and does not accomplish. It also
618 promotes further confusion in the field as to the distinction between experimental
619 and observational analysis.

620 To reiterate, pre-treatment variables known from previous research to be highly
621 predictive of the outcome should always be included in the model as covariates. To
622 fail to do so is to reduce power so that only the strongest effects will be seen. It
623 should not take a failed balance test to reveal such a variable, and the fact that a
624 balance test fails for a particular variable makes it no more likely that this variable
625 is in fact related to the dependent variable.

626 Finally, the question of whether a variable is adequately “controlled for” is a
627 non sequitur in experimental research. Control variables exist for good reasons in
628 observational studies (potential spuriousness), but treating a covariate as a control
629 variable in the experimental setting makes no sense. Nonetheless, this practice is
630 currently widespread. Including a variable in the statistical model because it has
631 been found to be out of balance is also precisely the wrong reason to include a
632 variable and should not increase our confidence in findings.

633 Taken at face value, the Standards Report promotes randomization checks strictly
634 as a way of “evaluating the integrity of the randomization process” (Gerber et al.,
635 2014: 92). They suggest that imbalances due to chance are distinguishable from
636 imbalances due to faulty random assignment mechanisms. But if a fear of faulty
637 mechanisms is the real reason for doing them, then the typical response (adding new

638 variables to the model) is completely inadequate; if a randomization mechanism
639 fails, the researcher needs to start over from scratch.

640 To summarize, failed balance tests cast doubt on experimental results; as a
641 result, one seldom if ever finds a published experimental study with a “failed”
642 randomization; instead they are routinely dubbed “successful” (Malhotra and
643 Popp 2012: 39) and even “highly successful” (Butler and Broockman 2011: 467).
644 Moreover, if an author admits a “failed” balance test, it is strictly on one or two
645 “unbalanced” variables that are, as a result of the balance test, included as covariates.
646 This practice does not fix the problem if the randomization mechanism was, indeed,
647 broken.

648 The real harm in this practice is the possibility of a Type II error when a skeptical
649 referee or editor causes a correct finding to be suppressed or uses it as a reason to
650 alter the statistical model to include more covariates in order to suggest that they
651 have “adjusted” for a bad or unlucky randomization. This practice implies that the
652 reader’s ad hoc estimates of treatment effects and confidence might be superior to
653 the researcher’s stated estimates and confidence. As mentioned above, changing the
654 model voids confidence statements.

655 At times, this kind of misunderstanding of randomization is made explicit, even
656 within our top journals. For example, as an article in the *Journal of Politics* explains,

657 In order to ensure that the experimental conditions were randomly distributed—thus
658 establishing the internal validity of our experiment—we performed difference of means tests
659 on the demographic composition of the subjects assigned to each of the three experimental
660 conditions. . . . As Tables 1a and 1b confirm, there were no statistically significant differences
661 between conditions on any of the demographic variables. . . . Having established the random
662 assignment of experimental conditions, regression analysis of our data is not required; we
663 need only perform an analysis of variance (ANOVA) to test our hypotheses as the control
664 variables that would be employed in a regression were randomly distributed between the
665 three experimental conditions (Scherer and Curry 2010: 95).

666 A test of mean differences across five demographic variables is not what gave this
667 study internal validity; proper use of random assignment did. Moreover, controlling
668 for these variables in a regression equation or using them as covariates would not
669 have fixed a failed randomization, nor would it have increased the power of the
670 study, unless those variables were chosen in advance for the known strength of their
671 relationships with the dependent variable rather than for their relationships with
672 the independent variable, as is suggested above.

673 Many researchers do not appear to understand that the alpha value used in
674 statistical tests *already incorporates the probability of the unlucky draw*. As Hyde
675 (2010: 517) suggests in another experimental study,

676 In theory, the randomization should produce two groups that are equivalent except that one
677 group was assigned to be “treated” with international election observation. Although it is
678 unlikely, it is possible that randomization produces groups of villages/neighborhoods that

679 are different in important ways, and could potentially generate misleading results. Therefore,
680 I also check the degree to which the two groups are similar . . .

681 Here again, a randomization check is being used to try to uncover the unlucky
682 draw in order to increase confidence in the findings as opposed to presenting
683 “misleading results.” This is a well-intentioned impulse, but one should not update
684 his or her confidence in the findings on this basis.

685 Using balance tests on a subset of observed variables as a way of establishing
686 group equivalence promotes further confusion because of the current popularity of
687 matching techniques in analyzing observational data. If, for example, a researcher
688 matches treated and untreated subjects on five demographic characteristics, there
689 is a tendency to see this as equivalent to an experiment in which a balance test has
690 been performed on these same five variables. What is lost here is an understanding
691 of the fundamental importance of random assignment. Matching techniques, no
692 matter how complex, cannot accomplish the same strength of causal inference as a
693 true experiment. Only random assignment collectively equates subjects on observed
694 and unobserved characteristics.

695 The lure of the unlucky draw, however, goes far beyond this. There is a strong urge
696 to believe that one can test for occurrences of the exceptional event: that not only
697 does Type I error have a visible signature but also that we can sense it, and therefore
698 should look at balance tests even though we are not able to prescribe an acceptable
699 response to what we see. This may be what is responsible for the heightened concern
700 about errors in randomization.

701 **Sub-Optimal Statistical Models**

702 Randomization checks notwithstanding, a more serious and widespread problem
703 in political science experiments is confusion surrounding analyzing experimental
704 versus observational data. By the Standards Committee’s own count, 75% of the
705 experimental studies published in political science do not show the unadulterated
706 effects of treatments on outcomes (Gerber et al. 2014: 88). In other words, 75%
707 of experimental results never show the reader the dependent variable means by
708 experimental condition or a regression including only treatment effects; instead,
709 they present multiple regression equations in which effects of treatment are already
710 adjusted by many other “control” variables, or they present *predicted means* as
711 a function of a multivariate regression equations including other variables (e.g.,
712 Hutchings et al. 2004: 521–2).

713 For example, in one analysis of experimental results, in addition to dummy
714 variables representing six different experimental treatments, the author includes in
715 his experimental regression analysis nine different “control variables” including if
716 the respondent follows politics “most of the time,” if he/she is a college graduate, age,
717 female, minority, employed, internet connection speed, conservative ideology and
718 liberal ideology. The rationale for this particular set of variables when predicting the

719 dependent variable—nightly news exposure—is unclear (see Prior 2009). Likewise,
720 in another experiment, in addition to dummy variables for experimental treatment
721 effects, the author includes 13 additional predictors of the outcome, none of which
722 significantly predicts the dependent variable, and the reader is instructed that these
723 variables are included as “control variables” (Ladd 2010: 39).

724 What is particularly unfortunate about this practice is that reviewers and authors
725 often seem to be under the impression that an experimental finding is more robust
726 if it survives the inclusion of a large number of “control variables” when nothing
727 could be further from the truth. Instead of encouraging this practice, reviewers and
728 editors should look at such large models with suspicion and demand justifications
729 for the particular statistical model that is used. Findings can be coaxed over the line
730 of statistical significance by virtue of what is included or excluded.

731 We are not suggesting that social scientists are dishonest when such variables are
732 included in a model. In fact, many authors find themselves compelled to include
733 them in an analysis specifically because of a reviewer or editor’s request. Even when
734 they are aware that their finding is more valid without excessive and unjustified
735 variables in the model, they comply in order to achieve publication. Adding variables
736 after the fact invalidates the reporting of confidence levels. Moreover, the proper
737 reporting of confidence is not an idle exercise; in fact, some suggest that it has large
738 scale consequences (Ioannidis 2005).

739 In some cases, these additional variables in models testing experimental effects
740 even include items assessed after the treatment. For example, in an article in the
741 *American Journal of Political Science*, a study of income distribution norms and
742 distributive justice promotes the inclusion of a variable assessed post-treatment as
743 a means of strengthening confidence in the experimental findings.

744 By asking participants in the post-experimental questionnaire about their own perception
745 of the relationship between merit and income, and then entering that information as
746 an independent variable in our regression analyses, we are able to determine that our
747 experimental manipulations rather than participants’ pre-existing perceptions explain our
748 results. This test shows how using multiple regression analysis to enter additional controls
749 can strengthen experimental research” (Michelbach et al. 2003: 535).

750 In short, what is most common within political science is for researchers to
751 analyze experimental data as if it were observational data, often including “control
752 variables” inappropriately. If there is no reason to think variables will increase
753 efficiency in the estimation of treatment effects, and no reason to think that they
754 are even correlated with the outcome, they should not be in the model, regardless of
755 what they are called. Which other variables are or are not included is unsystematic
756 and typically unjustified with some models including one set, and another model
757 within the same paper including a different set, thus opening the floodgates for all
758 kinds of foraging for results through their inclusion and exclusion.

759 We are not the first to note the logical problems inherent in randomization
760 checks. Psychologist Robert Abelson (1995: 76) dubbed the practice of testing for

761 differences between experimental groups a “silly significance test”: “Because the null
762 hypothesis here is that the samples were randomly drawn from the same population,
763 it is true by definition, and needs no data.” Senn (1994: 1716) calls the practice of
764 performing randomization tests “philosophically unsound, of no practical value,
765 and potentially misleading.” In the context of political science, Imai and colleagues
766 (2008: 482) echo this sentiment, suggesting that any other purpose [than to test the
767 randomization mechanism] for conducting such a test is “fallacious.”

768 The field of political science is populated with researchers primarily trained in
769 observational modes of research. For those trained exclusively in observational
770 methods, the source of confusion is obvious. If one treats experimental data as if
771 it were observational, then of course one would be worried about “controlling for”
772 variables, and about imbalance in any variable not used as a covariate. We believe
773 the Standards Committee should take a stand on whether they believe “control”
774 variables are sensible in experimental studies and/or whether they are an acceptable
775 fix for the broken random assignment mechanisms that balance tests are supposedly
776 designed to root out.

777 So, how can researchers be certain that randomization was done properly? The
778 CONSORT guidelines already provide guidance as to the kinds of details that
779 can help reviewers and readers judge the randomization process (see Moher et al.
780 2010; Schulz et al. 2010). Notably, because randomization is a *process* rather than
781 an outcome, what is more useful than tables of means is a description of that
782 process. Political scientists should test randomization mechanisms in advance of
783 studies if there are concerns, and promote transparency by describing the process
784 of randomization for each study.

785 When debating the utility of randomization checks, one argument we have heard
786 a number of times is “Why not just do both and let the reader decide?” In other
787 words, why not present both the original analysis and one adjusted by including the
788 covariates that fail a specified balance test? Assuming the randomization mechanism
789 is not faulty, there remain several good reasons not to do this. We elaborate on four
790 such reasons.

791 1. *Incorrectness*. Significance statements and size comparisons for the estimated
792 treatment effect will be wrong. To see why, consider the process by which the
793 adjusted estimate is computed. After the random assignment to experimental
794 conditions, a set of covariates exhibiting imbalance is added to the model. An
795 estimated treatment effect is computed by regressing onto the treatment variable
796 and this larger set of covariates. Confidence intervals and p -values for such an
797 estimator do not coincide with confidence intervals and p -values for a model in
798 which the same covariates are chosen before the units are assigned to conditions
799 (see Permutt 1990).

800 2. *Intractability*. Computing correct confidence statements for a model in which
801 covariate selection is not fixed in advance has, to our knowledge, never been
802 undertaken. An idealized example is worked out in Permutt (1990). Whether or

- 803 not such a computation is feasible, it is certainly not included in any standard
 804 statistics package. We can therefore be fairly certain that the correct computation
 805 was not carried out.
- 806 3. *Inefficiency*. Even if confidence was computed correctly for the adjusted estimate,
 807 the new estimator would not be an improvement over the old one. Any selection
 808 of covariates, whether chosen in advance or based on imbalance due to random
 809 assignment, leads to an estimator. The better estimator is the one with the least
 810 variance. For any pre-treatment measure, Z , one might choose to include Z in
 811 the model, exclude it, or include it only if it is unbalanced. The last of these is
 812 never the best choice. One always does better by deciding up front whether to
 813 include Z as a covariate. The mathematical proof supporting this is discussed in
 814 greater detail in Mutz and Pemantle (2011).
- 815 4. *Irrelevance*. One might argue that presenting both estimators and allowing the
 816 reader to choose is best because it reports everything that would originally have
 817 been reported, plus one more piece of data which the reader is free to ignore.
 818 Reporting a second conclusion, however, casts doubt on the first conclusion; it
 819 does not merely add information. It leads to “the wrong impression that we need
 820 balance, which is one of the many myths of randomization” (Statisticalmisses.nl
 821 2013). Recommendation E2 of the Standards for Experimental Research calls
 822 for an analysis to be specified prior to the experiment, and deviations from this
 823 come at a cost in credibility. Furthermore, if given a choice between two models,
 824 many would automatically choose the model with more covariates based on a
 825 (faulty) belief that such models are more robust. The researcher’s job is to present
 826 the best data analysis, not to present them all and allow the reader to choose.

827 CONCLUSION

828 The goal of the *Journal of Experimental Political Science* should be not only
 829 promoting the more widespread use of experimental methods within the discipline,
 830 but also promoting best practices and a better understanding of experimental
 831 methods across the discipline. Toward that end, we hope the Standards Committee
 832 will consider changing the standards with respect to manipulation checks, reporting
 833 of response rates, and randomization checks as part of the ongoing process
 834 of making what was historically an observational discipline more appropriately
 835 diverse in its approaches to knowledge. More specifically, we suggest the following
 836 adjustments:

- 837 1. Recommend manipulation checks for latent independent variables; that is,
 838 independent variables in which the operationalization and the causal construct
 839 are not identical;
- 840 2. Require response rates only for studies that claim to be random probability
 841 samples representing some larger population;

- 842 3. If tables of pretreatments means and standard errors are to be required, provide
 843 a justification for them. (Note that the following are not suitable justifications:
 844 (a) Confirmation of “successful” randomization, (b) Supporting the validity of
 845 causal inference, and (c) Evidence of the robustness of inference.)
- 846 4. If the inclusion of balance tests/randomization checks is described as desirable as
 847 in the current document, prescribe the appropriate response and interpretation
 848 of “failed” tests.

849 Evidence of widespread misunderstandings of experimental methods is plentiful
 850 throughout our major journals, even among top scholars in the discipline. As
 851 a result, future generations of political scientists are often not exposed to best
 852 practices. *The Journal of Experimental Political Science* should play a lead role in
 853 correcting these misunderstandings. Otherwise, the discipline as a whole will be seen
 854 as less methodologically sophisticated than is desirable. *The Journal of Experimental*
 855 *Political Science* could play an important role in raising the bar within the discipline
 856 by including requirements that are both internally coherent and statistical defensible.

857 REFERENCES

- 858 Abelson, R. 1995. *Statistics as Principled Argument*. Hillsdale, NJ: L. Erlbaum Associates.
- 859 Arceneaux, K. 2012. “Cognitive Biases and the Strength of Political Arguments.” *American*
 860 *Journal of Political Science* 56(2): 271–85.
- 861 Arceneaux, K. and R. Kolodny. 2009. “Educating the Least Informed: Group Endorsements
 862 in a Grassroots Campaign.” *American Journal of Political Science* 53(4): 755–70.
- 863 Barabas, J., W. Pollock and J. Wachtel. 2011. “Informed Consent: Roll-Call Knowledge,
 864 the Mass Media, and Political Representation.” Paper Presented at the Annual Meeting
 865 of the American Political Science Association, Seattle, WA, Sept. 1–4. ([http://www.
 866 jasonbarabas.com/images/BarabasPollockWachtel_RewardingRepresentation.pdf](http://www.jasonbarabas.com/images/BarabasPollockWachtel_RewardingRepresentation.pdf)), ac-
 867 cessed September 1, 2014.
- 868 Berinsky, A. J., M. F. Margolis and M. W. Sances. 2014. “Separating the Shirkers from
 869 the Wokers? Making Sure Respondents Pay Attention on Self-Administered Surveys.”
 870 *American Journal of Political Science* 58(3): 739–53.
- 871 Berinsky, A. J. and T. Mendelberg. 2005. “The Indirect Effects of Discredited Stereotypes in
 872 Judgments of Jewish Leaders.” *American Journal of Political Science* 49(4): 845–64.
- 873 Boers, M. 2011. “In randomized Trials, Statistical Tests are not Helpful to Study Prognostic
 874 (im)balance at Baseline.” *Lett Ed Rheumatol* 1(1): e110002. doi:10.2399/ler.11.0002.
- 875 Butler, D. M. and D. E. Broockman. 2011. “Do Politicians Racially Discriminate Against
 876 Constituents? A Field Experiment on State Legislators.” *American Journal of Political*
 877 *Science* 55(3): 463–77.
- 878 Cover, A. D. and B. S. Brumberg. 1982. “Baby Books and Ballots: The Impact of
 879 Congressional Mail on Constituent Opinion.” *The American Political Science Review*
 880 76(2): 347–59.
- 881 Cozby, P. C. 2009. *Methods of Behavioral Research*. (10th ed.). New York, NY: McGraw-Hill.
- 882 Fisher, R. A. 1935. *The Design of Experiments*. London: Oliver and Boyd.
- 883 Foschi, M. 2007. “Hypotheses, Operationalizations, and Manipulation Checks.” Chapter 5,
 884 In *Laboratory Experiment in the Social Sciences*, eds. M. Webster and J. Sell, (pp.113–140).
 885 New York: Elsevier.

- 886 Franco, A., N. Malhotra and G. Simonovits. 2014. "Publication Bias in the Social Sciences:
887 Unlocking the File Drawer." *Science* 345(6203): 1502–5.
- 888 Franklin, C. 1991. "Efficient Estimation in Experiments." *Political Methodologist* 4(1):
889 13–15.
- 890 Gerber, A., K. Arceneaux, C. Boudreau, C. Dowling, S. Hillygus, T. Palfrey, D. R. Biggers
891 and D. J. Hendry. 2014. "Reporting Guidelines for Experimental Research: A Report
892 from the Experimental Research Section Standards Committee." *Journal of Experimental*
893 *Political Science* 1(1): 81–98.
- 894 Gosnell, H. F. 1942. *Grass Roots Politics*. Washington, DC: American Council on Public
895 Affairs.
- 896 Green, D. P. and A. S. Gerber. 2002. "Reclaiming the Experimental Tradition in Political
897 Science." In *Political Science: The State of the Discipline*, 3rd Edition. eds. H. V. Milner
898 and I. Katznelson, (pp.805–32). New York: W.W. Norton & Co.
- 899 Harbridge, L., N. Malhotra and B. F. Harrison. 2014. "Public Preferences for Bipartisanship
900 in the Policymaking Process." *Legislative Studies Quarterly* 39(3): 327–55.
- 901 Hiscox, M. J. 2006. "Through a Glass and Darkly: Attitudes Toward International Trade
902 and the Curious Effects of Issue Framing." *International Organization* 60(3): 755–80.
- 903 Hutchings, V. L., N. A. Valentino, T. S. Philpot and I. K. White. 2004. "The Compassion
904 Strategy: Race and the Gender Gap in Campaign 2000." *Public Opinion Quarterly* 68(4):
905 512–41.
- 906 Hyde, S. D. 2010. "Experimenting in Democracy Promotion: International Observers and
907 the 2004 Presidential Elections in Indonesia." *Perspectives on Politics* 8(2): 511–27.
- 908 Imai, K., G. King and E. Stuart. 2008. "Misunderstandings between Experimentalists and
909 Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series*
910 *A*, 171(2): 481–502.
- 911 Ioannidis, J. P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine*
912 2(8): e124. doi:10.1371/journal.pmed.0020124.
- 913 Kessler, J. B. and S. Meier. 2014. "Learning from (Failed) Replications: Cognitive Load
914 Manipulations and Charitable Giving." *Journal of Economic Behavior and Organization*
915 102(June): 10–13.
- 916 Ladd, J. M. 2010. "The Neglected Power of Elite Opinion Leadership to Produce Antipathy
917 Toward the News Media: Evidence from a Survey Experiment." *Political Behavior* 32(1):
918 29–50.
- 919 Malhotra, N. and E. Popp. 2012. "Bridging Partisan Divisions over Antiterrorism Policies:
920 The Role of Threat Perceptions." *Political Research Quarterly* 65(1): 34–47.
- 921 Michelbach, P. A., J. T. Scott, R. E. Matland and B. H. Bornstein. 2003. "Doing Rawls
922 Justice: An Experimental Study of Income Distribution Norms." *American Journal of*
923 *Political Science* 47(3): 523–39.
- 924 Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux,
925 D. Elbourne, M. Egger and D. G. Altman. CONSORT. 2010. "Explanation and
926 Elaboration: Updated Guidelines for Reporting Parallel Group Randomised Trials."
927 *Journal of Clinical Epidemiology* 63(8): e1–e37.
- 928 Morgan, K. and D. Rubin. 2012. "Rerandomization to improve covariate balance in
929 experiments." *Annals of Statistics* 40(2): 1263–82.
- 930 Mutz, D. C. and R. Pemantle. 2011. "The Perils of Randomization Checks in the
931 Analysis of Experiments." Paper presented at the Annual Meetings of the Society for
932 Political Methodology, (July 28–30). ([http://www.math.upenn.edu/~pemantle/papers/](http://www.math.upenn.edu/~pemantle/papers/Preprints/perils.pdf)
933 [Preprints/perils.pdf](http://www.math.upenn.edu/~pemantle/papers/Preprints/perils.pdf)), accessed September 1, 2014.

- 934 Newsted, P. R., P. Todd and R. W. Zmud. 1997. "Measurement Issues in the Study of
935 Human Factors in Management Information Systems." Chapter 16, In *Human Factors in*
936 *Management Information System*, ed. J. Carey, (pp.211–242). New York, USA: Ablex.
- 937 Perdue, B. C. and J. O. Summers. 1986. "Checking the Success of Manipulations in Marketing
938 Experiments." *Journal of Marketing Research* 23(4): 317–26.
- 939 Permutt, T. 1990. "Testing for Imbalance of Covariates in Controlled Experiments." *Statistics*
940 *in Medicine* 9(12): 1455–62.
- 941 Prior, M. 2009. "Improving Media Effects Research through Better Measurement of News
942 Exposure." *Journal of Politics* 71(3): 893–908.
- 943 Sances, M. W. 2012. "Is Money in Politics Harming Trust in Government? Evidence from
944 Two Survey Experiments." (<http://www.tessexperiments.org/data/SancesSSRN.pdf>),
945 accessed January 20, 2015.
- 946 Sansone, C., C. C. Morf and A. T. Panter. 2008. *The Sage Handbook of Methods in Social*
947 *Psychology*. Thousand Oaks, CA: Sage Publications.
- 948 Saperstein, A. and A. M. Penner. 2012. "Racial Fluidity and Inequality in the United States."
949 *American Journal of Sociology* 118(3): 676–727.
- 950 Scherer, N. and B. Curry. 2010. "Does Descriptive Race Representation Enhance Institutional
951 legitimacy? The Case of the U.S. Courts." *Journal of Politics* 72(1): 90–104.
- 952 Schulz, K. F., D. G. Altman, D. Moher, for the CONSORT Group. CONSORT 2010.
953 "Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials."
954 *British Medical Journal* 340: c332.
- 955 Senn, S. 1994. "Testing for Baseline Balance in Clinical Trials." *Statistics in Medicine* 13:
956 1715–26.
- 957 Senn, S. 2013. "Seven Myths of Randomisation in Clinical Trials." *Statistics in Medicine*
958 32(9): 1439–50. doi: 10.1002/sim.5713. Epub 2012 Dec 17.
- 959 Statisticalmisses.nl, 2013. ([http://www.statisticalmisses.nl/index.php/frequently-asked-
960 questions/84-why-are-significance-tests-of-baseline-differences-a-very-bad-idea](http://www.statisticalmisses.nl/index.php/frequently-asked-questions/84-why-are-significance-tests-of-baseline-differences-a-very-bad-idea)),
961 accessed January 21, 2015. As attributed to Senn (2013).
- 962 Thye, S. 2007. "Logic and Philosophical Foundations of Experimental Research in the
963 Social Sciences." Chapter 3, In *Laboratory Experiments in the Social Sciences*, (pp.57–
964 86). Burlington, MA: Academic Press.
- 965 Valentino, N. A., V. L. Hutchings and I. K. White. 2002. "Cues That Matter: How Political
966 Ads Prime Racial Attitudes during Campaigns." *The American Political Science Review*
967 96(1): 75–90.