PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/rsta

Opinion piece

Cite this article: Dreber A, Toussaert S. 2026 Scientific workflow in experimental economics. *Phil. Trans. R. Soc. A* **384**: 20240606.
https://doi.org/10.1098/rsta.2024.0606

Received: 29 March 2025
Accepted: 9 October 2025

One contribution of 15 to a theme issue 'Statistical workflow'.

Subject Areas:
statistics

Keywords:
scientific workflow, lab experiments, online experiments, pilots, experimental economics

Author for correspondence:
Anna Dreber
e-mail: anna.dreber@hhs.se

THE ROYAL SOCIETY PUBLISHING

Scientific workflow in experimental economics

Anna Dreber[1] and Séverine Toussaert[2]

[1]Stockholm School of Economics, Stockholm, Sweden
[2]University of Oxford, Oxford, UK

AD, 0000-0003-3989-9941; ST, 0000-0002-2491-6713

Lab and online experiments are widely used tools in economics, as well as in other areas of the social sciences. They are typically designed either to test treatment effects or to elicit parameter values for concepts such as economic preferences. While the role of pre-registration of designs and analyses is increasingly discussed in workflows, many other aspects of the research process are less visible in the communicated output. We outline what we view as the common workflow for lab and online experiments in economics, highlight the steps that we believe are often 'missing' and discuss how these omissions may undermine the replicability, credibility and generalizability of published findings.

This article is part of the theme issue 'Statistical workflow'.

1. Background

Until not so long ago, it was widely believed that experiments had no place in the economist's toolbox. For example, Nobel laureates Paul Samuelson and William Nordhaus [1] wrote in their introductory textbook (cited in [2]): '*Economists have no such luxury when testing economic laws. They cannot perform the controlled experiments of chemists or biologists because they cannot easily control other important factors. Like astronomers or meteorologists, they generally must be content largely to observe*'. Over the past few decades, however, lab experiments—and increasingly also online experiments (alongside field experiments, which we do not cover here)—have become a standard part of economics, as well as of related social science fields not traditionally considered 'experimental'. Experiments are now used

to test and refine theories, isolate specific mechanisms and generate new insights that require a controlled environment such as the lab (e.g. [3]).

As with many other methodological approaches (e.g. [4,5]), there is growing evidence that a substantial share of experimental results in the social sciences, including economics, is selectively reported and fails to replicate [6–10]. This has led to increasing interest in how researchers generate hypotheses, design studies and conduct analyses in ways that yield more or less credible, replicable and generalizable results.[1]

There are many reasons why published experimental results may not replicate with new data. First, even in the absence of selective reporting, studies with low statistical power are less likely to replicate simply owing to sampling variability [11,12].

Second, 'researcher degrees of freedom' in data collection and analysis—such as testing multiple conditions or outcomes but reporting only significant results (as discussed in e.g. [13,14])—together with publication bias, distort the sampling distribution of estimators. This leads to over-rejection relative to nominal significance levels and confidence intervals that fail to achieve their intended coverage, particularly when true effects are small [15].

Third, some results may fail to replicate simply because the underlying hypotheses—whether selected *ex ante* or *ex post*—had low prior credibility. Even if the results are statistically significant, a Bayesian agent would still assign a low posterior probability. When hypotheses are generated after the results are known but presented as if they were specified in advance (a practice known as HARKing [16]), readers may be misled into overestimating the strength of the evidence.

Fourth, results may fail to replicate because of differences in implementation between the original study and the replication. Such differences can be unintended, arising from incomplete protocols in attempted 'direct' replications, or intentional, as in conceptual replications where researchers deliberately vary aspects of the design or analysis. In both cases, limited knowledge about heterogeneity and generalizability makes it difficult to interpret divergent findings.

Finally, while most discussions focus on researcher degrees of freedom in analysis and reporting, there is also considerable flexibility at the design stage of lab and online experiments. This flexibility is rarely documented in published work.

The purpose of this article is not to catalogue all the degrees of freedom available to experimentalists, as that would be impossible. Rather, we aim to encourage researchers to reflect on the many decisions they face, what is currently communicated about those decisions and how greater transparency could improve both inference (i.e. what readers take away) and research investments (i.e. the paradigms that researchers develop and extend).

We pursue this aim by examining the typical workflow of lab and online experiments in economics. Section 2 describes each phase of production and highlights aspects that are not fully transparent in published papers. Section 3 discusses how transparent workflows can enhance replicability. Section 4 concludes with a discussion of the challenges to enabling greater transparency and potential solutions.

Our focus is on economics, both because this is where we conduct most of our own research and because we have previously studied issues of replication, generalizability and prediction in this context. At the same time, we draw selectively on insights from neighbouring disciplines to situate our discussion in a broader perspective, without attempting a comprehensive review of related work.

2. Decomposing the scientific workflow in experimental economics

The workflow of an experimental economics project involves a series of interconnected steps. These typically include selecting a research question, formulating a hypothesis, designing the experiment to operationalize the test, pre-testing the design in some form, collecting the main data, conducting statistical analyses, deciding which data and results to present, choosing a

journal and revising the manuscript in response to audience or reviewer feedback. The process is often iterative.

In this article, we adopt a broad view of the workflow, focusing on the process up to journal acceptance. We take this as a natural boundary, since production and evaluation are typically intertwined. At the same time, we limit our analysis to the production and publication of a single experiment. In practice, many experimental papers include additional elements such as multiple experiments, simulations, observational data, theoretical modelling or structural estimations. These added components introduce further complexity and raise important questions for transparency, but fall outside the scope of this discussion.

While not universally followed, practices such as piloting and pre-registering designs or analyses prior to data collection have become increasingly common in experimental economics. For this reason, we devote particular attention to these practices. Other approaches, including the use of simulations or large language models to select experimental parameters, remain relatively uncommon and are left for future work.

Despite growing norms around open science, only a limited number of workflow elements can typically be inferred from published papers. The connections between different stages of the research process are often unclear. Although posting instructions, software files and data are common practice in experimental economics and related fields, these materials usually reflect only the final implementation. Academic articles rarely describe the full data-generating process in enough detail to reconstruct how the experiment was conceived or how decisions evolved over time. Moreover, reporting practices remain highly variable, making comparisons across studies difficult.

In what follows, we outline what we believe to be common practice at each stage of the experimental workflow and identify aspects that are often underreported in published work. Where relevant, we also reflect on differences between lab and online experiments. To move beyond purely anecdotal impressions, we refer, where applicable, to reporting patterns observed in the 23 articles published in the three standard issues of *Experimental Economics* in 2024, the leading field journal for research in experimental economics.[2]

(a) Choosing what to study

The experimental setting limits what is possible to study while also enabling researchers to generate their own data on new and exciting topics. In many other areas of economics, research topics are heavily shaped by data availability and, in some cases, the feasibility of causal inference. In experimental economics, by contrast, the methods and setting place their own constraints on what can be meaningfully studied, even as they offer greater control.

Consider the topic of competition and moral behaviour—a question of interest not only to economists but also to political scientists, psychologists and sociologists. Randomizing participants to a lifetime of exposure to a capitalist versus a socialist system, holding everything else constant and exploring how this affects their lifetime moral behaviour is beyond what the experimentalist can achieve in a lab or online setting. Such questions might be studied using so-called 'natural experiments', e.g. by comparing individuals who grew up in West versus East Germany, but these approaches typically rely on different and often more demanding identifying assumptions that can be problematic. What is feasible in the lab or online, by contrast, is a short-run intervention: for example, participants can be randomly assigned to perform a task under either a relative performance incentive or a piece-rate scheme (thus manipulating the degree of competition), followed by a decision involving moral behaviour. Hence, the credibility of causal identification is achieved, but at the expense of making narrower and less general claims.

Research questions can remain stable throughout the course of a project or evolve significantly over time, shaped by feasibility constraints, empirical results and feedback from collaborators or reviewers. In most published papers, however, it is difficult to determine

whether, or to what extent, the research question changed during the project. For instance, did the literature review meaningfully shape the topic and approach, or were most cited papers simply used to justify a pre-existing research focus? Did the paper's framing or scope change as a result of peer review? Such questions about the evolution of ideas are rarely addressed directly in the final write-up.

Another important dimension of 'choosing what to study' is selecting the population to sample from. Convenience samples—such as university students in lab settings or semi-professional online participants recruited via platforms like Amazon Mechanical Turk or Prolific Academic—are commonly used in experimental economics. In some cases, more targeted populations, such as finance professionals or politicians, have been studied. However, the rationale for the choice of sample and the recruitment strategy is often not discussed at all, making it difficult to assess potential selection bias.

Lab and online experiments face different constraints that shape the research questions they can address. Lab experiments allow for complex, interactive and high-stakes environments but are typically limited to smaller and less diverse samples. Online experiments, by contrast, make it possible to recruit large and diverse samples, but are better suited to simpler, lower-stakes designs where reduced control and limited scrutiny are less problematic. As shown in previous work [17], lab and online pools can substantially differ in terms of data quality.

(b) Formulating a hypothesis

Hypothesis formulation is a central step in experimental research. We distinguish several dimensions along which hypotheses can vary: importance, specificity, timing and epistemic status.

Importance refers to whether a hypothesis is central to the study (primary) or peripheral (secondary). Primary hypotheses typically inform the overall design of the experiment, particularly the main treatment and control conditions. Secondary hypotheses may be motivated by theory or prior evidence and often relate to subgroup analyses or auxiliary outcomes.

Specificity captures how precisely a hypothesis is stated. Some hypotheses are broad (e.g. 'competition affects moral behaviour'), while others are more narrowly defined (e.g. 'relative performance incentives increase cheating in a dice game'). Many papers begin with a broad hypothesis for motivation and then proceed to test a more specific version. Often, little justification is provided for the level of specificity chosen, or how the broader and narrower hypotheses are logically connected.

Timing concerns when a hypothesis is formulated: either *ex ante*, prior to data collection, or *ex post*, in response to the data. The timing can affect both the interpretation and credibility of a finding. For instance, a hypothesis may appear confirmatory but, in practice, has been refined after observing the data, e.g. by restricting its scope to particular subgroups or outcomes.

Relatedly, epistemic status refers to whether a hypothesis is presented as confirmatory (pre-specified and theoretically motivated) or exploratory (often generated post hoc, though it could also be pre-specified as exploratory). In most published work, this distinction is not made explicit, making it difficult for readers to assess the evidentiary status of a finding. This is distinct from HARKing, where researchers may present hypotheses as confirmatory even though they were not pre-specified (sometimes without realizing it), as discussed in the introduction. Both forms of ambiguity can contribute to misperceptions about the strength of the evidence.

Pre-analysis plans and pre-registration can help clarify the epistemic status of hypotheses. When researchers register their hypotheses in advance, readers gain insight into which claims are confirmatory and which are exploratory. Pre-analysis plans also help reduce the risk of unintentional hypothesis refinement in response to observed results. While some readers may place more weight on confirmatory evidence, others may value exploratory findings equally.

Our view is that exploratory analysis plays a crucial role in generating new hypotheses and should not be discouraged, but that distinguishing between confirmatory and exploratory claims is important for honest and cumulative science.

Finally, most experimental papers draw on some form of theory (whether verbal or formal) to justify their hypotheses and designs. Theory is usually presented prior to the empirical analysis in the paper and serves as a source of *ex ante* justification. However, the timing of theoretical development relative to data collection is often unclear. In some cases, theory may be developed after the results are known, in a way that resembles HARKing. This practice, sometimes referred to as 'theory-hacking', involves tailoring the theoretical model to fit the data without making this process transparent to the reader. While we do not focus on this issue in the present analysis, we view it as an important and understudied dimension of research transparency.

(c) Conceptualization and operationalization

Moving to the next stage, the researcher must decide how to conceptualize and operationalize the research question and hypotheses for empirical testing. Continuing with our earlier example, consider the hypothesis 'relative performance incentives lead to more cheating'. What task should be used to measure cheating? Even when narrowed to a specific setting—such as the commonly used dice game [18]—and a more precise hypothesis like 'relative performance incentives lead to more cheating in a dice game', multiple design choices remain.

First, how should the relative performance task be structured? For instance, how many competitors should be included, and what information should be provided about them to induce sufficient competitive pressure? Second, what level of monetary incentives should be used in the dice game to allow for meaningful variation in cheating across conditions? Third, should the study employ a between-subject or within-subject design to estimate the treatment effect?

While such decisions are routinely made, the rationale behind them is rarely explained in published papers. Yet, these choices can substantially shape the distribution of treatment effect estimates. Researchers often have private information (e.g. gained through piloting, prior studies or domain knowledge) about how specific experimental parameters affect outcomes. When this knowledge is not disclosed, readers cannot assess whether the featured design reflects a best-case scenario, a typical case or something else entirely. Greater transparency around these design decisions would help clarify the boundary conditions of research paradigms and offer more informative guidance for future work.

(d) Pilots

Pilots are closely related to the steps of conceptualization and data collection and are probably among the most undisclosed elements of the experimental workflow. Based on our reading of published experimental economics papers over the years, one might assume that pilots are rarely used at all. For example, among the 23 *Experimental Economics* articles mentioned above, only one explicitly mentions a pilot. Yet, we know from personal experience and discussions with colleagues that pilots are commonly used, suggesting that transparency may be an issue.

Pilots can refer to any form of data collection or testing carried out prior to what is ultimately labelled the main study. However, this distinction is often endogenous to the results: if a pilot 'works out' particularly well, it may simply become the main data collection. Pilots can serve various purposes, including testing recruitment strategies, assessing comprehension of instructions, verifying that the software functions as intended or evaluating whether a manipulation achieves its intended effect, e.g. ensuring a task is sufficiently boring or exciting (see [19] for a more complete list of example uses).

The various forms of piloting can be broadly categorized as 'implementation piloting' and 'data piloting' [20]. Implementation piloting refers to checks related to study logistics and clarity and is typically unproblematic and often helpful, though it should still be disclosed. Data piloting, by contrast, involves using early data to estimate parameters such as means and variances for power calculations, or to assess whether the treatment appears effective under specific design choices. This type of piloting is more problematic, as such early stage data are typically underpowered and may yield unreliable inferences. Still, if conducted, it should be transparently reported.

'Design-hacking' has been discussed in this context, referring to the practice of running multiple pilots and tweaking the design until one produces 'interesting' results [3]. Returning to our earlier example, a research team studying competition and moral behaviour might test several closely related designs, varying some parameter values, until they observe promising results in one condition and then scale up that version as the main study. This approach limits the generalizability of the findings, but if the iterative piloting process is not disclosed, these limitations only become apparent when others attempt conceptual replications or extensions. In such cases, claims of existence ('this phenomenon can occur') may be mistaken for claims of prevalence ('this phenomenon occurs readily').

A recent study on attitudes towards pre-registration among 519 experimental economists at all career stages finds that a majority of respondents would like pilots to be reported in pre-analysis plans [21]. In follow-up work, a majority of presenters at large experimental economics conferences also reported using pilots for the project they presented. While this is in line with our prior that pilots are often used, we did not expect that the support for reporting them would be so high, given that they often remain unreported in papers. Beyond helping others learn from effective piloting strategies, such transparency would also clarify the boundary conditions of reported results. In sum, we believe—as emphasized by [19] in the context of psychology—that piloting, when used wisely, can be extremely valuable, but transparency about its use is essential.

Our prior, based on anecdotal evidence, is that pilots are more common in online studies than in lab studies, and that these are often not disclosed. Online pilots are cheap and can be performed very quickly with large samples, potentially blurring the lines between pilots and main data collection.

(e) Data collection

In the data collection phase, researchers must make a series of decisions about sampling, sample size, number of conditions and how participants are allocated to those conditions. Some of these choices, such as the population from which participants were drawn, can typically be inferred from the final paper. Others are more difficult to discern. For example, conditions or participants may be dropped from the analysis (as discussed further below), leading to incomplete or potentially misleading accounts of the actual workflow.

In the absence of an *ex ante* power calculation, papers in economics often do not explain whether data collection followed a predefined stopping rule or how sample size was determined. For instance, it may have been shaped by funding constraints or limitations on subject pool availability. Details such as the timing of data collection, the implementation of experimental conditions or the method of participant assignment are also often left unspecified. Were participants randomized between sessions or within sessions? Could participants self-select into sessions that were themselves randomized to conditions? If randomization occurred at the session level, were standard errors adjusted accordingly? Reasons for including specific conditions at different stages of a paper's life—whether this is due to reviewer comments, journal requirements or the authors' own decisions—can sometimes, but not always, be inferred.

Sharing experimental instructions has long been standard practice in experimental economics, enabling others to build on prior work and assess the effect of variation in instructions. But it is also possible to share more detailed information about other parts of the data collection process, such as when the data was collected, how participants were allocated to conditions and how conditions were selected. For online experiments, which are typically fully computerized, sharing survey files (e.g. from platforms like Qualtrics) offers a direct way to document the participant experience. These files make it possible to see not only the exact instructions but also the structure and flow of the study, including randomization and branching logic. They can be easily included as supplementary materials and offer greater transparency about how the experiment was implemented. Among the 23 *Experimental Economics* papers we reviewed, 14 report the year of data collection, and all 23 include the experimental instructions.

(f) Analysis

When the analyses presented in a paper are based on a pre-analysis plan that is closely followed, this step of the workflow is typically transparent. When there is no such pre-analysis plan and/or reporting is not transparent, it becomes difficult to determine whether the analyses were specified in advance or shaped in response to the data. In practice, many papers probably reflect a mix of both: some analyses were pre-planned, whereas others emerged during the research process, influenced by new ideas, feedback from colleagues, reviewer and editor requests, robustness concerns or a desire for more convincing or statistically significant results.

To anchor ideas, consider our running example: a study on relative performance incentives and cheating in a dice game. Suppose the experiment includes two conditions (treatment and control), and background data are collected on participants' gender, age and self-reported competitiveness. If the analysis reveals no main treatment effect on cheating, the researcher might restrict the sample to women and observe a significant effect, rationalizing it by suggesting that women are more sensitive to contextual cues. Alternatively, an effect may appear only among participants who are more competitive than the median, and this might be explained by arguing that these individuals are more responsive to tournament incentives. In another scenario, the researcher might find a significant main effect in the full sample and, satisfied with this result, choose not to conduct any subgroup analyses (see [14] for a broader discussion of such analytical flexibility and forking paths).

Given how central the analysis and results are to a study's contribution, it is striking that this stage of the workflow often remains relatively opaque. Pre-analysis plans are increasingly used in experimental economics and can provide greater transparency by clarifying which analyses were planned in advance. However, uptake is still far from universal, and even with a pre-analysis plan, communication about deviations can be much improved (more on this in §2h). Tools such as multiverse analyses and specification curves offer additional ways to communicate the robustness of findings, but are still rarely used in practice. Among the 23 papers published in *Experimental Economics* that we reviewed, only seven mention a pre-analysis plan or pre-registration, none mention the word 'confirmatory' in relation to the study, hypothesis or test, while seven mention the word 'exploratory' in that context.

(g) Write-up and dissemination

This final stage of the workflow is closely connected to all preceding steps, as it determines how the research narrative is ultimately constructed and presented. As the paper evolves—shaped by the authors' own changing views, feedback from seminars or reviewer comments—the link between the original workflow and the version conveyed in the manuscript often becomes increasingly opaque. In some cases, this disconnect may already emerge by the time the first draft is written; in others, it deepens further during the dissemination process.

Consider a preprint that frames the study narrowly, for example, as a test of whether tournament incentives increase cheating in a dice game. The published version may instead present it as a broader contribution to the literature on competition and moral behaviour. Alternatively, a null result for the overall sample might be moved to an appendix, while a subgroup finding (such as a gender difference) is promoted to the main text. These kinds of shifts may reflect reasonable narrative refinements, but without transparent documentation, it becomes difficult to assess how the framing evolved and what prompted the changes.

One way to improve transparency would be to systematically link preprints to their published versions, enabling readers to track changes in framing, analysis or results over time. At the same time, such practices could affect researchers' willingness to post early versions in the first place, especially if discrepancies become a focus of scrutiny.

(h) Pre-registration and pre-analysis plans

While the concept of pre-registration often entails a detailed pre-analysis plan when discussed in a field like psychology (e.g. [22]), pre-registration in economics is a more fuzzy concept and can be as vague as a registration that a study will be conducted (see [20] for more discussion). This is the reason why we mainly refer to 'detailed pre-analysis plans' in this subsection.

If studies were pre-registered with detailed pre-analysis plans and papers clearly indicated when they followed or deviated from those plans, many of the current gaps in the reported scientific workflow could be avoided. For example, it would become more transparent whether the research question or hypotheses evolved over time. We have already highlighted some benefits of pre-analysis plans in previous sections. However, they are not always used in a way that maximizes these benefits. A pre-analysis plan might be vague, leaving considerable researcher degrees of freedom [23]. Conversely, a plan might be highly detailed, but the resulting paper may cite it merely as a signal of credibility while substantially deviating from it without informing the reader.

In economics, there is no clear norm on how to report and discuss pre-analysis plans and eventual deviations (e.g. [24] who study pre-analysis plans and papers in economics and political science and document substantial lack of transparency). Pilots may also remain undisclosed, since their existence is difficult to verify if not reported. Some researchers may use extensive piloting precisely to inform detailed pre-analysis plans they are unlikely to deviate from. In addition, if pre-registered studies are more likely to yield null results than non-pre-registered ones, they may be less likely to be written up, especially if authors expect difficulty publishing them. The Registered Reports format [25], in which the study is reviewed before data collection begins, may help mitigate this concern.

Several recent studies examine researchers' perceived costs and benefits of pre-analysis plans. For example, a survey of 355 researchers, mainly in psychology, finds that pre-registration is often seen as increasing both project duration and work-related stress; nonetheless, experiences and expectations are generally positive, with most respondents believing that pre-registration improves research quality [26]. In a survey of 519 researchers, primarily in experimental economics, a majority report being somewhat or very favourable to pre-analysis plans, while about a quarter are somewhat or very unfavourable [21]. This can be contrasted with attitudes towards replications, which are almost unanimously positive in the same sample.

3. Relation to replications

We argue that a more transparent workflow can enhance replicability not only by increasing the credibility of original results but also by reducing the likelihood of unintended differences between an original study and its replication.

For some studies, a direct replication may not be appropriate because certain aspects have changed over time. To continue with our running example on competition and moral

behaviour, suppose we are now interested in a related study on the effect of peacefulness—which might be considered the opposite of competition—on cheating. An original study used a particular movie clip to induce emotions, and the replicators attempted a direct replication with the goal of staying as close as possible to the original study; with another movie clip, the replication would instead be a conceptual one. The original movie clip was, at the time of the original study, considered extremely peaceful, and the goal of the movie is to induce feelings of calmness and peacefulness. However, the main actor has since been convicted of repeated aggravated assault, and the clip no longer evokes peacefulness owing to this association. If the replication fails to reproduce the original result, it might be due to a shift in perception of the movie clip or because the original finding was not robust.

Better documentation of how the original movie clip was selected could have helped replicators identify a replacement that matched the spirit of the original study as closely as possible. For instance, if the original researchers had recorded that the clip was chosen based on participant ratings of peacefulness, that procedure—not just the stimulus—could be replicated. More broadly, whether the element in question is a movie clip, task or another contextual feature, it is not enough to state *what* one did—it is better to explain *how* one did it.

Piloting and design-hacking also have implications for replicability. These practices may yield results that are more likely to replicate successfully in a direct replication but also more likely to fail in a conceptual replication, suggesting limited generalizability. In many cases, we believe conceptual replications can be more informative than direct ones. Still, direct replication is often the natural first step: if an effect does not hold up even under near-identical conditions, it may not be worth pursuing further conceptual extensions. Beyond replications, re-analyses of existing experiments (e.g. as part of a meta-analysis) can also generate valuable insights, provided the underlying data are relatively free from selective reporting and publication bias.

4. Discussion

Why did we keep returning to the example of a study on the effects of competition on moral behaviour? This is an area where previous studies have reported statistically significant results in opposite directions, with considerable variation in designs, settings and samples. In a recent project aimed at holding everything but the design constant, multiple research teams were asked to develop an incentivized online experiment with a control group and a treatment group to test the effects of competition on moral decision-making [27]. More than 18 000 participants were randomized across 45 experimental designs, randomly selected from 95 submissions. As expected, there was wide variation in both the competition manipulations and the outcome tasks. The resulting design heterogeneity was substantial—about 1.6 times larger than the average standard error of the effect size estimates. Scaling up this kind of initiative could help us better understand how workflow decisions shape the results for a given hypothesis.

Why are certain aspects of the workflow not disclosed? We see several plausible reasons:

(1) *Improper documentation*. Researchers often forget the many decisions made during a project and, therefore, may not think to document them. This underscores the value of maintaining 'lab notebooks', as is common in the natural sciences, to record the rationale behind design choices, failed attempts, meeting notes with co-authors and other relevant information. Researchers could also adopt a 'lab handbook', as described by [28], to share with PhD students and collaborators. While economists typically do not work in formal lab groups to the same extent as researchers in other fields, such a handbook could still serve as a living document covering scientific principles, roles and expectations, open science practices, data management, ethics, safety and more—especially in settings where not all research activities are visible to every member of a research team [28].

(2) *Opaque peer review*. Journals also bear some responsibility for the lack of transparency. Reporting guidelines are often vague and rarely require disclosure of important aspects

of the workflow, such as whether pilots were conducted or how specific design and analysis choices were made. Moreover, the peer review process itself can directly shape the research workflow by prompting authors to add new conditions, collect more data or deviate from their pre-registered analyses. Yet, it is often unclear which changes stem from reviewer or editorial demands. Journals could improve transparency by requiring authors to include a brief summary of major changes made during peer review, published alongside the final article. More broadly, a shift towards a more open peer review process could help, and survey evidence suggests some support for this among experimental economists [29].

(3) *Self-image concerns*. Reflecting on failed pilots or design missteps can be ego-threatening, especially when researchers view these setbacks as personal shortcomings. This may discourage honest self-assessment or the disclosure of lessons learnt and undermine the efficacy of post-mortems. Third-party involvement (such as a mentor, co-author or senior colleague) can help reframe the post-mortem process as a shared opportunity for growth rather than a personal reckoning. Over time, fostering a culture that values thoughtful reflection, e.g. through opinion pieces or open discussions, may encourage researchers to engage more openly with their own workflows.

(4) Researchers may fear that disclosing failed pilots or flawed design choices will make others question their competence or professionalism ('Why couldn't they anticipate this? I would have never run such a stupid treatment!'). To this date, we do not know if these fears are actually warranted, although existing incentives clearly tend to reward polished success over transparent reflection. Journals, funders and universities could help shift this norm by encouraging brief methodological reflections, soliciting 'what didn't work' sections or recognizing workflow transparency through open materials badges. Enabling citation of protocols (e.g. by assigning DOIs) could further increase discoverability and help align transparency with researchers' professional incentives. Dedicated journals (such as the *Journal of Trial and Error*) could also support this effort, though we believe these issues should be discussed in general economics journals for broader effect. How to build effective reward systems is still an open question, but it is worth noting that some universities (for example, in the UK) now include open research outputs in promotion criteria and research assessments [30].

Building on our discussion in §3, we argue that better documentation of the workflow is important for at least five reasons:

(1) *Replicability*. Better workflow documentation can help original authors build on their own work and enable others to replicate or extend studies. Without sufficient detail, it may be difficult to distinguish between a direct and a conceptual replication.

(2) *Efficiency*. Transparency about the workflow allows researchers to identify which steps are essential, which can be streamlined and which might be unnecessary, making future projects more efficient.

(3) *Equity*. Sharing the 'recipe' for successful projects can level the playing field, giving researchers with fewer resources or less access to informal knowledge the tools to succeed.

(4) *Value of information*. While researchers may worry that disclosing failed attempts will make them look like 'low types', the reverse is often true: if a novel task or paradigm emerges after several iterations, knowing what did not work can increase appreciation for the final result. Without this context, especially given our natural hindsight bias, the outcome may appear obvious or trivial.

(5) *Credibility*. More complete documentation of the workflow enhances credibility by reducing researcher degrees of freedom and limiting opportunities for selective reporting.

While we have focused on what we see as key building blocks in the workflow of experimental economists, there are additional steps and disclosures that we have not covered here—many of which could generate substantial benefits if shared. For instance, the growing use of artificial intelligence is likely to shape multiple stages of the research process, and transparency

about how these tools are used will become increasingly important. Similarly, lab and online experiments typically require ethical approval or a waiver from an institutional review board or equivalent body. Yet discussions of how ethical concerns were handled—if any arose—are almost never reported. Finally, information on the financial and time costs of running experiments remains scarce, even though such information is critical for evaluating the feasibility and value of research. In sum, we believe there should be more attention to what happens 'behind the scenes'.[3]

We also recognize that many of our own projects have not followed fully transparent workflows. Writing this text has prompted us to reflect on what we could do better as experimental researchers, and we hope it does the same for readers.[4]

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. The data are publicly available on https://osf.io/gysf5/.

Declaration of AI use. We used ChatGPT-4o and ChatGPT-5 to improve the readability of the paper.

Authors' contributions. A.D.: conceptualization, funding acquisition, investigation, methodology, project administration, writing—original draft, writing—review and editing; S.T.: conceptualization, investigation, methodology, project administration, writing—original draft, writing—review and editing.

Both authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was supported by the Knut and Alice Wallenberg Foundation, the Marianne and Marcus Wallenberg Foundation (Wallenberg Scholar grant to Anna Dreber) and the Jan Wallander och Tom Hedelius Foundation (Grants No. P23-0098 and P25-0210).

Endnotes

[1]While this paper focuses on experimental economics, we believe there are even stronger reasons to be concerned about the credibility of non-experimental quantitative research in economics and the social sciences (e.g. [4,5]). Experimental studies have been the focus of replication efforts in part because replication is actually feasible in this context, unlike for many other methods and data sources where replication is difficult or impossible.

[2]The data are available on https://osf.io/gysf5/.

[3]One of us (Séverine Toussaert) has set up a 'Behind-the-Scenes' seminar series with Vatsal Khandelwal, the goal of which is to learn about the production of research papers. The seminar allows students and researchers in all fields and at all career stages to discuss their research process and share challenges and solutions (https://www.bts-seminar.net/).

[4]In the spirit of this article, we briefly outline our own workflow. Anna Dreber produced the first draft, which was then edited to incorporate additional ideas from Séverine Toussaert. Reviewer feedback prompted us to substantiate our claims with further references and to include descriptive statistics on reporting practices in recently published experimental economics papers. The manuscript underwent extensive editing with the assistance of ChatGPT-4o and ChatGPT-5. All remaining errors are our own.

References

1. Samuelson PA, Nordhaus WD. 1985 *Economics*, 12th ed. New York: McGraw-Hill.
2. List JA. 2011 Why economists should conduct field experiments and 14 tips for pulling one Off. *J Econ Perspect.***25**, 3–16. (doi:10.1257/jep.25.3.3)
3. Fréchette GR, Sarnoff K, Yariv L. 2022 Experimental economics: past and future. *Annu. Rev. Econom.* **14**, 777–794. (doi:10.1146/annurev-economics-081621-124424)
4. Brodeur A, Lé M, Sangnier M, Zylberberg Y. 2016 Star wars: the empirics strike back. *Am. Econ. J Appl. Econ.* **8**, 1–32. (doi:10.1257/app.20150044)
5. Brodeur A, Cook N, Heyes A. 2020 Methods matter: p-hacking and publication bias in causal analysis in economics. *Am. Econ. Rev.* **110**, 3634–3660. (doi:10.1257/aer.20190687)
6. Open Science Collaboration. 2015 Estimating the reproducibility of psychological science. *Science* **349**, aac4716. (doi:10.1126/science.aac4716)

7. Camerer CF *et al*. 2016 Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436. (doi:10.1126/science.aaf0918)
8. Camerer CF *et al*. 2018 Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644. (doi:10.1038/s41562-018-0399-z)
9. Klein RA *et al*. 2018 Many labs 2: investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490. (doi:10.1177/2515245918810225)
10. Holzmeister F *et al*. 2025 Examining the replicability of online experiments selected by a decision market. *Nat. Hum. Behav.* **9**, 316–330. (doi:10.1038/s41562-024-02062-9)
11. Ioannidis JPA. 2005 Why most published research findings are false. *PLoS Med*. **2**, e124. (doi:10.1371/journal.pmed.0020124)
12. Gelman A, Carlin J. 2014 Beyond power calculations: assessing type S (Sign) and Type M (Magnitude) errors. *Perspect. Psychol. Sci.* **9**, 641–651. (doi:10.1177/1745691614551642)
13. Simmons JP, Nelson LD, Simonsohn U. 2011 False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366. (doi:10.1177/0956797611417632)
14. Gelman A, Loken E. 2013 The garden of forking paths: why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time. See https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf.
15. Andrews I, Kasy M. 2019 Identification of and correction for publication bias. *Am. Econ. Rev.* **109**, 2766–2794. (doi:10.1257/aer.20180310)
16. Kerr NL. 1998 HARKing: hypothesizing after the results are known. *Pers. Soc. Psychol. Rev.* **2**, 196–217. (doi:10.1207/s15327957pspr0203_4)
17. Snowberg E, Yariv L. 2021 Testing the waters: behavior across participant pools. *Am. Econ. Rev.* **111**, 687–719. (doi:10.1257/aer.20181065)
18. Abeler J, Nosenzo D, Raymond C. 2019 Preferences for truth - telling. *Econometrica* **87**, 1115–1153. (doi:10.3982/ECTA14673)
19. Handley-Miner IJ *et al*. 2025 A call for greater transparency in piloting. See https://osf.io/preprints/psyarxiv/q95zn_v1.
20. Coffman LC, Dreber A. 2025 Running replicable experiments. In *Handbook of experimental economics methods*. Amsterdam, The Netherlands: Elsevier.
21. Imai T, Toussaert S, Baillon A, Dreber A, Ertac S, Johannesson M, Neyse L, Villeval MC. 2025 Pre-registration and pre-analysis plans in experimental economics. See https://docs.iza.org/dp17821.pdf.
22. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. 2018 The preregistration revolution. *Proc. Natl Acad. Sci. USA* **115**, 2600–2606. (doi:10.1073/pnas.1708274114)
23. Brodeur A, Cook NM, Hartley JS, Heyes A. 2024 Do preregistration and preanalysis plans reduce *p* -hacking and publication bias? Evidence from 15,992 test statistics and suggestions for improvement. *J. Polit. Econ. Microecon.* **2**, 527–561. (doi:10.1086/730455)
24. Ofosu GK, Posner DN. 2023 Pre-analysis plans: an early stocktaking. *Perspect. Polit.* **21**, 174–190. (doi:10.1017/S1537592721000931)
25. Chambers CD. 2013 Registered reports: a new publishing initiative at cortex. *Cortex* **49**, 609–610. (doi:10.1016/j.cortex.2012.12.016)
26. Sarafoglou A, Kovacs M, Bakos B, Wagenmakers EJ, Aczel B. 2022 A survey on how preregistration affects the research workflow: better science but more work. *R. Soc. Open Sci.* **9**, 211997. (doi:10.1098/rsos.211997)
27. Huber C *et al*. 2023 Competition and moral behavior: a meta-analysis of forty-five crowd-sourced experimental designs. *Proc. Natl Acad. Sci. USA* **120**, e2215572120. (doi:10.1073/pnas.2215572120)
28. Mehr S. 2020 How to... write a lab handbook. In *The Biologist* pp. 26–29
29. Charness G, Dreber A, Evans D, Gill A, Toussaert S. Open Science, Closed Peer Review? *J. Econ. Sci. Assoc.*
30. University of Bristol. 2024 *A toolkit for recognising and rewarding open research*