Using independent covariates in experimental designs:

Quantifying the trade-off between power boost and Type I error inflation

Y. Andre Wang,[a] Jehan Sparks,[a] Joseph E. Gonzales,[b] Yanine D. Hess,[c] & Alison Ledgerwood[a]

[a]University of California, Davis [b]University of Massachusetts, Lowell

[c]State University of New York at Purchase

Manuscript in press, JESP

Please address editorial correspondence to:
Andre Wang or Alison Ledgerwood
Department of Psychology
University of California, Davis
One Shields Avenue
Davis, CA 95616.
E-mail: aledgerwood@ucdavis.edu

Abstract

The practice of using covariates in experimental designs has become controversial. Traditionally touted by statisticians as a useful method to soak up noise in a dependent variable and boost power, the practice recently has been recast in a negative light because of Type I error inflation. But in order to make informed decisions about research practices like this one, researchers need to know more about the actual size of their benefits and costs. In a series of simulations, we compared the Type I error rates and power of two analytic practices that researchers might use when confronted with an unanticipated, independent covariate. In the baseline practice, a researcher only analyzes the effect of the manipulation on the dependent variable; in the flexible-covariate practice, she analyzes both the effect of the manipulation on the dependent variable and the effect adjusting for the unanticipated covariate. We show that the flexible-covariate (vs. baseline) practice inflates Type I error by a small amount, and that it boosts power substantially under certain circumstances. The flexible-covariate practice tends to be most beneficial when the covariate is strongly correlated with the dependent variable in the population, and when the experimental design would have been only moderately powered (40%–60%) without including the covariate in the analysis. We offer concrete recommendations for when and how to use independent covariates in experimental designs, and contextualize our findings within the movement toward quantifying tradeoffs in choosing among research practices and optimizing the choice of practice within a given research context.

Keywords: covariate; false positives; power; tradeoff; exploratory analysis

Using independent covariates in experimental designs:

Quantifying the trade-off between power boost and Type I error inflation

Traditionally, statisticians have promoted the covariate as a useful tool for soaking up extra noise in a dependent variable, thereby boosting the statistical power of an experiment (e.g., Cohen, 1988; Maxwell & Delaney, 1990). More recently, however, this practice has been recast in a negative light because the flexible inclusion of a covariate in an analysis can lead to Type I error inflation (e.g., Simmons, Nelson, & Simonsohn, 2011; Simonsohn, Nelson, & Simmons, 2014). Thus, covariates—once cast as power-boosting heroes of experimental research—have been reimagined as error-inflating villains. These contrasting perspectives have led to considerable confusion as researchers attempt to distill these (often nuanced) statistical discussions into concrete and straightforward guidelines for best practices. Indeed, a quick skim of various articles and chapters on research practices can find covariates both touted for their power-boosting capabilities and frowned upon for their error-inflating potential within the same paper (e.g., Asendorpf et al., 2013; Ledgerwood, Soderberg, & Sparks, 2017).[1]

One easy way to reconcile these contrasting narratives is simply to advocate including a covariate in the analysis of experimental data if and only if the covariate is specified *a priori*—for instance, in a pre-analysis plan (see Ledgerwood et al., 2017; Simonsohn et al., 2014). By choosing a covariate ahead of time and analyzing the data only with the covariate included, a

[1] Of course, there are two different reasons why a researcher might use a covariate (see Supplemental Materials for a detailed discussion). In the first—the focus of this article—a researcher includes a covariate measured before an experimental manipulation in order to soak up some of the noise in her dependent variable. That is, the covariate is *independent* from the manipulation, and it accounts for a portion of the variance in the dependent variable that is due to stable individual differences. In the second, a researcher includes a covariate measured *after* an experimental manipulation, or measured in a nonexperimental study, in an attempt to control for a potential confound; in such cases, the purpose of the covariate is not reducing unexplained variance in the dependent variable but rather accounting for some portion of the explained variance. We focus here on the first context; a host of other issues are relevant for the second context and have been discussed in detail elsewhere (e.g., Westfall & Yarkoni, 2016; Yzerbyt, Muller, & Judd, 2004).

researcher can avoid Type I error inflation while still realizing any power-boosting benefits of including a covariate in their analysis. This *a priori* approach to including a covariate therefore represents an ideal strategy for researchers to use in order to maximize what they can learn from their data.

But of course, even the most careful planning cannot foresee every possible circumstance, and in the real and inevitably messy world of everyday research, the possibility of including a covariate in a design is sometimes unanticipated. A researcher may run an experiment, analyze her data, and only then realize that she could have included a particularly promising covariate that a colleague assessed at the beginning of the semester. Consider a scenario that mirrors one we have often seen in our own labs: A social psychologist conducts an experiment to test the effect of a social comparison manipulation on self-esteem, finds a non-significant effect, and then presents the “failed” study at a lab meeting to ask for advice on next steps. A student might then suggest: Why not include extraversion, which was measured in the departmental prescreen, as a covariate in the analysis to boost power, since extraversion correlates relatively strongly with self-esteem (Robins, Tracy, Trzesniewski, Potter, & Gosling, 2001) and is therefore likely to soak up a substantial amount of noise in the dependent variable of interest? When researchers stumble upon a promising covariate unexpectedly like this, should they ever consider using it in the hope of boosting power and learning more from their already collected data? Or should they shun unexpected covariates entirely, to avoid rampant Type I error inflation?

To answer these questions, we need to know more about (a) how beneficial is the research practice of including a promising but unplanned covariate for boosting power, as well as (b) how costly is this practice for Type I error. Like many other research practices, flexibly analyzing one’s experimental data both with and without a covariate poses a trade-off (Brewer & Crano, 2014; Finkel, Eastwick, & Reis, in press; Ledgerwood & Shrout, 2011): in this case, between boosting statistical power on the one hand and increasing the likelihood of a false positive on the other. If researchers are to make thoughtful and well-informed choices about their research practices—and to engage in empirically grounded discussions and debates about the merits and drawbacks of various approaches—we need to start quantifying these kinds of trade-offs and exploring the conditions under which they are bigger or smaller (Miller & Ulrich, 2016).

In this article, we assume that most researchers share the goal of maximizing what they can learn from their data—an aim that has long been recognized as involving a balancing act between minimizing Type I error on the one hand and maximizing statistical power on the other (Keppel & Wickens, 2004; Lakens & Evers, 2014; Ledgerwood, 2014). We set out to evaluate various approaches to using covariates that researchers might consider in terms of their usefulness in helping researchers to achieve that goal (see Table 1 for a preview of our recommendations). In addition to considering the *a priori* research practice noted above, we conducted a series of simulations to quantify the power boost and Type I error inflation produced by flexibly including a single, promising covariate when analyzing experimental data. We also explore and discuss a range of other possible approaches a researcher might take when confronting an unanticipated covariate, including flexibly including multiple covariates and flexibly testing interactions between a covariate and the key independent variable of interest. We conclude with concrete recommendations for researchers wishing to make informed decisions about tradeoffs when selecting among possible approaches to learning from their data.

Table 1

Possible Approaches to Using Independent Covariates When Analyzing Experimental Data

Research Practice	**Recommendation**
Baseline: When confronted with a single, unanticipated, promising covariate after conducting the primary analysis, ignore it. Stick to the original analysis plan.	Recommended if your priority is minimizing Type I error regardless of power
Flexible covariate: When confronted with a single, unanticipated, promising covariate after conducting the primary analysis, flexibly include it in a second analysis.	Recommended if you want to balance Type I error and power considerations.
Kitchen sink: When a primary analysis doesn't reach significance, try a series of data-dependent tests (e.g., multiple covariates, including the interaction between the covariate and the IV) until something reaches significance.	Not recommended
A priori: Before conducting a study, carefully choose a promising covariate and record a pre-analysis plan.	Recommended and ideal: Boosts power without inflating Type I error

The Current Research

A simple and straightforward way to quantify the effect of a given research practice on both Type I error rate and power is by using Monte Carlo simulations. Simulations allow us to create datasets that are sampled from a known population model. We can then analyze these datasets using different analytic practices to see how often a given practice incorrectly detects an effect when none is present (which tells us the practice's Type I error rate) and correctly detects an effect when one is present (which tells us the practice's statistical power) in the known population model. In our simulations, we focused on a simple two-group experimental design in which a researcher wants to test the effect of their manipulation, X, on their dependent variable, Y. We also simulated a covariate, C, that was independent from X in the population, reflecting the scenario of interest where a covariate is measured before the manipulation in an experiment.

We then compared the Type I error rate and power produced by the first two research practices listed in Table 1. In both cases, we imagine that a researcher tests the effect of X on Y, and then only later discovers that it would be possible to include a promising covariate in the analysis.[2] In the first, baseline research practice, the researcher forgoes including the covariate because it was not specified *a priori* and only infers the presence of an effect if the initial test of X on Y is significant at $p < .05$. In the second, flexible-covariate practice, the researcher conducts both the initial analysis without the covariate and then also a second analysis with the covariate included, and infers the presence of an effect if either the initial test of X on Y and/or the subsequent test of X on Y adjusting for C is significant at $p < .05$.[3] We assessed both the degree of Type I error inflation and the power boost produced by the second research practice (vs. the first) across varying levels of (1) sample size, (2) the true effect size of X on Y, and (3) the true correlation between C and Y in the population.

Method

We describe our simulations in detail in the Supplemental Materials; here, we simply highlight the key elements of our approach. We generated random samples of data consistent with a two-condition experiment involving a manipulated variable X, a measured dependent variable Y, and a measured covariate C. Across these simulations, we systematically varied three factors: (1) The total sample size of the simulated experiment (from $N = 40$ to 200 in increments

[2] Again, if a researcher can identify a promising covariate ahead of time, he can maximize power while holding Type I error rate at 5% by setting and recording a pre-analysis plan to conduct a single test of the effect of interest with the covariate included. This is clearly the ideal strategy, as noted in Table 1, and we return to highlight this point in the General Discussion. But here, we are interested in quantifying the tradeoffs that may confront researchers in the messy real world, where sometimes the possibility of including a promising covariate does not occur to a researcher until after he has already conducted his study and analyzed his results.

[3] This simulated strategy of conducting both tests and inferring an effect if either is significant may sound slightly different from the sequential strategy described earlier, in which a researcher tests the effect of X on Y and then pursues one of two courses of action depending on the significance of this initial test: (1) if the initial test is significant, she infers the presence of an effect, or (2) if the initial test is not significant, she continues on to test the effect of X on Y adjusting for C and infers an effect if this second test is significant. Note, however, that they are mathematically equivalent.

of 40); (2) the true effect size of X on Y (*d* = 0 to 0.80 in increments of .10); and (3) the true correlation between C and Y (ρ = 0 to .6 in increments of .10). We then analyzed each dataset using the two research practices described above: a baseline practice (a single analysis testing whether X affects Y) and a flexible-covariate practice (one analysis testing whether X affects Y and one analysis testing whether X affects Y when C is included as a covariate; this practice produces a significant result if either or both analyses are significant).

The Type I error is given by the percentage of simulated samples in which a given research practice returns a significant result (*p* < .05) when there is no effect of X on Y in the population (*d* = 0). Power is provided by the percentage of simulated samples in which a given research practice returns a significant result (*p* < .05) when there is a true effect of X on Y in the population (*d* > 0). We calculated two values of interest in order to compare them across the various simulations: *Type I error inflation*, referring to the increase in the Type I error rate produced by using the flexible-covariate (vs. baseline) practice, and *power boost*, referring to the increase in power produced by using the flexible-covariate (vs. baseline) practice.

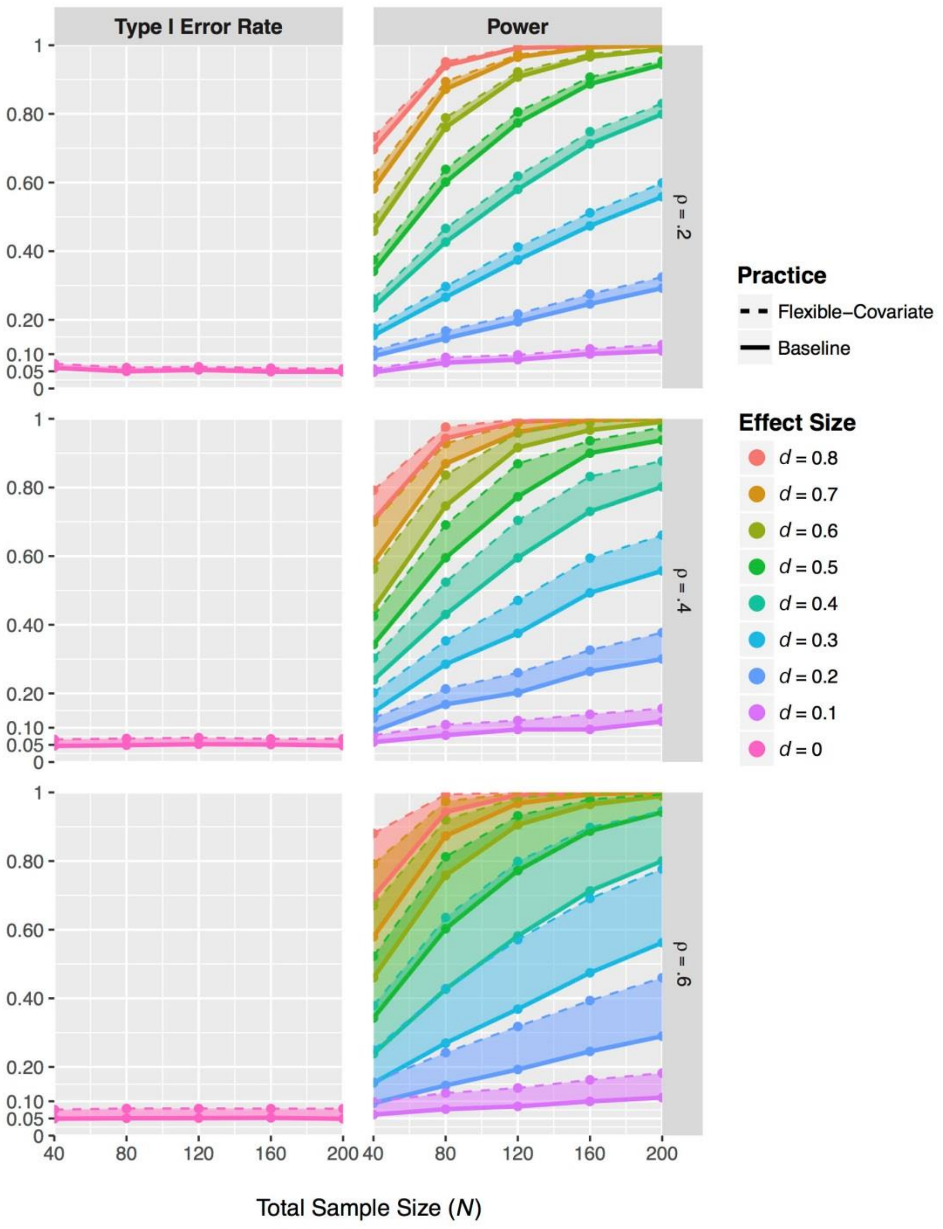

Figure 1. Type I error rate and power produced by the baseline and flexible-covariate practices as a function of the total sample size (*N*), the true effect size of X on Y (*d*), and the true correlation between C and Y in the population (ρ). The Y-axis represents the probability of finding a significant result, which is marked by solid lines for the baseline practice and dotted lines for the flexible-covariate practice across sample size and effect size. The dotted and solid lines reflect Type I error rate when *d* = 0 (left panel), and power when *d* > 0 (right panel). The width of the shaded band between the solid line and the dotted line for each color represents the extent to which the covariate (vs. baseline) practice inflates Type I error (left panel) and boosts power (right panel). Whereas Type I error inflates slightly as ρ increases, power receives a boost that becomes substantially larger as ρ increases. (For more details, see Supplemental Materials.)

Results

Figure 1 illustrates the trade-off between Type I error inflation and power boost across varying levels of sample size, true effect size of X on Y, and true correlation between C and Y in the population. Unsurprisingly, the flexible-covariate (vs. baseline) practice resulted in Type I error inflation, which increased as the true correlation between C and Y increased. However, this inflation was relatively small: Even when C and Y correlated as high as $\rho = .6$ in the population, the Type I error inflation produced by the flexible-covariate practice remained below 3% (i.e., an error rate below 8%), regardless of sample size. When C and Y correlated at .4 (akin to the correlation between neuroticism and subjective well-being or extraversion and self-esteem; DeNeve & Cooper, 1998; Robins et al., 2001), Type I error inflation was only around 2%.

Power, on the other hand, increased—often substantially—under the flexible-covariate practice. The extent of power boost depended on the true correlation between C and Y, as well as the baseline power (a function of sample size and the effect size of X on Y). As the true correlation between C and Y increased, power boost increased. As baseline power increased (because of an increase in sample size and/or a larger effect of X on Y), power boost increased, reaching a peak at moderate levels of baseline power, and then decreasing again as statistical power approached its ceiling (see Figure 2). Thus, using the flexible-covariate practice should generally give researchers the greatest power boost when their experiment would have been only moderately powered (e.g., 40%–60%) without the covariate.

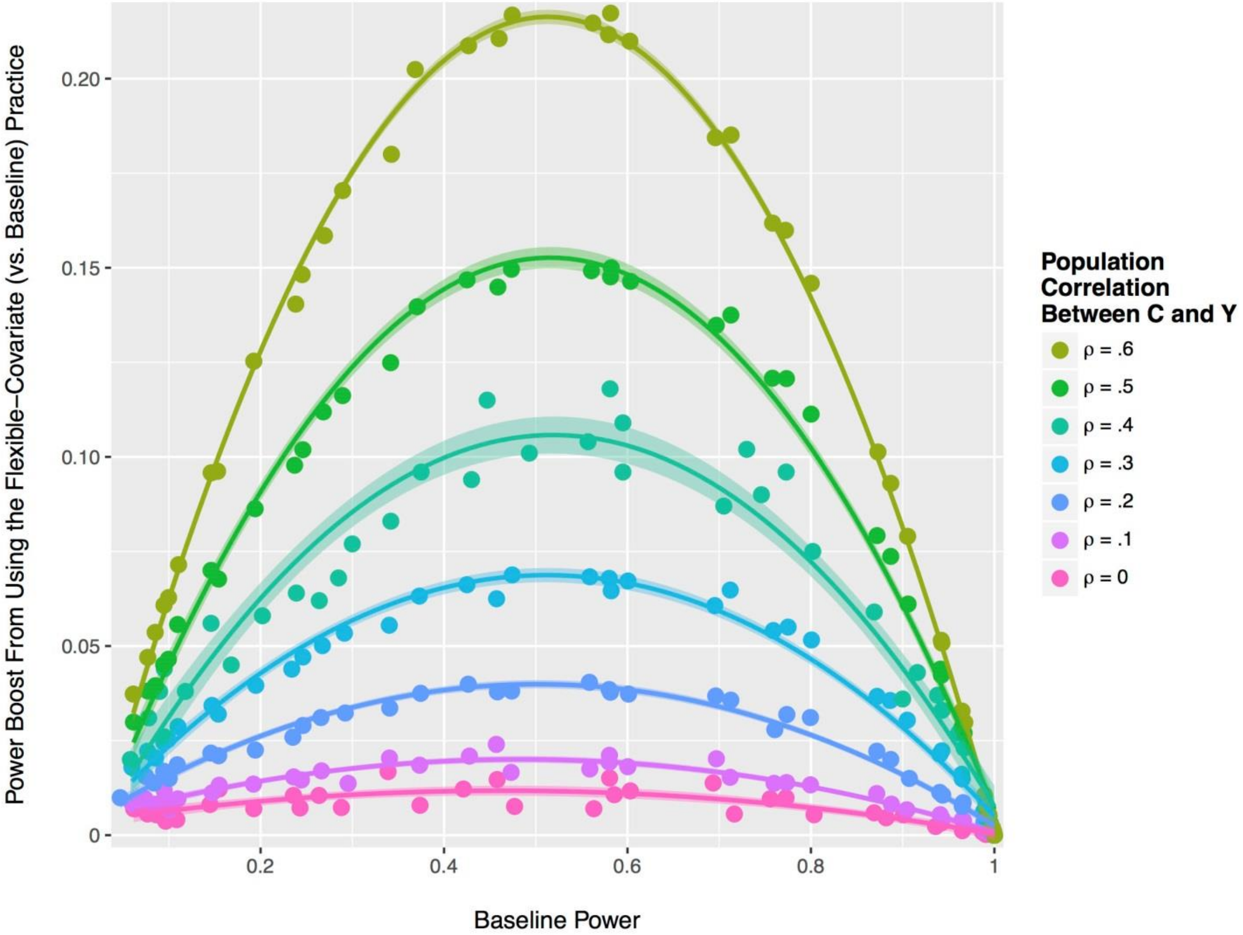

Figure 2. Power boost from using the flexible-covariate (vs. baseline) practice as a function of baseline power, shown across a range of true correlations between C and Y in the population. Each dot indicates one population that we simulated (i.e., one combination of the three factors we varied in our study design: sample size, true effect size of X on Y, and true correlation between C and Y in the population). Quadratic regressions between baseline power and power boost are fitted for each level of ρ, and regression lines and 95% confidence intervals are plotted. Power boost peaks at moderate levels (around 40%–60%) of baseline power and drops off on either side of the mid-range.

To demonstrate the trade-off more concretely, we zoom in on one typical scenario as an example (Table 2). When the true effect size of X on Y is *d* = 0.4 (approximately the average reported effect size in social-personality psychology; Richard et al., 2003; Fraley & Vazire, 2014) and the true correlation between C and Y in the population is ρ = .4, the flexible-covariate practice produces Type I error inflations of around 2%, while yielding sizable power boosts. In Table 2, we illustrate the practical impact of this research practice by converting power boost to number of additional participants needed to achieve comparable levels of power with the baseline practice. For example, with a sample size of N = 160, flexibly including a covariate results in a power boost of around 10%, which is equivalent to running 44 more participants.

Table 2

Type I Error Inflation and Power Boost Produced by the Flexible-Covariate (vs. Baseline) Practice at d *= 0.4 and* ρ *= .4.*

Total Sample Size (*N*)	Type I Error Inflation	Power Boost	Number of Additional Participants Needed for Comparable Power Boost (Δ*N*)
40	2%	6%	14
80	2%	10%	26
120	2%	10%	32
160	2%	10%	44
200	2%	8%	50

Note. The numbers of additional participants needed to achieve a comparable power boost were calculated as the difference between the total sample size in the first column and the minimum sample size required by the baseline practice to reach or exceed the (boosted) power of the flexible-covariate practice, as computed in G*Power.

Other Potential Approaches for Unanticipated Covariates

Flexibly including both C and the X*C interaction. Readers might reasonably wonder how our results compare to those of Simmons, Nelson and Simonsohn (2011), who suggested that including a covariate in an analysis is a form of “researcher degrees of freedom” (p. 1359) that can result in a potentially dramatic inflation of Type I error rate. In their simulation studies, Simmons and colleagues examined the Type I error rate for a research practice in which a researcher tests (1) the effect of X on Y, (2) the effect of X on Y controlling for C, (3) the interactive effect of X and C on Y, and (4) the effect of X on Y controlling for both C and the interaction between X and C; the researcher then reports a finding if *any* of these four tests reaches significance. We conducted additional simulations examining this four-part approach across the various sample sizes, true effect sizes of X on Y, and true correlations between C and Y included in our design. Consistent with the illustrative findings reported by Simmons et al. (2011), we found that this practice substantially inflated Type I error rates across the conditions in our design, such that Type I error regularly exceeded 10% (see Supplemental Materials for further details). In contrast, the power boost provided by the Simmons et al. approach was similar to that of the flexible-covariate practice we tested in our original simulations. For example, in the research scenario described in the preceding paragraph (where N = 160, d = 0.4, ρ = .4), using the Simmons et al. practice (vs. the baseline practice) would provide a power boost of around 11%, while more than doubling the Type I error rate (from 5% to 12%). In comparison, using our flexible-covariate (vs. baseline) practice would provide a power boost of around 10%, while inflating Type I error rate by only 2% (from 5% to 7%). In other words, allowing C to interact with X and reporting any significant effect substantially inflates Type I error beyond the flexible-covariate practice we tested in our main simulations, without providing a substantial boost to power. We therefore recommend against using the four-part practice.[4]

Flexibly including multiple covariates. What happens to Type I error rate and power when researchers consider flexibly including more than one covariate in their analysis? For example, in the scenario we described at the beginning of this paper, a researcher wonders

[4] If researchers wish to test the homogeneity of regression assumption (see e.g., Cohen, Cohen, West, & Aiken, 2003, pp. 350–351) before relying on the results of an ANCOVA model, we recommend that upon encountering a promising but unanticipated covariate, they set and record the following exploratory analysis plan: (1) Run an ANCOVA testing the effect of X on Y adjusting for C and the X*C interaction. If the X*C interaction is significant, stop. Do not use the flexible covariate strategy, and do not interpret the X*C interaction as meaningful (there is a relatively high chance that it is a false positive). If the X*C interaction is not significant, proceed to: (2) Run an ANCOVA testing the effect of X on Y adjusting for C (i.e., the flexible-covariate strategy) and follow the recommendations outlined in the Discussion below.

whether to flexibly include a single, especially promising covariate—extraversion—in an analysis testing the effect of social comparison information on self-esteem because large datasets suggest a fairly strong correlation between extraversion (the potential covariate) and self-esteem (the dependent variable). But what would happen if instead of carefully choosing this one particularly promising covariate, a researcher instead adopted a "kitchen sink" approach and decided to flexibly include each of the other Big Five factors as a covariate as well? Although it might be tempting for researchers to flexibly include an entire battery of potential covariates in their analysis, hoping to find one or another that will produce a favorable *p*-value, we discourage this kitchen-sink approach because it can inflate Type I error rates well beyond the carefully chosen flexible-covariate practice described in our main simulations. This happens because a kitchen-sink practice substantially increases the number of significance tests that are conducted on the data.

We conducted a series of additional simulations to illustrate this problem. In these simulations, the baseline practice again involved a single analysis testing whether X affects Y; the kitchen-sink practice, on the other hand, included six tests: one analysis testing the effect of X on Y, and five additional analyses testing the effect of X on Y when controlling for each of five potential covariates (analogous to a scenario in which a researcher systematically tests whether any of the Big 5 factors "works" to produce a significant effect of X on Y). The extent of Type I error inflation in such a scenario depends on how strongly each of the covariates correlates with Y, and it can get quite a bit higher than the original flexible-covariate practice we simulated above. For example, when two of the five covariates strongly correlated with Y (ρ = .4) while the other three covariates correlated more weakly with Y (ρ = .2), using the kitchen-sink practice inflated the Type I error rate to 10%. The inflation became even worse as the correlations between the five covariates and Y grew stronger: For example, the Type I error rate could reach as high as 16% when C1–C5 each correlated with Y at $\rho = .5$.[5]

The Type I error rate produced by a kitchen-sink approach also depends on how many flexible tests are considered (e.g., is each covariate tested one at a time or does the researcher also include them in pairs, trios, fours, and/or all together? Are any or all of the possible covariates allowed to interact with X?). In general, as the number of possible tests increases, so too does the Type I error rate. Thus, the Type I error rate associated with using a kitchen-sink approach in the real world becomes nearly impossible to estimate because of the many potential researcher decisions and population parameters that can affect it.[6] This means that when considering whether to adopt a kitchen-sink practice, researchers cannot make informed decisions about the likely tradeoff between Type I error inflation and power boost, because so many unknown parameters can influence the tradeoff in this context. We therefore recommend against flexibly including multiple covariates in an analysis, both because doing so can substantially inflate Type I error rates (as illustrated by the scenarios examined above), and because researchers will not have a good sense of the tradeoffs they are making between Type I error inflation and power boost.

Using the observed correlation in the sample to decide whether to flexibly include a single, promising covariate. Flexibly including a single, unanticipated covariate in an analysis

[5] In these examples, the sample size was set at $N = 200$, and the intercorrelations among the five covariates were set as .1 or below. Varying the sample size or the covariate intercorrelations (e.g., setting them to be the same as those observed in a recent meta-analysis of the Big Five intercorrelations; van der Linden, Nijenhuis, & Bakker, 2010) did not substantially change the pattern of Type I error rates (difference within 1%).

[6] The kitchen-sink strategy we considered, for instance, can produce at least 9 (effect size of X on Y: Cohen's $d = 0$–0.80 in increments of .10) × 5 (total sample size: $N = 40$–200 in increments of 40) × 7^5 (magnitude of correlation between each C and Y: $\rho = 0$–.6 in increments of .10) × 7^{10} (magnitude of correlation between any two Cs: $\rho = 0$–.6 in increments of .10) $\approx 2.56 \times 10^{14}$ potential conditions, even when we do not consider additional variations on the strategy (e.g., including two or three covariates at a time). Given the astronomical number of potential conditions, it would be exceedingly difficult for a researcher to gauge the likely tradeoffs involved in adopting a kitchen-sink strategy for a particular study by estimating all the relevant population parameters.

seems to produce a relatively reasonable tradeoff between Type I error inflation and power boost, compared to the even more flexible practices described above. But could we improve upon the single flexible-covariate practice by somehow retaining the power boost it provides while minimizing Type I error rate inflation? Given that greater power boost tends to occur at higher correlations between C and Y in the population, researchers might intuitively expect that they can simultaneously maximize their power boost and reduce Type I error inflation if they decide, upon first encountering an unanticipated but promising covariate, to only include it in their analysis if the covariate correlates with Y at above a certain level in their sample. For instance, the researcher in our example scenario may decide to only include extraversion as a covariate if it correlates with self-esteem at or above $r = .3$. In other words, could researchers use the sample correlation between C and Y to figure out which color curve in Figure 2 they are likely to be on, and then only use the flexible-covariate practice if they are on one of the higher curves?

However, additional simulations suggested that this approach does not improve the tradeoff between error inflation and power boost. Because correlations tend to fluctuate wildly when using the sample sizes typically employed in social psychology studies ($N < 250$; Schönbrodt & Perugini, 2013), testing the correlation between C and Y in a small sample will give researchers unreliable estimates of the true correlation in the population. Since the extent of both Type I error inflation and power boost depends on the correlation between C and Y in the population, *not* the correlation observed in the sample, referring to the sample correlation gives researchers a poor estimate of the tradeoff they face in reality. As a result, using the observed correlation between C and Y to decide whether to flexibly include a covariate does not improve the tradeoff between Type I error inflation and power boost (see Supplemental Materials for details). Thus, we recommend against trying to use the observed correlation between C and Y within a small (*N* < 250) sample as a decision rule for whether to include an unanticipated covariate in an analysis—it does not improve the Type I error inflation/power-boost tradeoff, and may mislead researchers into thinking they have a more precise estimate of that tradeoff than they really do.

Adjusting alpha to offset Type I error inflation. Finally, one might wonder whether reducing the alpha level for the flexible test of X on Y adjusting for C could allow researchers to hold their ultimate Type I error rate at 5% while still capitalizing on at least part of the power boost observed in our simulations. However, this approach confronts the same challenge we encountered above: In the real world, researchers do not know the actual population parameters that characterize their study (we can specify the true population correlations and effect sizes in a simulation, but in reality, of course, these numbers are unknown). Because the extent of Type I error inflation depends on these population parameters, it is impossible to know precisely what level of adjusted alpha would keep Type I error rate at exactly 5% when the population parameters are unknown. If researchers adopt an adjusted alpha value that holds the Type I error rate at or below 5% across a range of possible population parameters, this adjusted value will be too conservative under some conditions. Therefore, this approach results in a considerable loss of power—for instance, additional simulations showed that if researchers were to adopt an adjusted alpha value of .033, they would lose a substantial amount of the power boost provided by the flexible-covariate practice; in some conditions, researchers would even lose power relative to the baseline practice because the adjusted alpha value is too conservative (see Supplemental Materials for details). Therefore, for researchers who wish to prioritize holding their Type I error rate at 5% for a particular experiment, we advise against using the flexible-covariate practice with an adjusted alpha level. Instead, our recommendation for these contexts is for researchers to use the baseline practice (i.e., ignore the unanticipated covariate)—or to consider using the flexible-covariate practice as an exploratory analysis, acknowledging that it will slightly inflate Type I error for the current experiment, and to follow it up with a confirmatory replication study.

Discussion

Our results suggest that flexibly including a single, unanticipated covariate in an analysis can both slightly inflate Type I error and substantially boost power. Across a range of plausible sample sizes, true effect sizes of X on Y, and population correlations between C and Y, the Type I error rate produced by a flexible-covariate practice remained below 8%. Meanwhile, the power boost from the flexible-covariate practice was highest when the population correlation between C and Y was moderate to high, and when an experiment was moderately powered.

We hope that illuminating these tradeoffs enables researchers to make informed decisions about whether and when to consider using a flexible-covariate approach based on the relative importance of minimizing Type I error and maximizing power in a given research context. For instance, in contexts where data collection is easy and false positives are costly, researchers may wish to prioritize minimizing their Type I error rate. In such contexts, the Type I error inflation produced by the flexible-covariate practice may not be worth the boost to power, and researchers may choose instead to rerun their study with the promising covariate recorded ahead of time in a pre-analysis plan. On the other hand, in contexts where data collection is difficult and false negatives are costly, researchers may wish to prioritize increasing their power. In these contexts, the flexible-covariate practice may be worth considering as a tool for maximizing what researchers can learn from the data they have already collected.

We encourage researchers to weigh these tradeoffs for themselves, based on their own priorities and their own particular research context. With that caveat, we can offer some concrete recommendations for when and how to use independent covariates (i.e., covariates measured *before* the manipulation in experimental designs) that we think will be generally applicable to many researchers in many contexts:

1. *Whenever possible, carefully choose a promising covariate ahead of time.* If researchers (a) identify a covariate that is likely to correlate strongly with the dependent variable in the population, and (b) record their planned analysis ahead of time in a pre-analysis plan, they can gain a substantial power boost while preserving a 5% Type I error rate. This approach allows researchers to take advantage of the covariate's ability to soak up noise in the DV while entirely avoiding the Type I error inflation that results from data-dependent analytic decisions.

2. *When confronted with a serendipitous, unplanned covariate:*

a) *Consider adjusting for the covariate in the test of X on Y if it is likely to correlate strongly with the DV in the population.* Rely on large datasets or meta-analyses that do a good job of accounting for publication bias[7] whenever possible for evidence of strong correlations. Avoid relying on the observed correlation within the experimental dataset itself, since estimates can fluctuate widely from

[7] For example, look for meta-analyses that use selection methods to assess how robust the results are to different forms of publication bias (McShane, Böckenholt, & Hansen, 2016), or meta-analyses that incorporate large amounts of unpublished data (e.g., Eastwick, Luchies, Finkel, & Hunt, 2014), or meta-analyses that include data from many studies in which the result of interest to the meta-analysis was not the focal hypothesis test of the original paper (e.g., Emery & Levine, in press).

the true population parameter in smaller samples (Schönbrodt & Perugini, 2013; see Supplemental Materials).

b) *Decide whether the cost of inflating Type I error is worth the benefit of the potential power boost.* A flexible-covariate practice will be especially helpful when a study is only moderately powered and when it is difficult or impossible to collect more data.

c) *Consider registering your carefully chosen covariate before conducting the flexible-covariate analysis by downloading the simple form provided at this link: https://osf.io/pqk35*. Registering your covariate before including it in the analysis may be helpful for reassuring editors, reviewers, and readers that you conducted one and only one flexible analysis.

d) *Always transparently report all tests conducted and clearly label a flexible-covariate analysis as exploratory*. Researchers should always be transparent about the number and nature of the tests conducted. Since a flexible-covariate practice is by definition data-dependent, it should be clearly labeled as such and followed up with a replication before researchers assign high confidence to the results (van't Veer & Giner-Sorolla, 2016; Ledgerwood et al., in press). We offer a suggested template for reporting the flexible-covariate practice in Figure 3.

3. *When confronted with a scenario not modeled here:* Consider conducting your own simulations to model the tradeoffs involved in the specific scenario that you confront (e.g., encountering two promising but unanticipated covariates and conducting one flexible analysis that includes both), using the simulations reported here as a template (syntax available at osf.io/5d6hn).

Extraversion as an Exploratory Covariate. Following the recommendations of Wang et al. (2017), we included an unanticipated covariate: After conducting our planned analysis, we learned that a measure of extraversion was available, which theoretically should correlate with our dependent variable at around $\rho = .4$ (Robins, Tracy, Trzesniewski, Potter, & Gosling, 2001). In our results section, we report both the original, planned test as well as this exploratory, flexible-covariate analysis. The flexible-covariate strategy offers a tradeoff between a slightly increased Type I error rate and a substantial boost to power, and we reasoned that this tradeoff was worth making in this context to maximize what we can learn from this study.

Figure 3. Suggested Template for Reporting the Use of the Flexible-Covariate Practice in the Methods Section of a Manuscript.

Correspondingly, we also offer some notes of caution:

1. *The recommendations above apply only to experimental designs when the covariate is measured before the manipulation*. Our simulations and recommendations assume X and C are independent in the population and do not generalize to other contexts.

2. *Avoid testing the interaction between X and C or flexibly including more than one covariate*. As the number of tests increases, Type I error rate can increase substantially. For example, Simmons et al. (2011) found (and we replicated in additional simulations) that flexibly testing the effect of (1) X on Y, (2) X on Y adjusting for C, (3) the effect of X on Y adjusting for C and the X*C interaction, and (4) the interactive effect of X and C on Y can inflate Type I error rates to 11.7%. In contrast, in our simulations, flexibly including a single, carefully chosen covariate when testing the effect of X on Y inflated Type I error rates to only 6–8%, while offering a power boost comparable to that of Simmons et al.'s four-part approach.

3. *Resist the temptation to toss a covariate into an analysis "just to see if it helps,"* without a compelling reason to expect a strong correlation between C and

Y. A weakly correlated covariate will barely boost power, but can inflate Type I error—a poor trade-off to make.

By quantifying the trade-off between Type I error inflation and power boost associated with including an unanticipated covariate in an analysis, we respond to recent calls for understanding how researchers can optimize their choice of research practices (Finkel et al., in press; Ledgerwood, 2016; Miller & Ulrich, 2016), and provide recommendations for when and how to use independent covariates in experimental designs. Covariates are not simply error-inflating villains or power-boosting heroes—rather, our simulations show they have both benefits and costs. Researchers can weigh these trade-offs to make informed choices about optimal practices to use in their own research. We hope this paper can help researchers make thoughtful decisions about how to plan their experiments ahead of time as well as how to cope with unanticipated situations (i.e., stumbling upon a promising but unexpected covariate), thereby maximizing the informational value of their research.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108-119.

Brewer, M. B., & Crano, W.D. (2014). Research design and issues of validity. In H. T. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2^{nd} ed., pp. 11-26). New York: Cambridge University Press.

Cohen, J. (1988). *Statistical power for the behavioral sciences*. Hillside, NJ: Erlbaum.

Cohen, P., Cohen, J., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation for the behavioral sciences* (3rd ed.). Mahwah, NJ: Routledge.

DeNeve, K. M., & Cooper, H. (1998). The happy personality: A meta-analysis of 137 personality traits and subjective well-being. *Psychological Bulletin*, *124*, 197–229.

Eastwick, P. W., Luchies, L. B., Finkel, E. J, & Hunt, L. L. (2014). The predictive validity of ideal partner preferences: A review and meta-analysis. *Psychological Bulletin*, *140*, 623-665.

Emery, R. L., & Levine, M. D. (in press). Questionnaire and behavioral task measures of impulsivity are differentially associated with body mass index: A comprehensive meta-analysis. *Psychological Bulletin.*

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (in press). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology*.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th edition). Englewood Cliffs, NJ: Prentice-Hall.

Lakens, D., & Evers, E. R. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9*, 278-292.

Ledgerwood, A. (2014). Introduction to the special section on moving toward a cumulative science: Maximizing what our research can tell us. *Perspectives on Psychological Science, 9*, 610-611.

Ledgerwood, A. (2016). Introduction to the special section on improving research practices: Thinking deeply across the research cycle. *Perspectives on Psychological Science, 11*, 661-663.

Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology, 101*, 1174-1188.

Ledgerwood, A., Soderberg, C., & Sparks, J. (2017). Designing a study to maximize informational value. In J. Plucker & M. Makel (Eds.), *Toward a more perfect psychology: Improving trust, accuracy, and transparency in research.* Washington, DC: American Psychological Association.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data*. Belmont, CA: Wadsworth.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analyses: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*, 730–749.

Miller, J., & Ulrich, R. (2016). Optimizing research payoff. *Perspectives on Psychological Science*, *11*, 664–691.

Robins, R. W., Tracy, J. L., Trzesniewski, K., Potter, J., & Gosling, S. D. (2001). Personality correlates of self-esteem. *Journal of Research in Personality, 35*, 463-482.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609-612.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534-547.

Van ’t Veer, A. E., & Giner-Sorolla, R. (2016, September 6). Pre-registration in social psychology - A discussion and suggested template. Retrieved from osf.io/56g8e