Journal of Accounting Research

DOI: 10.1111/1475-679X.12208
Journal of Accounting Research
Vol. 56 No. 2 May 2018
Printed in U.S.A.

CHICAGO BOOTH

No System Is Perfect: Understanding How Registration-Based Editorial Processes Affect Reproducibility and Investment in Research Quality

ROBERT BLOOMFIELD,* KRISTINA RENNEKAMP,*
AND BLAKE STEENHOVEN*

ABSTRACT

The papers in this volume were published through a Registration-based Editorial Process (REP). Authors submitted proposals to gather and analyze data; successful proposals were guaranteed publication as long as the authors lived up to their commitments, regardless of whether results supported their predictions. To understand how REP differs from the Traditional Editorial Process (TEP), we analyze the papers themselves; conference comments; a survey of conference authors, reviewers, and attendees; and a survey of authors who have successfully published under TEP. We find that REP increases up-front investment in planning, data gathering, and analysis, but reduces follow-up investment after results are known. This shift in investment makes individual results more reproducible, but leaves articles less thorough and refined. REP could be improved by encouraging selected forms of follow-up investment

*Cornell SC Johnson College of Business.

Accepted by Christian Leuz. We are grateful for input from Sudipta Basu, Matt Bloomfield, Christopher Chambers, Ryan Guggenmos, Christian Leuz (Editor), Robert Libby, participants at Cornell University's Behavioral Economics and Decision Research Showcase, authors who provided feedback on our reading of their articles published in this volume, and the hundreds of conference participants and accounting researchers who responded to our surveys. An Online Appendix to this paper can be downloaded at https://research.chicagobooth.edu/arc/journal-of-accounting-research/online-supplements.

that survey respondents believe are usually used under TEP to make papers more informative, focused, and accurate at little risk of overstatement.

JEL codes: C18; I23; M40

Keywords: registered reports; reproducibility; editorial processes; research discretion; peer review

1. Introduction

The Traditional Editorial Process (TEP) begins when authors submit manuscripts reporting their conclusions from analyses they have already conducted on data they have already gathered. The articles in this special Conference issue, *Registered Reports of Empirical Research*, were published through a Registration-based Editorial Process (REP), which begins when authors submit proposals to gather and analyze data to test their predictions. Authors submitted 71 proposals in total. After at least two rounds of revision and resubmission, the eight proposals for the papers in this Conference issue received an "in-principle" acceptance from editors, guaranteeing publication as long as authors gathered and analyzed their data as promised, whether or not results supported their predictions.

How did registration affect the quality of the articles published in this issue? How could registration provide a more useful complement to the traditional process? We address these questions by examining the articles themselves; comments from conference attendees; survey responses from conference attendees, reviewers, and authors; and survey responses from hundreds of accounting colleagues who have published papers through the traditional process and share views on the challenges they faced.

Many who advocate for REP emphasize how it mitigates threats to the reproducibility of *p*-values generated by Null Hypothesis Significance Testing (NHST) (Nosek and Lakens [2014], Chambers et al. [2014]). TEP allows authors to cherry-pick their analyses, rewrite their hypotheses after they see their results, and engage in other questionable practices that overstate the predictive power of the authors' theory. TEP also biases the pool of published research because editors are unlikely to select papers that do not support predictions, and authors are unlikely to submit them.

REP mitigates questionable practices by requiring authors to gather and analyze data as they planned, and mitigates selection biases by requiring editors to make a publication decision before results are known. These changes make results more reproducible, but also change how authors invest in their projects. Under TEP, authors invest heavily in their studies after they have observed their results. They continue to gather data, refine their measures, adjust analyses to suit the distributions of their data, and rewrite to communicate more effectively. REP allows authors to make some of these follow-up investments, but provides little incentive for them to do so; by the time results are known, editors have already accepted the paper for publication. Authors are instead strongly motivated to make up-front investment in their proposal.

By our reading, the articles in this issue reflect REP's increased emphasis on up-front investment. Authors devote more time and effort to data gathering than seen in the typical paper published under TEP, proposing larger sample sizes, more intricate hand-collection and measure construction, and more challenging experimental settings. The papers also report weaker results than might typically be expected in published research. This is due, at least in part, to REP's de-emphasis of follow-up investment in unplanned analyses and revisions that can overstate predictive power—several papers' results could have *appeared* stronger if authors had strategically selected which hypotheses, subsamples, and analyses to highlight, as they could have under TEP. But de-emphasizing follow-up investment comes with a cost; most of the papers in this issue could have been more thorough and focused if the authors had continued to invest in unplanned analyses and rewriting.

Our survey of conference authors, reviewers, and attendees documents the challenges posed by REP, and points to how they might be overcome. Authors of accepted proposals found it difficult to plan studies to the level of detail required for REP, while reviewers found it difficult to evaluate proposals without the usual benefit of hindsight enjoyed under TEP, which allows them to see which analyses "worked." Many respondents believe that REP will generate higher quality articles if authors obtain more outside feedback before receiving in-principle acceptance, and if editors demand high standards for up-front investment, encourage investment in rigor over investment in scope or novelty, clarify the appropriate role of pilot data gathering, and set looser but clearer limits on the revisions authors can make after observing their results.

The net benefits of REP depend heavily on how authors actually use their discretion during follow-up investment in unplanned data gathering, analyses, and rewriting under TEP, and on how readers actually interpret authors' claims. To provide some empirical evidence on these matters, we sent a survey to recently published authors in six well-regarded accounting journals, soliciting their views and first-hand stories about how various forms of discretion in unplanned data gathering, analysis, and rewriting affect research quality. We received nearly 300 responses. Overall, respondents see discretion as improving the quality of published research in accounting. However, some forms of discretion are viewed as more harmful than others. Respondents are most concerned about overstatement when authors use discretion over which measures and analyses they report and highlight, and only slightly less concerned about how authors use discretion to exclude entire subsamples (e.g., industries, time periods). Respondents believe that authors tend to overstate their findings when they change their theories and predictions after seeing results, but see revisions to hypotheses as largely an improvement in exposition at little cost. Respondents also see substantial benefit and little cost when authors use their discretion to gather more data or exclude unusual observations.

Survey results show that authors care about far more than reproducibility, with free responses emphasizing the value of revising papers to incorporate what can be learned from observing results, respond intelligently to unanticipated outcomes, and write a more focused and understandable article. However, first-hand accounts exhibit a degree of finger-pointing that underscores the charged nature of these judgments. Authors tend to see their use of discretion as improving quality, but less so when their choice was driven by pressure from editors and reviewers. In contrast, editors and reviewers tend to see discretion as reducing quality, especially when it is driven by the authors. Responses suggest that accounting research is often not reproducible; of 113 respondents who provide a first-hand account of attempting a replication, 105 reported a failure.[1] As in a recent survey published in *Nature*, respondents most frequently attribute their replication failures to a lack of transparency that prevented them from following the same procedures used to generate the original claim (Baker [2016]). Thus, reproducibility should be enhanced by policies that encourage sharing of code and data, such as those incorporated in January, 2015 at *Journal of Accounting Research* (see current policies at JAR [2018]). But many accounts suggest that results were not reproduced because they reflected questionable practices or selection biases that would be difficult to address without registration. Many respondents note the difficulty of publishing replications, especially failed ones, further supporting concerns about selection bias and the value of accepting papers for publication before results are known.

We found it challenging to summarize the thoughtful and nuanced responses to our survey questions, and strongly encourage anyone interested in this topic to read them in full and discuss them with colleagues, students, and advisors. Many are reproduced below, and all are available in the online appendix. Our own discussions lead us to conclude that no system is perfect. Registration is a useful complement to the TEP, but is unlikely to be an adequate replacement. By encouraging authors to shift from follow-up investment to up-front investment, REP encourages careful planning and ambitious data gathering, and reduces the questionable practices and selection biases that undermine the reproducibility of results. But the reduction in follow-up investment leaves papers in a less-refined state than the traditional process, leaving useful work undone. Because accounting is a small field, with journals that typically publish a small number of long articles, subsequent authors may have no clear opportunity to make a publishable contribution by filling in the gaps.

[1] Our survey does not allow us to assess the proportion of results that are unreproducible in accounting. A poll by *Nature* suggests that across a range of fields, about half of scientists believe their field has a "significant reproducibility crisis," and about 70% think that at least half of the results in their field are not reproducible (Baker [2016]). Recent analyses found replicability rates of about 40% in psychology (Open Science Collaboration [2015]) and 10% in oncology (Begley and Ellis [2012]).

With experience, we expect that authors and editors will learn which editorial process is better suited to which types of studies, and learn how to draw inferences differently from papers produced by these very different systems. We also see many ways that both editorial processes could be improved by moving closer toward each other. REP would be improved by encouraging more outside input before proposals are accepted, and more extensive revisions after planned analyses are conducted, especially those relying on forms of discretion that our community sees as most helpful and least harmful. TEP would be improved by demanding more complete and accurate descriptions of procedures (as *Journal of Accounting Research* has been implementing for several years; updated JAR [2018]), not only those that help subsequent authors follow those procedures, but also those that help readers interpret *p*-values in light of the alternatives that authors considered and rejected in calculating them. REP and TEP would complement one another more successfully if journals would be more open to publishing short articles under TEP that fill in the gaps left by articles published under REP.

The remainder of this article is organized as follows. In section 2, we describe the accounting framework we use to evaluate the quality of scientific reports and to predict the impacts of REP. In section 3, we describe the editorial process for the conference. In section 4, we present our analyses of the papers in this volume. In section 5, we discuss our survey of conference authors, reviewers, and attendees. In section 6, we discuss our survey of authors who have published under TEP. In section 7, we summarize key lessons we take from the conference and our surveys.

2. *Background*

Almost all peer-reviewed articles in social science are published under what we call the TEP. Authors gather their data, analyze it, and write and revise their manuscripts repeatedly before sending them to editors. Editors send promising manuscripts to one or more reviewers and recommend revisions. Authors are given the opportunity to revise their manuscript in response, often multiple times, before it is either rejected or accepted.

Many disciplines have been introducing REPs as an alternative to TEP.[2] Under REP, authors propose a plan to gather and analyze data to test their predictions. Journals send promising proposals to one or more reviewers and recommend revisions. Authors are given the opportunity to revise their proposal in response, often multiple times, before the proposal is either rejected or granted in-principle acceptance, which guarantees publication as long as the authors live up to the commitments in their proposal, regardless of whether results support their predictions.

[2] See the Center for Open Science (https://cos.io/rr) for a list of journals adopting versions of REP and some of the reasons for doing so.

2.1 REP AND REPRODUCIBILITY

A common argument for REP, offered by Nosek and Lakens [2014] and Chambers et al. [2014], is that it improves the reproducibility of researchers' empirical claims, particularly those based on *p*-values from NHST. NHST and *p*-values have serious shortcomings, but these are largely independent of editorial processes.[3] Given their widespread use, we assume that authors present their claims in the form of *p*-values, however flawed this approach may be, and restrict our focus to how switching from TEP to REP addresses two threats to reproducibility: questionable research practices and selection biases.

2.1.1. Questionable Research Practices. A result is reproducible if an independent researcher can generate the same inferences using similar methods (Nosek et al. [2015]). While there is considerable controversy over what constitute the "same inferences" and "similar methods,"[4] there is broad agreement that reproducibility is undermined by authors' wide discretion over how they conduct research and what they report to readers, which allows them to engage in a variety of what are commonly called "questionable research practices" (Martinson, Anderson, and De Vries [2005]). Researchers can gather data on many variables and report only those that have low *p*-values (cherry-picking) and Hypothesize After Results are Known (HARKing) to exaggerate the predictive power of the theory supposedly driving those results. They can conduct many analyses and report only those that generate low *p*-values (*p*-hacking). They can continue gathering data until they generate a low *p*-value and then stop (the stopping problem). All of these "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn [2011]) increase the likelihood of generating results that are unlikely to be reproduced by an independent researcher who does not know how the original authors navigated the "garden of forking paths" (Gelman and Loken [2014]) that led to the result they are re-examining.

To understand how REP is likely to mitigate questionable practices, we apply the "fraud triangle" (Cressey [1953], Committee of Sponsoring Organizations of the Treadway Commission (COSO) [1992, 2013]), which can be used to assess the likelihood of any form of self-interested misbehavior, including those that fall far short of fraud. The triangle identifies three conditions that lead to misbehavior: pressure, opportunity, and rationalization. People are more likely to engage in misbehavior when doing so helps them address pressures they face, when they have the opportunity to

[3] NHST and *p*-values were criticized decades ago as "fallacious" by Rozeboom [1960] even when performed well, and current practice has been compared to religious rituals by Gigerenzer [2004]. Yet more recently, McShane et al. [2017] and the American Statistical Association (Wasserstein and Lazar [2016]) have issued strong statements cautioning against their use.

[4] See Nosek et al. [2015] for one approach to defining "same inferences." See Hamermesh [2007] for various definitions of "similar methods."

misbehave without detection or punishment, and when they can rationalize their choices as appropriate.

Under TEP, editors repeatedly evaluate how much readers will learn from a manuscript, and accept it for publication when they believe readers will learn enough, relative to some absolute standard and also relative to what readers would learn from additional revisions. This form of evaluation pressures authors to overstate their claims in pursuit of publication, and gives them the opportunity to do so because authors can revise their papers extensively after observing their results but before editors see their manuscript. Given that revisions with the potential to undermine replicability are so widespread and accepted (see, e.g., John, Loewenstein, and Prelec [2012]), authors will likely find ways to rationalize them. Thus, all three legs of the fraud triangle operate in ways that make misbehavior likely.

Under REP, authors face less pressure to overstate what readers will learn from their results because those results are not considered by editors in the publication decision. Authors face less opportunity, because they must stick to their proposal. They will also find rationalization more difficult, because the policies and norms surrounding REP highlight the importance of avoiding questionable practices. By addressing all three legs of the fraud triangle, REP is likely to limit the misbehaviors that undermine reproducibility.

2.1.2. Selection Biases. Reproducibility is also threatened by biases in the selection of which papers appear in print. There is a widespread belief, supported by evidence from Franco, Malhotra, and Simonivitz [2014], that editors tend to publish papers that support stated predictions in favor of those that do not (publication bias). Knowing this, authors may have a tendency to only submit papers with positive results (file-drawer bias). Authors can exploit selection biases by conducting studies with very low power, which increases the chance that they will provide support for a novel but incorrect theory simply due to random error (capitalizing on chance).[5] This behavior further undermines reproducibility by providing editors with many opportunities to accept papers with positive (but wrong) results, while rejecting negative (but correct) results.

REP enhances reproducibility by reducing selection bias. Editors cannot bias publication toward papers with positive results because they must offer in-principle acceptance before those results are known. Authors cannot bias their submissions toward papers with positive results because they must submit their proposals before observing results. REP also makes it unlikely that authors can capitalize on chance because proposals for low-power studies are unlikely to be accepted.

[5] https://www.sciencenews.org/blog/scicurious/blame-bad-incentives-bad-science

2.2 BEYOND REPRODUCIBILITY

Discussions of reproducibility tend to focus attention on individual tests that generate *p*-values below a threshold of significance (usually 0.05). These represent claims that can be reproduced. To understand the full impact of REP on the quality of published research, we must also consider how REP alters the investments authors make in their work, and the context in which readers interpret and extend their claims.

2.2.1. Investment. REP changes the nature and timing of authors' investment in their work. Authors provide contributions by planning, gathering data, analyzing data, and reporting results. Not every contribution is a *p*-value. Bloomfield, Nelson, and Soltes [2016] identify five distinct goals of an empirical literature and argue that most papers contribute to only some of them. A paper can specify theory more clearly or completely, document statistical associations predicted by their theory, attribute those associations to causal factors, generalize associations and attributions across settings, and place results in broader contexts that will allow others to specify better theories. All of these goals have value, and all of them require investment in planning, data gathering, and analysis; only the documentation of statistical associations involves *p*-values.

Following Governmental Accounting Standards Board (GASB) [2008], we view planning, data gathering, and analysis as "outputs" because they are largely under authors' control, and are undertaken with the hopes that they will lead to the "outcomes" of interesting contributions. REP is an output-based system of evaluation that rewards authors for proposing studies that reflect careful planning, ambitious data gathering, and thorough analysis. TEP is an outcome-based system that rewards authors for finding interesting results.

REP's output orientation is likely to encourage authors' investment in planning for ambitious data gathering and thorough analysis, because these are the primary basis for editors' evaluation. Ideally, the authors plan well enough and ambitiously enough that their study will find something interesting even if their guiding theory is wrong. But in the event a study fails to generate a contribution, in-principle acceptance guarantees that the manuscript will still be published. One might question how much benefit accrues to authors who publish "failed" studies in prestigious outlets, but we believe they benefit more than if the study is relegated to a file drawer (or a less prestigious outlet), as it likely would under TEP. By bearing the risk of publishing such studies, editors who adopt REP further encourage authors to make up-front investments in output. Increased upfront investment will increase quality to the extent that authors can invest wisely and editors and reviewers can evaluate those investments well.

REP is likely to reduce follow-up investment in outputs after results are observed. Under TEP, authors have tremendous incentive to refine how they sift through their data to find the most interesting contributions, and how they characterize their implications in light of the theory that guided

them, because doing so increases the likelihood of publication. Under REP, these follow-up investments do not increase the likelihood of publication, because authors already have an in-principle acceptance by the time they observe their results. REP also limits how much authors can use follow-up investment to persuade readers that their results are interesting and important, because it would largely undermine the purpose of REP to allow unplanned analyses and postresult rewriting to dominate the abstract, introduction, and conclusion of a registered report.

It is difficult to know how REP's reduction of follow-up investment will affect research quality. Follow-up investment provides many opportunities to overstate results. But it also allows authors to gather more and better data, create better measures, construct analyses that are more appropriate to the actual distribution of their data, write a more focused and clear manuscript, and refine their study in many other ways. The more authors use follow-up investment to overstate, rather than refine results, the more REP's restriction of follow-up investment will improve the quality of published research.

2.2.2. Context. REP's benefits to readers depend partly on whether they interpret *p*-values in isolation or in context. Advocates of reproducibility focus on readers who interpret reported *p*-values taken in isolation, since this is how reproducibility is assessed. If REP reduces overstatement at little cost of discouraging refinement, each reported *p*-value will be more reproducible. But reproducibility will provide less benefit to readers with the sophistication to interpret *p*-values in context, as they are encouraged to do by the American Statistical Association (Wasserstein and Lazar [2016]). Sophisticated readers draw inferences about an individual *p*-value by considering all of the methods and analyses that are reported, as well as those that are not reported. For example, a sophisticated reader who sees that authors have asked five postexperimental questions, but analyzed and reported only one, may conclude that the authors engaged in some cherry-picking, and view the reported *p*-value skeptically. They will also draw inferences from their understanding of the authors' incentives to misreport, and the quality of the editorial systems' defenses against such misreporting. To the extent these incentives and defenses are well known, readers can adjust their inferences to correct for the bias expected to arise through questionable practices, while still benefitting from the authors' additional investment.

REP's benefits over the longer term depend on how future authors and editors will extend the literature. Scholarship is expensive and time consuming, and page-space in respected journals is constrained. Authors and editors must therefore balance the value of writing and publishing an article against the opportunity cost of another article not written or published. Editors must also decide how much a paper must accomplish in order to be publishable. Relative to other fields, accounting has a small number of researchers, who publish few, but long, articles. This may be one reason that follow-up investment is so extensive under TEP. The growth of our

literature will be hindered if authors leave results on the table, but subsequent authors cannot meet the bar for publication by cleaning up what was missed.

2.3 IMPLEMENTATION AND IMPROVEMENT

The preceding analysis assumes a best-case scenario for the implementation of REP and TEP. But even after long experience with TEP, editors are still proposing ways to improve it. *Journal of Accounting Research* has recently revised policies to encourage sharing of programming code and data (JAR [2018]). The American Psychological Association has recently expanded the reporting requirements for methods, data, and analysis (Appelbaum et al. [2018]), while the *Journal of Financial Reporting* (JFR) has introduced different standards for evaluating the substance of a paper and evaluating its commentary (JFR [2017]). Given that authors, reviewers, and editors have limited experience with REP, it is likely to fall short of its best-case advantages and leave many opportunities for improvement.

As we examine responses to our surveys, we look for opportunities to improve REP's implementation. These include clarifying the standards for in-principle acceptance and follow-up revision, encouraging the types of planning and up-front investment that are best suited for REP, and allowing follow-up investments that pose little threat to reproducibility but offer significant improvements in other aspects of research quality. We also look for ways to make articles easier to read. REP forces authors to specify in detail their planned procedures and any revisions, encouraging prose that is long, dense, and peppered with discussion of false starts, all of which reduce readability (Zimmerman [1989]). We look for ways that editors might accommodate studies that REP makes more valuable, particularly those that fill in gaps left by limited follow-up investment. Finally, we look for opportunities to improve the reproducibility of articles published under TEP, without requiring a change as dramatic as registration.

3. Design and Implementation of REP

To summarize the goals and processes of REP as it was implemented for this conference, we rely on the Conference Call for Papers (CFP) and postconference correspondence with authors.

3.1 CALL FOR PAPERS

The appendix includes key excerpts from the CFP, the associated Policies for Authors and Reviewers, and text incorporated into postconference letters from editors to authors.[6] The CFP was designed to attract proposals over a wide range of methods and topics (section 1 in the appendix). A

[6] https://research.chicagobooth.edu/arc/journal-of-accounting-research/call-for-papers#simple2

link to detailed policies, included in the CFP, described the basic output-based structure of REP: proposals that receive in-principle acceptance are published regardless of whether results support predictions (section 2 in the appendix).[7] The policies emphasized three potential benefits of REP: to encourage up-front investment in data gathering (section 2 in the appendix), to enhance reliability by limiting author discretion in planned analyses (section 2-3 in the appendix), and to accelerate input from reviewers and editors so that it could be used to guide and improve data gathering (section 2-4 in the appendix). Investment was further categorized as being directed toward improving the rigor (section 2-5 in the appendix), scope (section 2-6 in the appendix), or novelty of data gathering (section 2-6 in the appendix).

Anticipating the challenges of planning analyses before observing data, the policies encouraged authors to anticipate contingencies in their analyses by adopting statistical techniques (such as robust regression or canonical correlations) that are robust to unexpected realizations of data distributions (sections 2-8 and 2-9 in the appendix), to use power analyses and Bayesian approaches to establish that their designs were likely to yield informative results (section 2-10 in the appendix), and to gather pilot data to demonstrate that the proposed data could be feasibly gathered without too many missing observations or other problems that would undermine the study (section 2-11 in the appendix).

3.2 REVIEWER GUIDANCE

The policies included the guidance provided to reviewers (section 3 in the appendix). In evaluating proposals, referees were directed to evaluate the study's motivation, investment, theory, feasibility, quality of planned analyses, and the likely importance of conducting additional analyses. Reviewers were also asked to assess the likelihood that the study would be informative regardless of the realization of the data, and to describe any concerns that might require additional conditions attached to in-principle acceptance. In evaluating final reports, referees were directed to evaluate how well the authors lived up to their commitments, how well the final report described what was actually done and found (with clear indications of any deviations from the proposal), whether additional analyses were adequate, and whether the interpretation of the results was appropriate.

3.3 POSTCONFERENCE REVISION POLICIES

Immediately after the conference, the Editors met to discuss the next step of the editorial process and to determine the extent of revisions that would be allowed. The Editors settled on a three-tiered approach to revision (section 4 in the appendix). Tier 1 items included the operational

[7] Editors were allowed to impose some data-related conditions for publication (e.g., achieving a high enough response rate). These conditions were imposed on only a few papers and were never invoked.

TABLE 1
JAR Conference Submissions by Method

Methodology	Number of Proposals	Percentage of Total Proposals	BNS (2016)
Empirical-archival	51	72%	63%
Laboratory experiments	10	14%	11%
Field experiments	4	6%	0%
Surveys	4	6%	10%
Analytical/simulation	2	3%	10%
Nonstatistical	0	0%	6%
Total	71	100%	100%

This table summarizes the number of proposals received, by methodology, and compares the percentage breakdown across methodologies to the percentages summarized in Bloomfield, Nelson, and Soltes [2016]. Bloomfield, Nelson, and Soltes [2016] compile the percentage of papers, by methodology, that are published between 2003 and 2013 in the *Journal of Accounting Research*, *The Accounting Review*, the *Journal of Accounting and Economics*, and *Accounting, Organizations and Society*.

details of data gathering and analysis, which authors were permitted to revise only sparingly, if at all. Tier 2 items included definition and description of theoretical constructs, which authors were permitted to revise as long as their motivation was not driven primarily by results.[8] Tier 3 items included motivation, literature review, and other discussions and speculations about the broader implications of the study. Authors were permitted wide latitude in such revisions, which were not viewed as essential to the design of REP and would be likely to improve readability. All tier 1 and tier 2 revisions were required to be acknowledged in the paper.

4. Analysis of Conference Submissions

JAR received 71 proposals. As indicated in table 1, the submissions covered a range of methods, with 51 (72%) proposing to hand-gather archival data, 10 (14%) proposing to conduct laboratory experiments, and 4 (6%) proposing to conduct field experiments. These proportions are fairly similar to those observed in recently published articles, as documented by Bloomfield, Nelson, and Soltes [2016], suggesting that authors viewed REP as an equally attractive publication channel regardless of their method. Just over half of the proposals (36) were sent to reviewers.[9] All 18 author teams receiving a revise-and-resubmit submitted a revision, of which 10 were sent on for a second review. Of those, eight received in-principle acceptance and were published in this volume.

[8] For example, the emphasis in Bernard et al. [2018] changed from examining the effects of "dispersion of ownership" to the effects of "direct investment" by investors, as the latter mapped better to the construct of interest, and more clearly described the contribution of the paper.

[9] The high desk-reject rate reflects the editors' recognition that even a promising proposal was unlikely to yield a high-quality publication without deep investment by the reviewers.

Our earlier discussion and the nature of the REP process outlined in the CFP lead us to expect that the papers in this volume are likely to differ from TEP-based publications in several ways. We expect more up-front investment in data gathering, analysis, and other outputs, since these are the basis for acceptance; weaker but less overstated results, since positive results are not a consideration for acceptance; and less follow-up investment after authors observed results of their planned tests, since these are not needed for publication and may be seen as violating the spirit of REP. We use the eight conference papers to evaluate each of these predictions.

4.1 INVESTMENT

Our reading of the papers leads us to conclude that REP's output-based approach encouraged authors to invest more up-front, relative to papers under TEP, in the "tedious and difficult task" (Vatter [1966]) of gathering data.

Three papers hand-collect archival data. Allee, Deangelis, and Moon [2018] develop a new process for measuring the degree to which a firm's disclosures are accessible through computer algorithms (scriptability), which they apply to approximately 1.3 million documents. Ertimur et al. [2018] hand-collect data on the employment dates of 2,036 external CEO hires from news articles and various SEC filings (of which 1,538 were included in their final sample). Hail, Tahoun, and Wang [2018] oversee dozens of research assistants who gathered information on press coverage of corporate scandals and regulations in 26 countries over 216 years to evaluate how those two time series are associated. While there are surely archival papers that reflect comparable or greater investment, all of these studies seem to exceed the typical level of investment in published articles using this method.

Three conference papers collaborate with organizations to conduct field experiments. Van Duin et al. [2018] collaborate with the Dutch Authority for Financial Markets to manipulate the degree of supportiveness conveyed by a letter they send to thousands of financial intermediaries. Eyring and Narayanan [2018] collaborate with designers and instructors of large-scale online statistics courses to test the effect of changing comparative reference points provided to tens of thousands of students. Li and Sandino [2018] work with a large Indian retail chain to examine how an information-sharing system affects employee creativity, engagement, and performance. We are not sure that these field experiments involve much more investment than the typical field experiment, but field experiments are famous for requiring a great deal of investment, which is one reason this method is used in only two of the 1,638 papers published in four top accounting journals from 2003 to 2013 (Bloomfield, Nelson, and Soltes [2016]).

Two papers are best categorized as laboratory experiments. Kowaleski, Mayhew, and Tegeler [2018] mostly follow the traditions of experimental economics, bringing cohorts of participants into the lab for hours at a time and offering cash incentives. With 264 participants earning an average

of $22 in payments, we see this as a high-investment study. Bernard, Cade, and Hodge [2018] also reflect unusually high investment, by endowing MBA students at three different universities with the equivalent of Starbucks stock, and testing whether they spend more money at Starbucks over several weeks. Rather than relying solely on self-reported spending, Bernard, Cade, and Hodge [2018] collaborate with Starbucks to collect actual spending data from the Starbucks card and smartphone app. See table 2 for a summary of the data gathering investment across each of the conference papers.

4.2 STRENGTH OF RESULTS

To assess the strength of the papers' results, we tabulate their predictions, analyses, and results in table 3. We exclude Hail, Tahoun, and Wang [2018] from our tabulation because it does not state formal hypotheses. Of the 30 predictions made in the remaining seven proposals, we count 10 as being supported at $p \leq 0.05$ by at least one of the 134 statistical tests the authors reported. The remaining 20 predictions are not supported at $p \leq 0.05$ by any of the 84 reported tests. Overall, our analysis suggests that the papers support the authors' predictions far less strongly than is typical among papers published in JAR and its peers.

Eyring and Narayanan [2018] is the only paper that reports results comparable in strength to the typical paper published under TEP, supporting six of its eight hypotheses. The paper provides solid evidence that reference points matched more closely to performance levels are more effective in encouraging effort. Not coincidentally, this study addresses one of the better studied questions in a large literature on goals and motivation, and the study uses a clean experimental design with fairly well-understood data and relatively high power. Thus, this study's strong results may reflect the fact that its investment primarily took the form of enhanced rigor, rather than scope or novelty.

Before attributing the weak results of the other papers to the usual culprits (the theories are wrong or the studies lack power), we must first consider how much of this weakness is driven by a reduction in overstatement. We find it easy to imagine revisions of several conference papers would allow them to report results of strength comparable to those found in most papers published under TEP, resorting only to practices well within common (if questionable) norms (Simmons, Nelson, and Simonsohn [2011], Gelman and Loken [2014]).

For example, Allee, Deangelis, and Moon [2018] conducted 72 planned statistical tests of their Hypothesis 1, based on three Scriptability measures (CompScript, IdentifyData, and DataToInformation), four types of disclosure (10-K, 10-Q, 8-K, and DEF 14A), three windows of time (5-minute, 60-minute, and 24-hour), and two measures of intraperiod timeliness (price-based and volume-based). Of these 72 tests, 13 (18%) are statistically significant at the 5% level with the predicted sign.

TABLE 2
Data Gathering Investment in Conference Papers

Paper	Data Gathering
Tone from Above (*van Duin, Dekker, Wielhouwer, and Mendoza*)	Field experiment spanning across 7,242 firms, yielding a sample size of 4,577 firms, in collaboration with the Authority for the Financial Markets, the supervisor of the financial markets in the Netherlands.
Disclosure "Scriptability" (*Allee, DeAngelis, and Moon, Jr.*)	Development and validation of a new measure of "scriptability," with empirical tests using this measure and its components applied to approximately 1.3 million documents across four types of disclosures, for two measures of intraperiod timeliness, and over three periods of time.
Benefits of Direct Stock Ownership (*Bernard, Cade, and Hodge*)	Field experiment including 280 graduate business students enrolled in introductory financial accounting courses across three universities with data collection involving solicitation of participants' credit and debit card statements, participants' Starbucks card information, and responses to an in-class poll and an online survey.
Bridging the Gap (*Ertimur, Rawson, Rogers, and Zechman*)	Hand-collection of data for 2,036 new CEO-firm pairs, including CEO employment gaps, existence and enforceability of noncompete constraints, and characteristics of turnover events.
Performance Effects of Reference Points (*Eyring and Narayanan*)	Two field experiments conducted over four online statistics courses with combined enrollment of 52,611, yielding samples of 15,171 for the main experiment and 4,460 for the supplemental experiment.
Corporate Scandals and Regulation (*Hail, Tahoun, and Wang*)	Hand-collection of data for 2,244 corporate scandals and 1,081 regulations enacted across 26 countries from 1800 to 2015, identified and classified using primary and secondary sources retrieved from library catalogues, online databases, Internet searches, and through contact with local experts.
Consulting Services on Audit Quality (*Kowaleski, Mayhew, and Tegeler*)	Interactive market-based laboratory experiment including 264 undergraduate and graduate students assuming various roles over multiple periods.
Information Sharing (*Li and Sandino*)	Field experiment conducted over a nine-month period across 36 stores owned by a mobile phone retail chain in India. Development and pilot-testing of an information-sharing app for use in the experiment by store employees. Creation of a customer rating app so that 683 posters could be rated eight times each, in batches of 25.

This table summarizes the data-gathering process, and investment in data gathering, for the eight papers that received an in-principle acceptance from submitting a proposal to the JAR conference.

However, the authors support 10 of 18 tests of their H1 for 8-K filings. Narrowing their focus further to the 60-minute window or the volume-based intraperiod timeliness measure would further increase the number of tests supporting their H1 to five of six (83%) or seven of nine (78%),

TABLE 3
Conference Paper Predictions Registered and Supported

Paper	Hypothesis	Statistical Tests	Significant Statistical Tests ($p \leq 0.05$)	Percentage of Total Tests of Hypotheses Supported
Tone from Above (*van Duin, Dekker, Wielhouwer, and Mendoza*)	1	8	0	0%
	2	4	0	0%
	3	8	1	13%
Disclosure "Scriptability" (*Allee, DeAngelis, and Moon, Jr.*)	1	72	13	18%
	2	24	3	13%
Benefits of Direct Stock Ownership (*Bernard, Cade, and Hodge*)	1	2	0	0%
	2	1	0	0%
Bridging the Gap (*Ertimur, Rawson, Rogers, and Zechman*)	1a	4	1	25%
	1b	4	0	0%
	2a	24	16	67%
	2b	15	0	0%
	3a	24	0	0%
	3b	15	0	0%
Performance Effects of Reference Points (*Eyring and Narayanan*)	1	1	1	100%
	2a	1	1	100%
	2b	1	1	100%
	2c	1	0	0%
	3a	1	1	100%
	3b	1	1	100%
	3c	1	1	100%
	4	1	0	0%
Corporate Scandals and Regulation (*Hail, Tahoun, and Wang*)	N/A	116	As this paper made no formal predictions, results of individual statistical tests cannot be interpreted as consistent or inconsistent with the authors' hypotheses.	
Consulting Services on Audit Quality (*Kowaleski, Mayhew, and Tegeler*)	1	1	0	0%
	2	1	0	0%
	3	1	0	0%
Information Sharing (*Li and Sandino*)	1a	2	0	0%
	1b			
	2a	1	0	0%
	2b			
	3a	1	0	0%
	3b			
Total	30	336	40	12%

This table summarizes the predictions that were registered versus supported in the eight papers that received an in-principle acceptance (seven conference papers, as well as Van Duin et al. [2018], which were not completed before the conference). For each paper, we outline the number of registered hypotheses, the number of statistical tests of each hypothesis, the number of these tests that were supported (at p-values ≤ 0.05), and the percentage of tests that supported each hypothesis.

respectively. Moreover, the authors could justify this decision by arguing that the forward-looking and voluntary nature of these 8-Ks makes them the disclosures most appropriate for testing their H1, and that trimming the other analyses would make the paper more understandable and less weakened by the errors that would arise from using data and measures ill-suited to their research question.

Li and Sandino [2018] yielded no statistically significant support for their main hypotheses. However, they found significant results in their planned additional analyses that are consistent with informal predictions included in the accepted proposal. Specifically, they find significant improvements in the quality of creative work for stores that accessed the information-sharing system more frequently and stores with fewer nearby same-company stores. They also found improvements in the quality of the creative work and employee engagement in stores in "divergent" markets where there are greater needs to adapt advertising to local conditions. The authors could justify highlighting these analyses by arguing that stores benefit only if they actually use the information sharing system, and that they will have the most power to detect effects in isolated or divergent stores that do so.

We are not suggesting that these particular authors would have or should have made these revisions under TEP. Our point is merely that they could have, without violating common norms, so we are not ready to conclude that the studies in this issue actually provide weaker support for their predictions than most studies published under TEP.

4.3 ADDITIONAL ANALYSES

We next examine whether the papers reflect less follow-up investment that would have improved the papers (rather than merely overstating results). To identify opportunities for refinement, we rely on our real-time notes from conference discussions.[10] We summarize what could be done to extend each of these articles, but do not indicate who, if anyone, should do that work. We believe that a more traditional process, which rewards outcomes, would likely have led the most worthy of these additional steps to be taken before publication, but see no reason to pass judgment on authors for choosing to leave these matters to future studies.

Conference attendees raised a number of directions to flesh out Ertimur et al. [2018] through additional data gathering and analysis. Additional data gathering could provide insight into how CEOs spent their time during gap years, which would shed more light on the causes and consequences of long gaps. A combination of economic modeling and analysis could help identify the "efficient" gap that would arise assuming perfectly functioning labor markets and optimal tradeoffs between the costs and benefits of gaps to CEOs. Additional data and analyses might also document associations

[10] Van Duin et al. [2018] was not complete in time to be presented at the conference.

between gaps and various CEO characteristics, such as risk-taking, overconfidence, and reporting aggressiveness.

Li and Sandino [2018] could be shored up with more analysis documenting whether the variance in the quality of posters changed due to the information-sharing system (i.e., whether the sharing of information reduces the number of both very low-quality and very high-quality posters). Interesting results could also be uncovered by analyzing the effects of employees' age and education level, which could be predictors for the level of underlying comfort with the technology used in the system.

For Allee, Deangelis, and Moon [2018], attendees asked a number of questions that could be addressed with additional processing of the text already gathered. How does scriptability relate to boilerplate disclosures? Are scriptable disclosures more similar to other scriptable disclosures in content or style? Are scriptable reports providing better information to everyone, or the same information to more people? Do the results hold for changes in scriptability, in addition to the levels reported in the paper? Attendees also raised questions about the appropriate equilibrium analysis; economic modeling could guide analyses based on assumptions about the costs of writing and implementing scripts that would accommodate less scriptable disclosures.

Bernard, Cade, and Hodge's [2018] results could be strengthened by gathering additional data from participants who are given larger or variable investments in Starbucks. With the existing data, results might be teased out through analyses that more completely exploit variation in some key student attributes (like wealth) and in the timing of when participants are spending money at Starbucks. Additional work could clarify the nature of brand awareness, and perhaps concentrate more effectively where the effects seem to be largest (e.g., with coffee and tea drinkers).

Eyring and Narayanan [2018] provide a very rich data set that would allow much more analysis of the structural relationship between effort and the distance between performance and reference points. The existing study also provides many opportunities for follow-up experiments that would clarify the mechanism and scope of the effects, manipulating factors like the nature of targets and feedback provided, or allocating targets dynamically based on performance.

Kowaleski, Mayhew, and Tegeler [2018] find results that suggest a number of follow-up questions. Why did auditors not cooperate in the expected manner (as discussed in footnote 19 of Kowaleski, Mayhew, and Tegeler [2018])? What is the role of strategic uncertainty? Additional experiments could confirm the exploratory finding on variance or explore when and why managers demand high versus low audit quality.

Hail, Tahoun, and Wang [2018] could be extended by testing for specific underlying drivers of both scandal and regulation (e.g., new technologies, business models, investment relationships) by remeasuring variables to capture the size and importance of the scandals and regulations currently represented as being identical. Additional work, probably requiring more data,

could clarify the relationship of particular scandals with particular regulations, distinguish regulations enacted for their substantive impact from those designed to protect regulators' reputations, and document spillover effects within and across countries. Some of these conference suggestions were implemented as additional analyses after the conference. However, due to their ambitious scope, further work in these areas may be warranted.

That the changes summarized above were not incorporated may reflect problems in the implementation of REP, rather than being inherent to its design. Follow-up investment was expected and even encouraged in the CFP, as long as their unplanned nature was transparently reported. Most of the conference articles did include unplanned follow-up analyses, some of which are extensive. However, editors cautioned authors that the results of unplanned analyses and reinterpretation not dominate the abstract, introduction or conclusion, which would undermine much of the purpose of registration. This policy may have made authors reluctant to invest in unplanned tests and interpretations, by encouraging them to see them as inconsistent with the spirit of registration. We address these possibilities in the next section by analyzing responses to a survey of conference participants.

5. *Conference Survey*

To validate and explain our assessments of the conference papers and to identify opportunities for improving REP's design and implementation, we conducted a survey of all conference authors (of both accepted and rejected proposals), reviewers, and attendees. In this section, we describe the survey and our analysis of responses.

5.1 SURVEY ADMINISTRATION

In November and December of 2017, we sent emails to all conference authors, reviewers, and attendees. As reported in table 4, we received a total of 78 responses from 312 potential respondents, for a response rate of 25%. Authors of accepted proposals were most likely to respond (16 of 23, 70%), followed by conference attendees (47 of 142, 33%), reviewers (11 of 42, 26%), and authors of rejected proposals (16 of 159, 10%). Note that these categories sum to greater than 100%, because many of those contacted played multiple roles. Conference survey questions are reproduced in table 5.

We coded the responses to identify issues that are relevant to the stated goals of the conference and the challenges of achieving them. Two authors independently coded responses to the survey to identify common themes. While coding individual responses involved a high level of subjectivity on the part of each author, there was strong agreement between authors on the themes identified. We present these themes along with responses that articulate them well, but encourage readers to view the complete set of responses made publicly available in the online appendix. We structure our

TABLE 4
Conference Survey Respondents

Conference Contribution	Potential Respondents	Actual Respondents	Response Rate
Author of an accepted proposal	23	16	70%
Author of a proposal that was not accepted	159	16	10%
Reviewer	42	11	26%
Conference attendee	142	47	33%
Total	312	78	25%

This table summarizes the number of potential respondents and actual respondents to our Conference Survey. The survey was sent to 312 individuals that were authors of accepted conference papers, authors of rejected conference papers, reviewers of submitted papers (regardless of whether the submissions were accepted or rejected), and attendees at the conference. The total of percentages in the last column adds up to greater than 100% because many respondents played multiple roles (e.g., reviewers and also conference attendees).

analysis to reflect the chronology of the editorial process: proposal development; review and revision; data gathering, analysis, and writing the reports; conference discussion; and suggestions for improvement.

5.2 PROPOSAL DEVELOPMENT

Of the 16 responding authors who submitted proposals that were ultimately accepted, a majority emphasized the high level of effort that went into developing and implementing a proposal. Several found it hard to propose a simple, definitive study while still being ambitious and having tension in their hypotheses, and many expressed the difficulty of anticipating all contingencies.

> C2: I believe the primary challenge was to not only think about, but to actually write down all of the possible outcomes of our data collection efforts and the follow-up analyses we would want to conduct under each possibility. However, in being forced to write things down, we ended up considering even more possibilities and analyses that we may not have otherwise, ex ante, which I believe was a huge benefit as we refined our research instrument prior to collecting data.

Authors of rejected proposals were not always sure what editors and reviewers were looking for.

> C26: I think that we did not have a well-defined idea of what the editors and reviewers were looking for, so we struggled to determine how much development of the project was too much versus too little versus just right. One of the things that the reviewers criticized us for was being "too developed." I think that we also struggled with designing a project that was appropriately ambitious. Again, this might have been related to the fact that we didn't know what the editors were expecting.

Fifteen of the 16 authors of rejected proposals indicated that they were pursuing their proposal. This high rate of continuation might indicate that REP did little to encourage up-front investment, or that authors felt their

TABLE 5
Questions from Conference Survey

Respondents	Question
Authors of accepted proposals	What challenges did you face in writing a proposal for a Registered Report that you do not face in writing a traditional manuscript?
	What challenges did you face in writing the final Registered Report that you do not face in writing a traditional manuscript?
	What would have made your experience as an author easier, or would have resulted in a better final report?
Authors of proposals that were not accepted	What challenges did you face in writing a proposal for a Registered Report that you do not face in writing a traditional manuscript?
	Have you pursued your project even though the proposal was not approved for the conference?
	If not, why not?
	If so, how did the ultimate project differ from the last proposal you submitted? Did you invest as much in the project as you proposed?
Reviewers	How would you compare your experience reviewing a Registered Report proposal to reviewing a typical manuscript?
	What would have made writing your review easier and/or more effective?
Conference attendees	How would you compare reading a Registered Report to a typical article published in a top accounting journal?
	What are your main takeaways from the conference?
	What advice would you give to the *Journal of Accounting Research* if they continue to offer Registered Reports as an additional path to publication?
All respondents	What advice would you give to those submitting proposals in the future?
	What advice would you give to reviewers and editors managing the process for registered reports?

This table summarizes, by type of respondent, the questions that were asked in our Conference Survey. The survey was sent to 312 individuals that were authors of accepted conference papers, authors of rejected conference papers, reviewers of submitted papers (regardless of whether the submissions were accepted or rejected), and attendees at the conference. All questions were open ended, and many respondents played multiple roles (e.g., reviewers and also conference attendees).

studies were worth pursuing, even in the absence of an in-principle acceptance, once they had made the up-front investment in developing a proposal. We see the latter as more likely, given that five respondents indicated that that they were pursuing a less ambitious version of their proposed study, and none indicated they were pursuing a more ambitious version.

5.3 REVIEW AND REVISION

5.3.1. Challenges for Authors in the Review Process. Several authors of rejected proposals described difficulty in conveying their level of investment to reviewers.

> C26: [. . .]Perhaps, in retrospect, the reviewers did not sufficiently appreciate the aggressiveness of our proposal. That is, it is easy to say, "these authors will pursue this no matter what" if you haven't thought through all the practical issues. So, I guess that I would advise authors to be more explicit about all of the challenges they will face. Authors could have an entire section of the proposal dedicated to the risks/difficulties and the ways to handle these issues. We had these sorts of things sprinkled throughout [our] proposal but it probably needed to be more "front and center."

Some authors were not sure how much work they could do up front.

> C28: In a normal manuscript, we could actually hand collect the data and screen out issues that come up along the way, or convince a reader that we did so. In a Registered Report, it was more difficult to convince the reviewer that the hand collection could be successful. We did collect a pilot sample, to show that collection was feasible, but it was difficult to strike the balance between collecting too much data and maintaining the integrity of the Registered Report. The reviewer and editor ultimately concluded that it was difficult to know whether the project would be successful, and decided to reject it.

5.3.2. Limited Feedback. Many authors had strong praise for the input they received from the editorial process, but that was offset by the difficulty in getting feedback from other colleagues through private discussions and public presentations.

> C21: A registered report proposal requires the authors, referee, and editors to carry most of the weight of identifying the theory, hypotheses, and model specifications. A more traditional manuscript spends time as a working paper. The authors can present it at many schools in the normal state for those presentations. [. . .] This allows peers to guide the manuscript's development. By contrast, a registered report proposal spends time in the review phase without results and, therefore, perhaps too undeveloped to present externally.

Like many authors, several reviewers expressed concern that papers did not receive the usual input from colleagues, which caused uneasiness with approving a proposal.

> C59: Separately, it seemed that the lack of other eyes on the paper (e.g., workshops, conference presentations, etc.) made the authors and editor over-reliant on one reviewer. No matter how smart one reviewer is, s/he can't match the wisdom of the crowd (especially one as high quality as the JAR audience).

5.3.3. Evaluating Studies Without Results. In describing the challenges of evaluating proposals under REP, reviewers appeared to share many of the concerns expressed by authors. Many reviewers mentioned the difficulty of anticipating contingencies, with several mentioning more explicitly the difficulty of evaluating a study without seeing the results.

> C60: The proposal was at an earlier stage, with many choices still to be made. I had to try to think about all the possible issues or reasonable approaches and decide what made the most sense. At times, any of a number of approaches would be fine, or sometimes, it's hard to know which makes the most sense without seeing initial data/descriptives and understanding the structure of the data more fully.

Some reviewers described the absence of results more positively, noting some of the pitfalls associated with the traditional review process.

> C68: The process was different in that reviewers needed to assess whether the proposal was sufficiently developed to lead to reasonably convincing hypotheses before any data were obtained. I believe this is a very useful exercise because after the data are available, hindsight bias leads authors, reviewers, and editors, to assess the quality of the hypotheses at least in part based on whether the results support the hypotheses.

5.3.4. Increased Role of Reviewers. Several reviewers mentioned that they appreciated the ability to provide suggestions before data were gathered and analyses were run, with one reviewer suggesting a potentially larger role for reviewers in the process.

> C63: There is a lot of merit in the idea where authors have to pre-commit to propose a series of analyses and that they are not allowed to extend their analyses. However, once these analyses [are completed] I believe that the reviewer could suggest that the authors conduct additional analyses. In that case there is little risk that authors influence the likelihood of them finding [results]. I found it difficult to accept that that task should be left to a next study. In a way I understand this as the reviewer is involved in the design of the study. But maybe it would then be better to seek alternative reviewers to evaluate the full paper.

In describing this larger role, several reviewers noted an increased amount of effort and expertise required under REP.

> C6: Being an expert and up to date on the topic would help greatly given the fact that you are helping set a "contract" for the authors to fulfill. Without knowledge on the area, it is difficult to do this effectively. I'm not sure how to reduce the investment required by a "good" referee. They are essentially required to become an author and think through the design and potential outcomes and responses without much reward (other than the thanks of the editor and anonymous thanks from the authors).

5.4 DATA GATHERING, ANALYSIS, AND WRITING THE REPORTS

5.4.1. Lack of Flexibility. Echoing earlier concerns about the difficulty of anticipating challenges, authors of accepted submissions struggled to respond to surprises that arose during the research process, speaking directly to the value of the follow-up investments that are encouraged more by TEP than REP.

> C3: We were not sure whether we could go beyond what was in our proposal. As we executed on our proposal, we learnt things. We would have liked to make midcourse corrections to our research design but we were hemmed in by our proposal.

> C22: Not knowing anything about the data was surprisingly difficult. We learned that some of the choices we make in our other research is driven by observed distributional properties or correlations. We were bound by certain choices that, in hindsight, we wished we hadn't made.

5.4.2. Writing a Focused Manuscript. Authors described several challenges associated with writing their final reports. Several authors indicated that writing a focused manuscript was more difficult under REP.

> C2: In traditional manuscripts, telling the most fluent and convincing story seems to sometimes require dropping original or adding new analyses and to sometimes require re-writing the motivation or refining the hypothesis development. In writing the final registered report, we were limited in our ability to make adjustments. I don't necessarily see this as a bad thing . . . but it was certainly a challenge[.]

> C9: The registered report process encouraged extensive collection of data and documentation of data generating processes. One downside of this was that the end manuscript was quite lengthy compared to a normal manuscript and throughout the process it was unclear how much editing was allowed after the initial acceptance.

5.4.3. Interpreting Null Results. Several authors expressed that a lack of experience made it difficult to interpret and present null results.

> C23: We had some null results in the paper, and once we ran the data we firmly believed there was little there. So we had the unusual experience of trying to argue "no no, really there's nothing here!" Not many times that you have to try to argue that side as an author.

> C21: [A] registered report is more likely than a traditional manuscript to require the authors to explain null results. This may require the authors to learn techniques for that type of reporting, such as Bayesian analysis. It likely also requires the authors to consider an array of theoretical and empirical explanations for null results.

5.5 CONFERENCE DISCUSSION

5.5.1. High Trust in Registered Reports. Overall, responses from conference attendees largely align with the conclusions drawn from our analyses of the conference papers. Starting with the most common observations, many attendees express a greater degree of trust in the registered reports than in papers published under TEP, typically emphasizing the transparency of the process, particularly regarding hypotheses and analyses.

> C83: I enjoy reading the initial idea and research plan, because I know that I am truly reading an outcome of the "clean" scientific method, before the design (and even research question) become influenced by the results.

> Stated differently, I like understanding the authors' thought process and creative thinking, and I feel like a Registered Report gives that to me in a purer sense.

5.5.2. Understandability. As with authors and reviewers, attendees' views of transparency are accompanied by complaints that the reports were not easy to understand.

> C60: The Registered Report is longer and more meandering than the typical article. There is less ability to select and interpret the most important information. However, there is less concern that there was manipulation of the results, which is helpful, and more transparency overall.

5.5.3. Difficulty Interpreting Null Results. Many attendees noted that the results appeared weaker than those typically published under the TEP, with several noting that they were difficult to interpret.

> C117: I was struck by the large number of [no-results] or insignificant results. The incidence was much higher than we normally see. The interpretation of this observation is not obvious. It could be that by giving more discretion to authors in the normal process, they can tease out the results. That is, the difference could reflect learning from the data and in the process. An alternative and less favorable interpretation is that this illustrates the amount of p-hacking that is normally going on.

> C100: I did not find the abundance of "null results" surprising. It could have been discovered from one's own experience. Research is an iterative process and it involves learning. I am not sure if there is anything useful that we discover in the research process by shutting down the learning channel; especially with the research questions that are very novel and we do not know much about.

5.6 SUGGESTIONS FOR IMPROVEMENT

The concerns above, and the following suggestions for improvement, reflect "a mixture of optimism in the format as [a]way of tackling bias and liberating researchers from the pressure to massage results, to concerns about it potentially stifling creativity in design and analysis" [C78].

5.6.1. Advice for Those Submitting Proposals in the Future. The most frequently given advice for those submitting proposals in the future was to obtain more external feedback prior to submitting the proposal.

> C9: [M]ore input on the proposal (between being accepted and collecting/analyzing data) would have been extremely helpful. Along the way of completing the project, it became less clear whether certain decisions that were made in the proposal were the "right" decisions. Earlier feedback, when there was still time to make minor changes, could have alleviated some of these concerns.

Several respondents, four of whom were accepted authors, stressed the importance of thoroughly considering possible issues and outcomes of data collection, analysis, and results.

> C15: Make as few assumptions as possible regarding how the data (and findings) would look. Think of the contingencies with data collection and data structure, and plan accordingly (this also has an impact of how you form the hypotheses—you often form the hypotheses with implicit assumptions of how the data would look like).

On a related note, submitters were also encouraged to clearly state predictions for all planned analyses in submitted proposals.

> C64: More generally, I think [it's] very important to clearly specify what analyses will be performed and how those analyses will be interpreted. I see absolutely no problem with preparing additional analyses that are conditioned on a preliminary view of the data—just have a section of the paper that clearly labels those analyses as such. [I] think this is a major step forward in academic publishing.

5.6.2. Advice for Editors and Reviewers of Registered Reports. Echoing the advice to submitters, respondents most frequently suggested that registered reports would benefit from additional feedback, both from reviewers and from external sources.

> C59: Encourage authors to get input from other researchers and experts MUCH earlier in the process, especially before the data collection, data cleaning and research design get locked down through registration. Use two or even three reviewers instead of one to get more expert eyes if you stick to the current process, especially for the pre-registration stage.

Respondents commonly expressed that reviewing a proposal for a registered report should entail a greater level of scrutiny, as more careful vetting is required than under the TEP.

> C78: For reviewers my main advice is to pay very close attention the methods at Stage 1, and the review criteria, because you will not be able to relitigate the study procedures or preregistered analyses at Stage 2 after results are in.

Another common recommendation was to be more clear about the criteria by which proposals will be evaluated and maintain these expectations (rather than "moving the goalposts").

> C24: One of the things that I found somewhat comical about our reviews was that one reviewer said, "This is so obvious. There is no way that these authors WON'T get results." The other reviewer said, "This is so NOT obvious. There is no way that the authors WILL get results." I think this "bet" about whether we would get results or not was part of what factored into the reviewers' recommendations despite the fact that the guidelines for reviewers did not mention the likelihood of results as a criterion for evaluating proposals. Perhaps editors can be more explicit about this in the future?

> C67: Be clear about the primary reason for soliciting registered reports. If it is to incentivize researchers to chase big, bold ideas, let the authors and

> reviewers know that. If it is intended to mitigate purposefully biased intervention in the analysis and interpretation phases of a study, then be clear about that. This information will allow reviewers to understand the types of manuscripts that would be good candidates for the registered report program.

These comments lead us to conclude that editors should think carefully about the forms of up-front investment they wish to encourage under REP, perhaps focusing more on investments in rigor that will clarify results in settings that are already well understood, and setting a higher bar for proposals that invest in scope or novelty.

5.6.3. Advice for the Journal of Accounting Research. When providing advice to JAR, 18 recommended that if JAR is to have another conference dedicated to REP, authors should present proposals before data gathering, rather than after editors have offered in-principle acceptance. Most expressed that feedback was less valuable to authors that had already received in-principle acceptance.

> C92: Future conferences on registered reports should focus on reports before registration, not on already completed registered reports that cannot—and should not be—changed.

Several respondents suggested decoupling registered reports from the conference altogether.

> C2: I believe [that] a registered report path to publication would be a phenomenal step forward for the Journal of Accounting Research. I think the path would be even more successful if it was decoupled from the conference, as then there is no time pressure to get proposals approved, and no specific number of proposals that need to be approved by a pre-specified date. This decoupling should allow for only the best proposals to be guaranteed publication, ex ante.

Some attendees argued that investment is not an appropriate goal for REP, or that standards should be extremely high for investment-oriented projects.

> C83: As I understand it, two key "justifications" for pursuing a Registered Report study are either costly data collection, or trepidation that there will be bias against a study in the review/publication process (i.e., someone may not like a particular result). I would suggest that JAR hold authors to a high standard on these qualifying dimensions, particularly the "costly data" dimension.

Other attendees questioned whether REP is appropriate for archival research and methods that are not controlled enough to limit exploratory data analysis and revision.

> C79: Registered reports do not work very well for empirical accounting research that tend to not use strictly controlled research designs. One of the benefits of the registered report as I understand it is that you can make

> a tight design, execute and experiment, and accept the results (whatever they may be) because to extend anything in the study would require another experiment. Accounting research is simply not done that way as we can almost extend our analyses indefinitely! While this has its own problems, the strictness of a registered report does not seem to be the way to go. It was a really interesting "test" though.

To all of these voices, we would like to add some advice of our own. The CFP explicitly recommended that authors conduct power analyses, consider analyses (like robust regression) that are designed to anticipate contingencies, and rely on Bayesian methods to articulate prior beliefs (see sections 2-9 and 2-10 in the appendix). Only a handful of proposals did so. Almost no survey respondents referred to these statistical practices, or to related concerns that *p*-values are poorly suited to evaluating contributions (e.g., Gelman [2013]). We suspect that improved statistical practices would help REP achieve its goals more effectively.

6. *Survey on Author Discretion Under TEP*

The conference papers and survey responses reported in sections 4 and 5 indicate that REP improves reproducibility by limiting authors' discretion to engage in the questionable practices and selection biases that arise under TEP. However, these limits to discretion come at the cost of less follow-up investment. Thus, those who are considering adopting REP must understand how discretion is actually used by authors under TEP and how readers interpret authors' claims. How much do authors use discretion to make useful follow-up investments, rather than overstating their results? How much do readers place *p*-values in context, rather than expecting individual *p*-values to be reproducible? To what extent are questionable practices driven by authors rather than editors? How reproducible are results published under TEP? We address these questions by surveying accounting researchers who have published under TEP.

Our survey approach complements that taken by Baker [2016], who solicited answers to many questions similar to ours, from over 1,500 scientists across a range of disciplines. We complement Baker by focusing on the single discipline of accounting, which allows us to be more specific in the questions we pose, and in how we interpret answers to free-text responses. We also complement Baker (and many other studies of reproducibility) by posing questions about how questionable practices might actually improve the quality of published research, rather than assuming their effect can only be damaging.

6.1 SURVEY ADMINISTRATION

We used the Web of Science database to identify every author who published between January of 2007 and August of 2017 in six highly regarded accounting journals that publish peer-reviewed papers under similar versions of TEP: *Accounting, Organizations and Society*; *The Accounting Review*;

TABLE 6
Researcher Discretion Survey Respondent Demographics

Panel A: Primary methodology of discretion survey respondents

Methodology	Number of Respondents	Percentage of Total Respondents	BNS (2016)
Preexisting data archives (CRSP, Compustat, IRS, BLS, etc.)	180	60%	63%
Laboratory experiments	66	22%	11%
Field studies/Field work	25	8%	10%
Surveys	7	2%	
Field experiments	12	4%	0%
Other	9	3%	16%
Total	299	100%	0%

Panel B: Rank of researcher discretion survey respondents

Rank	Number of Respondents	Percentage of Total Respondents
Emeritus Professor	5	2%
Full Professor	126	42%
Associate Professor	81	27%
Assistant Professor	82	27%
Doctoral Student	2	1%
Other	2	1%
Not Provided	1	0%
Total	299	100%

This table summarizes demographic information for the 299 participants in our Researcher Discretion Survey. Panel A summarizes the number of respondents, by their reported methodology, and compares the percentage breakdown across methodologies to the percentages summarized in Bloomfield, Nelson, and Soltes [2016] related to the methodologies used in published papers. Bloomfield, Nelson, and Soltes [2016] compile the percentage of papers, by methodology, that are published between 2003 and 2013 in the *Journal of Accounting Research*, *The Accounting Review*, the *Journal of Accounting and Economics*, and *Accounting, Organizations and Society*. Panel B presents the number of respondents, by academic rank, as well as the breakdown in percentages by rank.

Contemporary Accounting Research; *Journal of Accounting and Economics*; *Journal of Accounting Research*; and *Review of Accounting Studies*.[11]

Over the course of September–November of 2017, we sent emails to 2,650 authors, of which about 210 were returned as undeliverable. We received a total of 299 responses, for a response rate of 12%. We began with two demographic questions that allowed us to assess respondents' primary research method and career stage (see table 6). About 60% of our responses came from authors who primarily conduct archival research, compared to 22% conducting laboratory experiments and the remainder using other

[11] Contact information was first extracted from Web of Science, which had an email address for approximately half of our target audience. To the extent possible, we collected remaining email addresses manually via publicly available Web sites. However, we were not able to collect email addresses for all possible respondents. Further, some email messages were not delivered successfully (e.g., due to university email filters, out-of-office messages, changes in authors' affiliations, etc.).

methods. These percentages are roughly consistent with publication rates of these methods in the journals examined by Bloomfield, Nelson, and Soltes [2016]. Almost all authors are tenure-track faculty, including 27% assistant professors, 27% associate professors, and 42% full professors. These proportions are similar to those reported by Boyle, Carpenter, and Hermanson [2015] (27%, 32%, 41%), further suggesting that our respondents are roughly representative of the overall population. We did not ask questions about topical interests (e.g., financial, managerial, tax), as we did not expect those to affect views about author discretion and wanted to protect respondents' anonymity as much as possible.

6.2 FORMS OF DISCRETION

We first summarize respondents' views about six specific forms of discretion, labeled A–F in table 7, panel A. The first four categories pertain primarily to statistical analysis and data gathering: researchers can choose (A) which measures and analyses to report, (B) whether or not to report a sample or subsample, (C) which observations to exclude from primary analyses, and (D) to gather additional data after observing results. The last two categories pertain to the reporting of predictions and theories: (E) researchers can change their predictions and/or underlying theory, and (F) choose not to report or highlight hypotheses.

For each form of discretion, we asked questions 1 through 6 (table 7, panel B) to assess on a 5-point scale how much and how often respondents believe discretion causes reported results to improve quality or be overstated. For the categories pertaining to statistical analysis and data gathering, we characterized improvements in quality as making articles more informative; for hypotheses, as making articles more focused; and for predictions and theories, as making articles more accurate. Each category was accompanied by a free-text question: "For what other reasons do researchers use [this form of discretion], and what other effects does it have on published research using your primary method?" The two categories of discretion were ordered randomly, and the ordering of questions within each block was balanced by having half in one order and half in a reverse order.[12]

In table 8, we show the six forms of discretion, ranked based on how much respondents indicate that they cause results to be more informative, accurate, or focused (panel A) versus overstated (panel B).

As with our Conference survey, two authors independently coded free responses to the survey to identify common themes. While coding individual responses involved a high level of subjectivity on the part of each author, there was strong agreement between authors on the themes identified. In the following sections, we present these themes with references to representative responses, but encourage readers to view the complete set of responses made publicly available in the online appendix. Throughout this

[12] Responses do not vary by the ordering of the instrument.

TABLE 7
Questions from Researcher Discretion Survey

Panel A: Forms of discretion

ID	Form of Discretion	Examples Provided
A	Researchers can choose which measures and analyses to report and highlight.	For example, a researcher might gather many dependent, independent, and control measures, or compute *p*-values using both parametric and nonparametric techniques, but only report or highlight some of them. Researchers can use this discretion to make reported results more informative, but can also use it to overstate the strength of their results.
B	Researchers can choose not to report a sample or subsamples that were gathered as part of a published article.	For example, researchers may choose not to report subsamples of observations with a particular characteristic, such as firms from a particular industry or respondents with a particular level of experience. Researchers can use this discretion to make reported results more informative, but can also use it to overstate the strength of their results.
C	Researchers can choose which observations to exclude from primary analyses.	For example, a researcher might remove observations that fail manipulation checks or other tests for validity, that are highly influential (outliers), or that are otherwise unusual in some way. Researchers can use this discretion to make reported results more informative, but can also use it to overstate the strength of their results.
D	Researchers can choose to gather additional observations after observing results.	For example, researchers might gather more observations or expand their sample (using exactly the same protocol) because they believe they do not have enough data to draw clear conclusions. Researchers can use this discretion to provide more informative tests of their theories, but can also use it to overstate the strength of their results.
E	Researchers can change their predictions and/or underlying theory after observing results.	For example, researchers may change their reasoning for predicting the sign of an effect, when it should be strongest, or what theories justify the predictions. Researchers can use this discretion to improve the accuracy of their theoretical development, but they can also use it to overstate the extent to which they are testing prespecified hypotheses.
F	Researchers can choose not to report or highlight hypotheses.	For example, researchers may choose not to report predictions after observing results that were weak, surprising, or confusing. Researchers can use this discretion to make their write-ups more focused, but can also use it to overstate the extent to which they are testing prespecified hypotheses.

(*Continued*)

TABLE 7—*Continued*

Panel B: Questions on specific forms of discretion

For each of the six forms of discretion listed in this panel, participants were asked:

1) In general, *how often* does this form of discretion cause results reported in published research using your primary method to be more *informative* (*A–D*)/*accurate* (*E*)/*focused* (*F*)?
2) In general, *how often* does this form of discretion cause results reported in published research using your primary method to be more *overstated* (*A–F*)?
3) When researchers use [this form of discretion], *how much* does it increase the following qualities of results reported in published research using your primary method? (*informative* (*A–D*)/*accurate* (*E*)/*focused* (*F*))
4) When researchers use [this form of discretion], *how much* does it increase the following qualities of results reported in published research using your primary method? (*overstated* (*A–F*))
5) For what other reasons do people use this form of discretion, and what other effects does it have on published research using your primary method?

Panel C: General scale questions

6) For papers using your primary method, how do you believe researcher discretion ultimately affects the quality of the average individual published paper in accounting?
7) In general, how often do *you personally* exercise researcher discretion to *increase research quality*?
8) In general, how often do *you personally* exercise researcher discretion to *satisfy reviewers/editors*?
9) In general, when *others using your method* exercise researcher discretion, how often is the motivation for the discretion to *increase research quality*?
10) In general, when *others using your method* exercise researcher discretion, how often is the motivation for the discretion to *satisfy reviewers/editors*?

Panel D: General free response questions

11) Please think about a time when you had first-hand experience with researchers' use of discretion that improved or reduced research quality. For reference, these are situations where researchers may choose:
    - to select which measures and analyses to report and highlight.
    - to remove selected observations.
    - to gather additional data after observing results.
    - to not report samples or subsamples that did not "work."
    - to not report hypotheses that did not "work."
    - to change predictions and/or underlying theory after observing results.

It is extremely valuable to us to know about the first-hand experiences of researchers in these settings. We encourage you to share your experience with us in the space provided below. You may have first-hand experience as an author, reviewer, or editor. For each story you share, please indicate which role you played.

12) Have you or one of your PhD students failed to replicate a study published by someone else? If so, why do you think the replication failed, and how did you respond? Please share any thoughts you have on this matter.
13) What advice would you provide to editors and reviewers that would help them improve authors' use of their discretion in reporting empirical research?
14) What other forms of discretion are used by researchers, and what important effects do they have?

(*Continued*)

TABLE 7—*Continued*

Panel D: General free response questions
15) Besides the motivations noted here, what other motivations for use of researchers' discretion have you or your peers personally experienced?
16) Please use this space to share any other thoughts that authors or readers of this study might find useful.

This table presents the questions that were asked of participants in our Researcher Discretion Survey. We describe six forms of discretion to participants, which relate to either gathering data and analysis (four forms of discretion) or theory and exposition (two forms of discretion). Panel A describes each of the six types of discretion, and presents the examples that were given to respondents. Panel B presents the questions that were asked for *each* of the six forms of discretion. For Forms of Discretion A through D (from panel A), questions 1 and 3 asked about the effect of discretion on informativeness. For Form of Discretion E (F), questions 1 and 3 asked about the effect of discretion on research focus (accuracy). Questions 1 and 2 were asked on 5-point scales with end points of 1 = "Very Rarely" to 5 = "Very Frequently." Questions 3 and 4 were asked on 5-point scales with end points of = 1 = "Not at All" to 5 = "A Lot." Question 5 was asked as an open-ended free response question. Panel C presents more general questions that were presented to participants. Question 6 was asked on an 11-point scale with end points of 0 = "Greatly Reduces Quality" to 10 = "Greatly Improves Quality." Questions 7–10 were asked on 5-point scales with end points of 1 = "Very Infrequently" to 5 = "Very Frequently." Panel D presents the general free-response questions that were presented to participants at the end of the survey.

section, we refer to responses that articulate themes particularly well, identifying them with a letter to identify the form of discretion being discussed (see table 7 for the six forms) and a number to identify an individual respondent. For example, [A321] would refer to a comment on "Choosing Analyses to Highlight and Report" as a form of discretion (A in table 7), and respondent #321 on the survey.

6.2.1. Data and Analysis.

6.2.1.1. Choosing Analyses to Highlight and Report. TEP allows authors the discretion to choose which measures and analyses to highlight and report in a paper. Overall, authors believe that this form of discretion tends to overstate results without increasing informativeness (see table 8). Coding the free responses confirms this view, with many respondents suggesting it is a serious problem [A44]. Several responses indicate that this is a form of *p*-hacking, or cherry-picking in order to present stronger results [A95]. Positive responses generally emphasize that this form of discretion can make papers more focused [A145].

6.2.1.2. Excluding Subsamples. TEP allows authors discretion to exclude entire subsamples of data, such as industries and time periods in archival work, or participant groups or experimental sessions in nonarchival studies. Our scale responses suggest that respondents view this form of discretion as having limited impact on research quality, overstating results and increasing informativeness less than other methods (see table 8). Coding of free responses suggests that authors hold a mixed-to-favorable view on excluding subsamples. Several respondents see it as a good statistical procedure to address either theoretical or practical problems with measurement [B176], while others note that removing subsamples can help focus a manuscript [B61]. As with other forms of discretion, many stressed

TABLE 8

Researcher Discretion Survey Scale Questions Related to Discretion

Panel A: How much does this cause results to be more informative[I]/accurate[A]/focused[F]?

Form of Discretion	Mean Rating	*N*	Pairwise Comparisons[a]				Percentage of Respondents Who Answered: *Moderately* or *A Lot*	Percentage of Respondents Who Answered: *Not at All* or *Very Little*
Choosing Hypotheses to Highlight and Report [F]	0.31	288	A				47	20
Additional Data Gathering [I]	0.14	287	A	B			40	25
Excluding Unusual Observations [I]	0.03	288		B	C		31	27
Choosing Analyses to Highlight and Report [I]	−0.05	287		B	C		32	33
Changing Predictions [A]	−0.06	289			C		31	32
Excluding Subsamples [I]	−0.37	285				D	18	44

Panel B: How much does this cause results to be more overstated?

Form of Discretion	Mean Rating	*N*	Pairwise Comparisons[a]				Percentage of Respondents Who Answered: *Moderately* or *A Lot*	Percentage of Respondents Who Answered: *Not at All* or *Very Little*
Choosing Analyses to Highlight and Report	0.37	288	A				47	18
Changing Predictions	0.19	290		B			39	24
Excluding Unusual Observations	0.03	287		B	C		33	29
Choosing Hypotheses to Highlight and Report	−0.02	287			C		34	31
Excluding Subsamples	−0.03	285			C		30	29
Additional Data Gathering	−0.54	285				D	13	56

[a]Pairwise comparisons are of mean responses for each form of discretion, where those with overlapping letters are not significantly different from one another.

This table summarizes results of the scale-response questions in our Researcher Discretion Survey regarding the effects of six forms of discretion. Panel A summarizes responses to the questions asking participants how much each form of discretion causes results to be more informative, accurate, or focused (depending on the form of discretion). Panel B summarizes responses to the questions asking participants how much each form of discretion causes results to be overstated. See table 7 for the exact wording of the questions. All responses are elicited on 5-point scales with the following labels: –2 = Not at All, –1 = Very Little, 0 = Somewhat, 1 = Moderately, 2 = A Lot. For each form of discretion, the table shows mean responses to these scale questions, and the percentage of respondents who answered "Moderately" (1), or "A Lot" (2) versus "Very Little" (–1), or "Not at All" (–2). The table also presents pairwise comparisons of mean responses for each form of discretion using Tukey's Honest Significant Difference test, with the Tukey-Kramer adjustment for multiple comparisons. Responses not connected by the same letter are not significantly different from one another at $p \leq 0.05$, two-tailed.

the need for disclosure to allow readers to draw their own conclusions [B152].

6.2.1.3. Excluding Unusual Observations. TEP allows authors discretion to exclude observations based on arguments that they are in some way unusual. Many observations are excluded in archival research because they are said to be "outliers" that do not reflect a reasonable value of the variable or are disproportionately influential on parameter estimates. Experimentalists sometimes exclude observations because participants fail a particular requirement, such as attention checks and manipulation checks. This form of discretion was rated near the top in terms of *both* informativeness and overstatement (see table 8). Coding of free responses shows that many respondents see this form of discretion as a good statistical practice for addressing avoidable error [C123], while others argue that authors are cherry-picking observations to report better results [C98]. One uniting theme was that the exclusion of observations should be transparent, so readers can evaluate for themselves how removing observations may have affected results [C131].

6.2.1.4. Additional Data Gathering. TEP allows authors to gather additional data after they have seen their results. While authors could exploit stopping rules to bias *p*-values downward, respondents believe that it is generally used to make published results more informative and less overstated. As indicated in table 8, discretion to gather additional data is near the top of the list for increasing informativeness, but at the bottom for overstating results. In our coding of free-text responses, we find that many respondents see additional data gathering as primarily a response to reviewers and editors, often suggested because additional data have become available over the course of the project, with a mix of positive and negative views on its effect on research quality [D152]. Quite a few mention using discretion to gather additional data as a good way to increase power [D156], while others are more pessimistic, viewing the use of stopping rules as means of *p*-hacking [D36].

6.2.2. Theory. We focus on two aspects of discretion in how authors report the theoretical interpretation of their empirical results.

6.2.2.1. Changing Predictions. TEP allows authors to alter the predictions of a paper or its underlying theory. Overall, respondents see this type of discretion as one the most prone to overstatement, with little benefit from increased focus (see table 8). Several respondents view this as a form of HARKing, where authors decide on an underlying theory after observing results [E146]. Many respondents believe that this form of discretion is driven by reviewers and editors, with most viewing it negatively, oftentimes as a consequence of publication bias [E90]. Nevertheless, quite a few respondents had a favorable view of this form of discretion, expressing that it is natural (and informative) for authors to revise theory as the setting and underlying processes become more clear [E11, E183].

6.2.2.2. Choosing Hypotheses to Highlight and Report. The final form of discretion we consider deals with changing and omitting hypotheses. Overall, respondents believe that such discretion is used to improve the focus of published manuscripts, and do not see a tendency toward overstating the strength of results (see table 8). These views are more positive than those regarding theory and predictions, suggesting that respondents see the stated hypotheses as expositional devices to communicate effectively, rather than as rigorous claims about what authors intended to test. Free responses support this view. Some respondents suggested that this form of discretion can make a paper more succinct and coherent for readers [F264]. Several respondents mentioned that underlying theory is often unclear or underdeveloped when a project first begins, and that the research process naturally leads to revisions in predictions. Thus, altering the hypotheses in a paper may reflect learning by the authors as research progresses [F111]. Other responses suggest a more negative view of this form of discretion. While a few noted improved understandability, several suggested this occurs because researchers engage in HARKing, or deciding upon their hypotheses after seeing results [F118].

6.3 OVERALL IMPRESSIONS

After asking questions about specific forms of discretion, we ask a number of more general questions, listed in table 7, panel C. First, we ask on an 11-point scale how, on balance, papers using respondents' primary method are affected by researcher discretion. We follow up by asking about how often they and others employ all forms of discretion to improve research quality and to satisfy reviewers and editors. We then ask five free response questions (table 7, panel D) about personal experiences, advice for editors and reviewers to improve authors' use of discretion, and other forms of discretion and motivations for the use of discretion.[13]

Overall, respondents believe that discretion ultimately improves research quality in their area, with a mean rating of 0.437 on an 11-point scale ranging from –5 = "greatly reduces quality" to 5 = "greatly improves quality" (above the midpoint with $p = 0.001$, two-tailed). This overall improvement is driven primarily by nonarchival researchers, with a mean of 0.910 ($p < 0.001$, two-tailed), with archival researchers not indicating any significant improvement (mean of 0.167, $p = 0.312$, two-tailed).[14] This is one of the

[13] Because our six forms of discretion may not be comprehensive, we also asked about other forms of discretion and their effects, as well as what types of other motivations for discretion the respondents had personally experienced. While respondents provided many additional forms of discretion and motivations for using discretion, as no themes emerged, we do not attempt to summarize them here. However, we encourage readers to view these responses in the online appendix.

[14] Our "nonarchivalists" sample excludes nine participants who classify themselves as using "other" methodologies in our survey. For purposes of anonymity, we do not request additional detail on methodology when participants choose "other." However, we exclude these from

TABLE 9
Researcher Discretion Survey General Scale Response Questions

Panel A: Survey responses to the question: in general, how often do *you personally* exercise researcher discretion to:

Motivation	Average Rating	*N*	Percentage of Respondents Who Answered: *Frequently* or *Very Frequently*	Percentage of Respondents Who Answered: *Very Rarely* or *Rarely*
Increase research quality?	0.20	291	37	18
Satisfy reviewers/editors?	0.35	293	46	17

Panel B: Survey responses to the question: in general, when *others using your method* exercise researcher discretion, how often is the motivation to:

Motivation	Average Rating	*N*	Percentage of Respondents Who Answered: *Frequently* or *Very Frequently*	Percentage of Respondents Who Answered: *Very Rarely* or *Rarely*
Increase research quality?	0.30	288	40	12
Satisfy reviewers/editors?	0.71	288	62	7

This table summarizes results of the scale-response questions in our Researcher Discretion Survey related to the more general effects of discretion. Panel A (panel B) summarizes responses to the questions asking participants how often *they personally* (*others using their method*) exercise researcher discretion to increase research quality or to satisfy reviewers/editors. All responses are elicited on 5-point scales with the following labels: –2 = Very Rarely, –1 = Rarely, 0 = Sometimes, 1 = Frequently, 2 = Very Frequently. For each form of discretion, the table shows mean responses to these scale questions, and the percentage of respondents who answered "Frequently" (1), or "Very Frequently" (2) versus "Rarely" (–1), or "Very Rarely" (–2).

few significant differences we observe between archivalists and nonarchivalists ($p = 0.006$, two-tailed), which may reflect that these two groups of researchers face different challenges or have adopted different norms (or both).

Respondents see discretion as common and driven more by requests from reviewers and editors than by authors (table 9). Thirty-seven percent say that they frequently or very frequently use discretion to increase research quality, compared to 18% who say they rarely or very rarely do. 46% of respondents say that they frequently or very frequently use discretion to satisfy reviewers or editors, compared to only 17% who say this is rare or very rare. Responses about why *others* exercise researcher discretion indicate slightly more frequent use of discretion both to increase research quality (40% frequently or very frequently vs. 12% rarely or very rarely) and to satisfy reviewers or editors (62% frequently or very frequently vs. 7% rarely or very rarely).

6.3.1. First-Hand Accounts of Discretion. We asked respondents to share first-hand accounts with the use of discretion. We begin our analysis by

our scaled responses because the forms of discretion we analyze do apply most appropriately to hypothesis-testing studies, which free-text responses suggest is not pursued by respondents indicating focus on "other" methods.

identifying whether respondents portray discretion as improving or worsening quality, and whether they portray themselves as author, reviewer, or editor. Overall, discretion is portrayed as improving quality in 40 accounts and worsening quality in 53. Accounts written as authors were slightly favorable (36 improving quality and 31 worsening), while accounts written as editors and reviewers were sharply unfavorable (8 improving quality and 21 worsening quality).

We also find that respondents take credit for good uses of discretion but attribute blame to others for bad ones. Authors were more likely to think they drove a good use of discretion (26 of 51) than a bad use (18 of 51), but more likely to think editors and reviewers drove a bad use of discretion (23 of 44) than a good use (17 of 44). Editors and reviewers were more likely to think that an author drove a bad use of discretion (19 of 23) than a good use (5 of 23), and were roughly equally likely to think that they or another editor or reviewer drove a good use of discretion (9 of 15) or a bad use (8 of 15). This deflection of blame indicates that discretion poses issues that are charged and complex enough to support motivated reasoning.

The charged and complex nature of discretion is illustrated well by many of the first-hand accounts. Many accounts are depicted as reducing research quality by methods we would describe as either HARKing or *p*-hacking.

> I have seen most of these situations. I will share one that bothers me to this day. I reported an unsupported hypothesis in a submission (significant in the opposite direction) and shared an alternative theory that could have predicted the results. I had reviewers at two different journals imply that I should modify my theory and consider the hypothesis supported. They were very cautious in their wording and I do believe their intent was to improve the contribution, but the fact that reviewers would suggest this boggled my mind.
>
> Respondent 48, Associate Professor, Laboratory Experiments

> I refereed a paper that selectively chose whether to report one-tail or two-tail tests. The main variables of interest showed one-tailed *p*-values (and curiously were significant at the 10% level, but two-tail tests would not have been significant), whereas less important coefficients reported two-tailed *p*-values.
>
> Respondent 15, Assistant Professor, Field Studies

Other uses of discretion are depicted as improving research quality by gaining a better understanding of the theory or making the presentation in a paper more clear.

> I was serving as a reviewer for a paper at a top journal, and the original manuscript submitted by the authors had found conflicting results relating to the theory they had proposed—in other words, some of the results were consistent with expectations derived from the theory while others were contrary. The other reviewer suggested that the authors consider a different theory that was, frankly, a better fit for the situation and that

> explained the pattern of results very well—far better than the theory proposed by the authors. The question immediately arose as to whether it would be ethical and proper for the authors to rewrite the manuscript with the new theory in place of the old. This was a difficult situation because it was clear the authors had chosen a theory that didn't fit the situation very well, and had they been aware (or had thought of) the alternate theory suggested by the other reviewer, they would have been well advised on an a priori basis to select it instead of the one they went with, but I had concerns about a wholesale replacement of a theory after data had been collected to test a different theory. On the other hand, the instrument used in collecting the data actually constituted a reasonably adequate way to test the alternate theory, except, of course that it wasn't specifically designed to differentiate between the two. I don't recall exactly how the situation was resolved as it was a number of years ago, but my recollection is that the paper was published after some additional data was collected that pointed to the alternate theory.
>
> Respondent 84, Full Professor, Laboratory Experiments

> As an author, I have received feedback from an editor at a Top 3 journal that the economic significance of the results in the paper seemed a little too large to be fully explained by the hypotheses. My co-authors and I were informed by the editor of an additional theoretical reason why the effects sizes could be that large and we were encouraged by the editor to incorporate that additional discussion into the underlying theory in the paper. My co-authors and I agreed that the theory and arguments provided by the editor seemed reasonable. As a result of incorporating this suggestion, we believe the paper is more informative to readers.
>
> Respondent 280, Assistant Professor, Archival

> As a doctoral student, I ran a $2 \times 2 \times 2$ design on one of my studies. The 2×2 of primary interest worked well in one level of the last variable but not at all in the other level. I was advised by my dissertation committee not to report the results for participants in the one level of that factor that "didn't work" because that level of the factor was not theoretically very important and the results would be easier to explain and essentially more informative. As a result, I ended up reporting only the 2×2 of primary interest with participants from the level of the third variable where the 2×2 held up. To this day, I still feel a little uncomfortable about that decision, although I understood the rationale and thought it made sense.
>
> Respondent 85, Full Professor, Laboratory Experiments

We encourage readers to review the complete set of first-hand accounts provided in the online appendix, to fully appreciate the many difficult judgments required to determine whether discretion is being used to improve the quality of a paper or overstate results, and the professional pressures that make such judgments challenging. We also encourage readers to discuss these accounts with doctoral students, colleagues, and advisors; the variety of views in our survey responses suggest that such discussions will reveal substantial disagreement over dilemmas that most of us will face repeatedly throughout our careers.

6.3.2. First-Hand Accounts of Failed Replications. To provide more direct evidence on the reproducibility of accounting research under the TEP, we asked participants whether they or one of their PhD students had failed to replicate a published study, and if so, why they believe the replication was unsuccessful. A total of 113 respondents reported having tried to replicate a published study, with 105 saying that they had failed at least once (eight participants said that they had never failed to replicate a published study). We caution against using these numbers to estimate a proportion of how much accounting research is unreproducible, because we do not know how many replications were attempted by those reporting a failure, and 21 responses not included in our 113 simply answered "no," leaving us unclear on whether they had never attempted a replication or had always been successful. However, we believe 105 first-hand accounts of failed replication suggest that improvements in reproducibility would be a worthy goal for our field.

Not all failures to replicate would need to be addressed by registration. The most frequent theme in responses, expressed by 51 respondents, was that published papers, or the authors themselves, simply do not provide enough detail for readers to implement the procedures used to generate the original claim.[15]

> Yes, an archival study published in [a top accounting journal]. We followed the prior paper exactly but could not replicate the sample or their results for the test variable. We concluded the most likely reason was that there were sample selection or model specification details that were omitted from the published paper.
>
> Respondent 82, Full Professor, Laboratory Experiments

The second-most common theme, expressed by 32 respondents, is that the original results may have reflected some form of questionable practice or selection bias. We were fairly liberal in identifying responses as fitting this theme, because we expect respondents to be diplomatic, especially when they can only rely on informed speculation. We included in this theme any respondent who indicated that they could reproduce methods exactly but generated a null result, found the result was sensitive to questionable choices, or attributed the original positive result to noise or low power.

> My coauthors and I failed to replicate a published archival study by other researchers. We think the replication failed because we used a much larger and more recent sample whereas the published paper used a small sample collected many years before our sample period.
>
> Respondent 40, Associate Professor, Laboratory Experiments

[15] Four additional responses noted that publishers may update their databases retroactively, making it difficult for researchers to access data used in prior analyses.

> I have failed to replicate this form of study. Ultimately, we believe the failure to replicate is due to the initial study being unreplicable. This paper actually presents an abnormally detailed amount of information on the experimental results and we do not believe their findings are due to an inappropriate use of discretion. Rather, we believe they just had a weird day in the lab that gave them an odd result that had a plausible theory backing it.
>
> Respondent 42, Assistant Professor, Laboratory Experiments

> Yes, on one unpublished working paper a co-author and I were not able to replicate a now published, competing paper. I contacted the senior author of the competing paper and explained that we were not able to replicate their findings and I asked for help with our replication. The senior author forwarded our message to the individual running the data on their project. We received a verbal response from the "data person" that inadequately explained the steps taken in their data analyses. But we did not find their explanation helpful. We were never able to replicate their results, and our study was never published.
>
> Respondent 141, Full Professor, Archival

We may overstate the importance of questionable practices and selection bias by defining them so broadly. However, another common theme suggests that selection bias may be quite strong in our field, especially when it comes to failed replications. All three of the respondents quoted above noted that they were unwilling or unable to publish their findings. As another respondent put it, "No one gets published for replication studies" [Respondent 150, Full Professor, Archival]. Such attitudes cause the pool of published results to tilt strongly toward the positive, with little public evidence of the negative results found by others.

6.3.3. Advice to Editors and Reviewers. We next asked participants what advice they would give to editors and reviewers to help them improve authors' use of their discretion in reporting empirical research. One of the main themes in these responses was that reviewers and editors should be more forgiving of imperfect or null results, which would be one way to lower selection bias (and the pressure for *p*-hacking) within TEP.

> Do not fixate on "magical" *p*-values. As Rosnow & Rosenthal, say, "... surely, God loves the .06 nearly as much as the .05." Editors and reviewers are often guilty of over-emphasizing these cutoffs, leading to p-hacking.
>
> Respondent 50, Full Professor, Laboratory Experiments

> Well, this is a long conversation. Nothing will improve until editors stop being obsessed with perfect 'consistency' in every result in a paper. You often read papers that are so 'ridiculous' in that every single result works perfectly, perfectly fits with the rest, every agent immediately processes complex info perfectly, etc. so that the story becomes a simple perfect story. This is not the reality of the world, the reality of data. Acceptance of this imperfection would lead to more transparent studies showing [what] works and what doesn't, studies raising more questions rather than

> providing perfect answers, with the reader making his/her views on what to learn from it. But the current publication/review process forces each paper to be THE conclusive paper on a given topic where every partition works as expected etc. (of course only the partitions shown by the author, etc.).
>
> Respondent 146, Associate Professor, Archival

Consistent with the idea that there should be less emphasis on positive results, respondents encouraged editors and reviewers to be more open-minded as to what constitutes a contribution, to not force authors to deviate from the original intent of the project, and to encourage more replication.

> Try to evaluate the paper for what it is not what you want it to be. Often the main findings the authors present are interesting, even if we do not fully understand what those findings mean. It always takes more studies and different designs to fully understand.
>
> Respondent 61, Full Professor, Laboratory Experiments

Echoing earlier responses, several responses suggested increased transparency in reporting.

> Documentation and transparency are important attributes of accountants and auditors. We, as accounting researchers, should also highly value these. Choices are made all of the time, but we just need to [be] forthcoming about them.
>
> Respondent 43, Assistant Professor, Laboratory Experiments

Many journals are already moving to improve transparency while retaining TEP. For example, the *Journal of Accounting Research* has recently strengthened its requirements such that authors provide detailed information on their programming codes, and "ask whether the information they have provided will allow an informed researcher to understand and replicate the analyses reported in the paper" (JAR [2018]). The American Psychological Association has expanded the detail authors must provide in the body of papers they publish, to help authors reproduce procedures and also interpret *p*-values in context (Appelbaum et al. [2018]). However, some respondents suggest that these changes may not be sufficient, and recommend registration as a way to allow full transparency into how the paper has been revised.

> Actively look for, and carefully assess, the discussions provided by the authors regarding their research choices with respect to the selection of the sample, measures, and analyses. On the other hand, I think it would be very difficult, if not impossible, to identify whether the hypothesis (and the underlying theory) is laid out before the results are observed or it is the other way around. To address this, some sort of two-stage evaluation of a research study (something like what JAR proposed before) has to be implemented.
>
> Respondent 207, Assistant Professor, Archival

7. *Conclusion*

This paper documents lessons we learned from JAR's experiment in REPs based on our reading of the conference papers; a survey eliciting feedback on REP from authors, reviewers, and attendees; and a separate survey eliciting views on the use of discretion under the TEP.

Our conclusions will be familiar to most scholars of accounting, starting with the familiar adage that "no system is perfect."[16] In many settings, principals can reward agents for outputs under their control, or for the outcomes those outputs are intended to create (GASB [2008]). REP rewards authors with publication based on the outputs under their control, such as choosing a good research question, clearly articulating their theory and its predictions, gathering data tailored to their question, and analyzing their data appropriately. REP encourages up-front investment in these outputs but provides few incentives for follow-up investments that allow authors to refine their work after seeing their results. In contrast, TEP encourages follow-up investment, but results in less up-front investment and presents authors with pressure and opportunity to overstate their contributions. REP also reduces selection biases that keep readers from seeing many studies that failed to support authors' predictions, but forces editors to bear the risk that they will publish studies that, once results are known, provide little contribution.

Advocates of REP typically place very high value on the reproducibility of results, particularly of *p*-values taken in isolation. Our surveys show that readers of scholarly journals value many other qualities, much like readers of financial reports. To cast this claim in terms from the Conceptual Framework that accounting standard setters use to guide their deliberations (e.g., FASB [2010]), readers want articles to provide a faithful representation of what authors did and found, which requires articles to be complete, neutral, and free from error. But the Conceptual Framework also recognizes that readers want articles to be relevant, which we would define as having the potential to change readers' beliefs about issues that matter to them. REP encourages faithful representation, making results more likely to be reproducible, but does so by reducing the extent and flexibility of follow-up investment that might enhance relevance. Our survey responses suggest that readers understand this tradeoff. Much like readers of financial reports, they approach published research with a fair bit of sophistication, unwinding predictable biases by interpreting them in the context of what authors report and do not report about their study, and in the context of how the editorial process creates and defends against pressures toward overstatement.

Our analysis also suggests that future implementations of REP could improve the overall quality of final reports without much sacrifice of other

[16] See Bloomfield [2017] for elaboration of why accounting systems are imperfect and numerous examples from financial and managerial reporting contexts.

qualities. Editors could encourage investments in data that improve rigor and power, rather than scope and novelty, and encourage authors to obtain more input on proposals before and throughout the review process. Authors could be held to higher standards for investment in data gathering, piloting, power analyses, and statistical methods, and could be required to conduct more unplanned analyses once planned analyses are complete. Authors could be granted more flexibility to revise their papers to make them more understandable and focused, particularly on matters of exposition, while still being held to strict compliance on planned statistical analyses. Authors might even be allowed more leeway to exercise forms of discretion that are viewed as primarily beneficial, such as expanding sample sizes and excluding unusual observations, provided these are used transparently.

More generally, editors could improve both REP and TEP by identifying studies that are better suited to each process, allowing slightly more discretion under REP and slightly less under TEP, clarifying standards under REP, and demanding more transparency under TEP. However, more widespread adoption of REP will present editors and our scholarly community with some new and difficult questions. Should editors be more willing to publish short papers that flesh out results left on the table under REP? What about replications of papers whose reproducibility was potentially undermined by TEP? How should authors themselves be evaluated for publishing under REP, or publishing the shorter extensions and replications that REP might encourage? How much should such studies "count" as authors are evaluated in hiring, promotion, and other career-related decisions? Should a publication under REP count for less if exemplary upfront investment did not ultimately lead to positive results? Our analysis raises these questions, but much more work will be required to answer them.

APPENDIX

Call for Papers and Details of the Editorial Process

1	**Call for papers**
1-1	To encourage submissions across a broad set of research areas, the 2017 conference is not restricted to a particular method or topic. Authors may undertake any research method, as long as they are proposing to gather new data. Acceptable methods include, but are not limited to:
1-2	• Analyses of newly acquired structured archives; • Analyses of newly acquired or existing unstructured archives (text, images, audio, and events) that require structuring or "hand collection" before being analyzed; • Surveys, laboratory experiments, and laboratory studies; and • Field studies and field experiments.

1-3	Proposals can address any topic likely to be of interest to our readership, even if the topic might otherwise be more appropriately targeted to journals covering other areas of business and social sciences, such as finance, economics, social psychology, organizational behavior, and political science.
2	**Detailed policies provided by JAR**
2-1	Editors determine whether the initial proposal submission is promising enough to be sent to one or more referees for review, and if so, use the review(s) to determine whether the proposal should be rejected, returned to the authors for revision, or approved. Approval letters spell out the conditions under which they will accept the second-stage report for publication. These conditions will always require that authors fulfill their commitments to gather and analyze data as proposed, and never require that the results support any particular conclusion (such as the stated hypotheses). However, editors may also include other conditions to address specifically identifiable concerns about the informativeness of the data or the thoroughness of the additional analyses. To the extent possible, conditions will be crafted to allow authors to guarantee publication simply by living up to commitments under their control.
2-2	REP encourages investment. Without the commitments embodied in REP, authors are effectively creating research "on spec," speculating that the end result will be attractive to an editor. Authors are wise to limit their investment in such circumstances, since they cannot easily predict either editors' tastes or the end result of the study. The acceptance decision may also be far in the future. REP encourages authors to propose studies that are more ambitious (e.g., in the scope of their data gathering, by deviating from conventional tastes) because they can defer the bulk of their investment until after an editor has committed to publish the end result.
2-3	REP enhances reliability. In the traditional editorial process, authors offer their completed research results to editors for evaluation. This leaves authors with both the incentives and the ability to overstate their results by choosing statistical tests that indicate strong support for their predictions, and revising those predictions to make their theory seem more powerful. REP reduces the incentives and ability to pursue such strategies, by making publication of an accepted proposal contingent on whether the author lived up to their commitments to gather data and analyze it appropriately, rather than on the outcome of those results.
2-4	REP accelerates input. An editor's goal is to publish good papers. Editors accomplish this goal partly by making wise decisions about which papers to publish and which to reject, and partly by providing input that helps authors improve their initial submissions. Such input is often very painful in the traditional editorial process, because it comes after the author has made crucial and sometimes irrevocable decisions on how to gather data. REP provides authors with input before they gather data, improving the likelihood that the editors (and authors) can publish good papers.
2-5	Rigor. [S]tudies are more likely to be successful when authors tailor data gathering closely to the theory they are testing and use tight research designs, possibly allowing authors to draw clear causal inferences. Laboratory and field experiments are well suited to this type of ambition, as are analyses of newly hand-collected archives that can be appended to familiar archives, particularly when they allow for strong research designs. Such studies are also more likely to be effective when data gathering provides a large but carefully selected sample.

2-6 Scope. Authors can gather extensive new data that are likely to shed light on a wide variety of related questions. Such studies are more likely to be successful when authors gather new data from settings of substantial interest to many researchers, and that can be used to extend multiple studies. Field studies, surveys, and extensive hand-collection of archives are well suited to this type of ambition. While the authors may be able to provide rigorous tests of one theory, such studies can also make a contribution by providing rich contextual information that allows readers to interpret prior research in new ways and propose new theories to test. Such proposals are more likely to be approved when the authors agree to make the data publicly available.

2-7 Novelty. Authors can address novel theories in novel ways, knowing that they need not undertake the effort of data gathering if their innovations are not to the taste of reviewers and editors. Any method is well suited to this type of ambition. However, novelty makes it more likely that planned hypothesis tests will be subject to unanticipated features of data, such as floor and ceiling effects, skewed distributions, outliers and the like. Such proposals are most likely to be successful if they include pilot data to rule out such concerns, and propose robust methods of analysis that accommodate a variety of distributional assumptions.

2-8 All proposals must spell out planned analyses clearly enough that editors can evaluate whether the authors adhered to them when reported in their stage 2 manuscript. However, good statisticians understand that the optimal method of analysis often depends on features of the realized data, like the skewness of distributions and correlations among variables. Authors are encouraged to propose analyses that handle data contingencies without having to tailor the approach too finely to realizations of the data. For example:

2-9
- Robust regression limits the weight on influential observations based on observed distributions.
- Canonical correlation accounts for multiple comparisons by identifying associations between vectors of dependent and independent variables.
- Support vector machines and other machine learning methods use predetermined algorithms to identify complex nonlinear associations from a relatively small set of assumptions (e.g., tuning parameters).

2-10 Proposals are more likely to be successful if they allow readers to interpret null results as well as positive ones. Thus, authors should carefully think about what we learn from their analysis independent of the specific realization of the data and the results. Authors who propose to use Null Hypothesis Significance Testing (NHST) are encouraged to include a power analysis that calculates the likelihood of rejecting the null hypothesis given the prediction of an expected or reasonable effect size and an estimation of variability in the data. Authors may also find it helpful to replace (or complement) NHST with Bayesian methods of data analysis. Bayesian methods provide two clear advantages for Registered Reports. While NHST can provide evidence only that an effect was or was not reliably observed, Bayesian analysis can indicate how the new data changed the relative likelihood of the null and alternative hypotheses. A Bayes Factor of 3 means that the alternative hypothesis is three times as likely (relative to the null), while a factor of 1/3 means that the null hypothesis is three times as likely as it was before any data were collected. Either result is likely to indicate that the data support a substantial change from prior beliefs. Bayesian analyses also avoid the inferential problems that arise when researchers can observe data before choosing when to stop collecting more. In a well-designed study, additional data points increase the Bayes factor if the alternative hypothesis is true, and reduce it if the null hypothesis is true. Inferences are therefore not affected by stopping decisions.

2-11	Authors are encouraged to consider collecting and reporting pilot data to demonstrate that they are able to gather data and analyze them as proposed. Authors proposing to hand-collect archival or field data are more likely to be successful if they can use pilot data to demonstrate that the data can be obtained, coded and analyzed in the proposed fashion, and that there is enough variation in key measures to allow reasonable power. Authors proposing experiments or surveys are more likely to be successful if they can use pilot data to demonstrate that subjects can understand the stimuli, can (and will) complete the task as proposed, and provide responses that are not heavily encumbered by floor and ceiling effects or other issues that make inference difficult.
2-12	Note that the role of pilot data in REP is not to demonstrate that the study is likely to generate positive results, but instead to demonstrate that the data gathering and the analysis are feasible and that the results are likely to be interpretable, whether or not they are positive. In fact, if the pilot analysis essentially provides the proposed analysis with fewer data and hence forecasts the results, then it would undo the purpose of REP.
3	**Reviewer guidance**
3-1	For all proposals (stage 1 submissions), reviewers will be asked to write a referee report that addresses the following questions:
3-2	• How important, relevant, and innovative is the research question? • How substantial is the investment in data gathering? • How novel is the data gathering exercise? • Are the predictions well grounded? • How clear and detailed are the descriptions of data gathering methods and planned analyses? • How likely is it that the authors will be able to fulfill their commitments to gather and analyze data, within the timeline established for the Conference? • How likely is it that additional analyses will be necessary upon gathering and seeing realizations of the data? • What specific concerns do you have that might make the data or the analysis uninformative, even if the authors live up to their commitments to gather and analyze data? Are there ways to address these concerns by changing the proposal or by imposing specific conditions for publication of the ultimate report?
3-3	For all reports of approved proposals (stage 2 submissions), reviewers will be asked to address the following questions:
3-4	• Are the theory and hypotheses largely consistent with the proposal? • Do the data gathering methods and planned analyses fulfill the authors' commitments? Are all deviations clearly stated? • Does the report fulfill any additional conditions specified in the approval of the proposal? • Are the data gathering and the research design explained clearly? • Are additional (unplanned) analyses appropriate given the realization of the data? • Are stated results and interpretations justified by the actual methods, analyses, data realization, and results?

4	**Post-conference revision guidance to authors**
4-1	• Tier 1 includes the description of proposed data gathering, data coding, hypotheses about statistical associations, and interpretations of statistical results from planned analyses. In the spirit of the REP, this material should be revised sparingly, if at all, particularly when it comes to material that addresses the operationalization of the study's methods and measures, as opposed to theoretical constructs. Revisions to Tier 1 matters must be described in detail in an appendix, and should also be discussed in relevant spots in the text.
4-2	• Tier 2 includes the definition and descriptions of theoretical constructs that are being examined. The conference discussion indicated that there is room for improving the exposition on these matters in many papers, because constructs can be fuzzy in accounting, and reasonable people disagree about the meanings of terms. We do not believe that revisions that help readers understand the paper better violate the spirit of REP, unless of course those revisions are driven primarily by observing the paper's results, in which case you should refrain from them or clearly state that they came after the fact. Revisions to tier 2 matters should be described in text or footnotes when they are substantive. For some papers this might take the form of footnoting the first use of a construct term to indicate "Our approved proposal referred to the construct as "X"; we now refer to it as "Y" for the following reasons . . . ."
4-3	• Tier 3 includes the arguments for why the study is worth doing (its motivation), its broader implications, how it relates to the literature, and other more speculative judgments. Our view is that material that was primarily written to justify why the proposal was worth doing can be revised without violating the spirit of REP; authors should make the best current case to readers about why they should read the paper and how it ties into the literature, in light of conference comments, and final results. However, authors should avoid basing final interpretations and implications too heavily on unplanned analyses; essential to the spirit of REP is that results of planned analyses are given substantially more weight than the results of unplanned analyses. Revisions to tier 3 matters need not be disclosed in the final paper; readers who are interested in the evolution of the manuscript will always be able to refer to the approved proposal, which will be permanently available on the JAR Web site. However, any revisions that are motivated by the results of unplanned analyses should be communicated in a discussion toward the end of the paper. The introduction can refer to these results, but again, they cannot dominate the introduction and the interpretation of your paper.

REFERENCES

ALLEE, K. D.; M. D. DEANGELIS; AND J. R. MOON, JR. "Disclosure "Scriptability." *Journal of Accounting Research* 56 (2018): 363–430.

APPELBAUM, M.; H. COOPER; R. B. KLINE; E. MAYO-WILSON; A. M. NEZU; AND S. M. RAO. "Journal Article Reporting Standards for Quantitative Research in Psychology: The APA Publications and Communications Board Task Force Report." *American Psychologist* 73 (2018): 3–25.

BAKER, M. "Reproducibility Crisis." *Nature* 533 (2016): 452–54.

BEGLEY, C. G., AND L. M. ELLIS. "Drug Development: Raise Standards for Preclinical Cancer Research." *Nature* 483 (2012): 531–33.

BERNARD, D.; N. L. CADE; AND F. HODGE. "Investor Behavior and the Benefits of Direct Stock Ownership." *Journal of Accounting Research* 56 (2018): 431–66.

BLOOMFIELD, R. J. "What Counts and What Gets Counted (2nd edition)." 2017. Available at SSRN: https://ssrn.com/abstract=2899141 or http://doi.org/10.2139/ssrn.2899141.

BLOOMFIELD, R.; M. W. NELSON; AND E. SOLTES. "Gathering Data for Archival, Field, Survey, and Experimental Accounting Research." *Journal of Accounting Research* 54 (2016): 341–95.

BOYLE, D. M.; B. W. CARPENTER; AND D. R. HERMANSON. "The Accounting Faculty Shortage: Causes and Contemporary Solutions." *Accounting Horizons* 29 (2015): 245–64.

CHAMBERS, C. D.; E. FEREDOES; S. D. MUTHUKUMARASWAMY; AND P. ETCHELLS. "Instead of "Playing the Game" It Is Time to Change the Rules: Registered Reports at AIMS Neuroscience and Beyond." *AIMS Neuroscience* 1 (2014): 4–17.

COMMITTEE OF SPONSORING ORGANIZATIONS OF THE TREADWAY COMMISSION (COSO). *Internal Control—Integrated Framework*. Jersey City, NJ: AICPA, 1992.

COMMITTEE OF SPONSORING ORGANIZATIONS OF THE TREADWAY COMMISSION (COSO). "Internal Control—Integrated Framework." 2013. Available at http://www.coso.org/ic.htm

CRESSEY, D. R. *Other People's Money; A Study of the Social Psychology of Embezzlement*. New York, NY: Free Press, 1953.

ERTIMUR, Y.; C. RAWSON; J. L. ROGERS; AND S. L. C. ZECHMAN. "Bridging the Gap: Evidence from Externally Hired CEOs." *Journal of Accounting Research* 56 (2018): 521–79.

EYRING, H., AND V. G. NARAYANAN. "Performance Effects of Setting a High Reference Point for Peer-Performance Comparison." *Journal of Accounting Research* 56 (2018): 581–615.

FINANCIAL ACCOUNTING STANDARDS BOARD (FASB). *Qualitative Characteristics of Useful Financial Information*. Statement of Financial Accounting Concepts No. 8. Norwalk, CT: FASB, 2010. Available at http://www.fasb.org/resources/ccurl/515/412/Concepts%20Statement%20No%208.pdf.

FRANCO, A.; N. MALHOTRA; AND G. SIMONOVITS. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (2014): 1502–505.

GELMAN, A. "Commentary: P Values and Statistical Practice." *Epidemiology* 24 (2013): 69–72.

GELMAN, A., AND E. LOKEN. "The Statistical Crisis in Science Data-Dependent Analysis—a "Garden of Forking Paths"—Explains Why Many Statistically Significant Comparisons Don't Hold Up." *American Scientist* 102 (2014): 460–65.

GIGERENZER, G. "Mindless Statistics." *The Journal of Socio-Economics* 33 (2004): 587–606.

GOVERNMENTAL ACCOUNTING STANDARDS BOARD (GASB). *Concepts Statement No. 2 as Amended by Concepts Statements No. 3 and 5: Service Efforts and Accomplishments Reporting*. Norwalk, CT: GASB, 2008.

HAIL, L.; A. TAHOUN; AND C. WANG. "Corporate Scandals and Regulation." *Journal of Accounting Research* 56 (2018): 617–71.

HAMERMESH, D. S. "Viewpoint: Replication in Economics (Réplication En Science économique)." *The Canadian Journal of Economics/Revue Canadienne D'Economique* 40 (2007): 715–33.

JOHN, L. K.; G. LOEWENSTEIN; AND D. PRELEC. "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling." *Psychological Science* 23 (2012): 524–32.

JOURNAL OF ACCOUNTING RESEARCH (JAR). "Author Guidelines." 2018. Available at http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291475-679X/homepage/ForAuthors.html.

JOURNAL OF FINANCIAL REPORTING (JFR). "Guidance for Authors & Reviewers." 2017. Available at https://www.dropbox.com/s/32jjeslh0x3sdkf/JFR%20Guide%20to%20Authors%20and%20Reviewers.pdf?dl=0.

KOWALESKI, Z. T.; B. W. MAYHEW; AND A. C. TEGELER. "The Impact of Consulting Services on Audit Quality: An Experimental Approach." *Journal of Accounting Research* 56 (2018): 673–711.

LI, S. X., AND T. SANDINO. "Effects of an Information Sharing System on Employee Creativity, Engagement and Performance." *Journal of Accounting Research* 56 (2018): 713–47.

MARTINSON, B. C.; M. S. ANDERSON; AND R. DE VRIES. "Scientists Behaving Badly." *Nature* 435 (2005): 737–38.

MCSHANE, B. B.; D. GAL; A. GELMAN; C. ROBERT; AND J. L. TACKETT. "Abandon Statistical Significance." 2017. Preprint available at https://arxiv.org/abs/1709.07588.

NOSEK, B. A.; G. ALTER; G. C. BANKS; D. BORSBOOM; S. D. BOWMAN; S. J. BRECKLER; S. BUCK; C. D. CHAMBERS; G. CHIN; G. CHRISTENSEN; AND M. CONTESTABILE. "Promoting an Open Research Culture." *Science* 348 (2015): 1422–25.

NOSEK, B. A., AND D. LAKENS. "A Method to Increase the Credibility of Published Results." *Social Psychology* 45 (2014): 137–41.

OPEN SCIENCE COLLABORATION. "Estimating the Reproducibility of Psychological Science." *Science* 349 (2015): aac4716.

ROZEBOOM, W. W. "The Fallacy of the Null-Hypothesis Significance Test." *Psychological Bulletin* 57 (1960): 416–28.

SIMMONS, J. P.; L. D. NELSON; AND U. SIMONSOHN. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (2011): 1359–66.

VAN DUIN, S. R.; H. C. DEKKER; J. L. WIELHOUWER; AND J. P. MENDOZA. "The Tone from Above: The Effect of Communicating a Supportive Regulatory Strategy on Reporting Quality." *Journal of Accounting Research* 56 (2018): 467–519.

VATTER, W. J. "Critical Synthesis of Conference Papers." *Journal of Accounting Research (Empirical Research in Accounting: Selected Studies)* 1966 (1966): 228–33.

WASSERSTEIN, R. L., AND N. A. LAZAR. "The ASA's Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70 (2016): 129–33.

ZIMMERMAN, J. L. "Improving a Manuscript's Readability and Likelihood of Publication." *Issues in Accounting Education* 4 (1989): 458–66.