

EVALUATING AND IMPROVING EXPERIMENTAL DESIGNS IN ECONOMICS

Daniel Evans

Luca Henkel

Brian Jabarian

John A. List

October 16, 2026

Preliminary version: Feedback welcome

Abstract

Conducting experiments requires researchers to make numerous design choices while respecting statistical, practical, and ethical constraints. We develop the Experimental Design Assessment Tool (EDAT) to help researchers navigate this complexity when generating and evaluating evidence from experiments. EDAT is based on a comprehensive framework that distills experimental designs into 30 components and provides guidance on optimal design choices for each. We operationalize this framework with an evaluation instrument that guides evaluators through experimental designs using targeted questions and generates quantitative quality scores. We evaluate EDAT's performance by applying it to experiments published in eight top economics journals in 2000, 2001, and 2020. Evaluators' scores show high reliability and strongly predict citation counts, publication outcomes, and other proxies for impact and quality. This suggests that a quantitative assessment of experimental design quality is feasible. They further reveal that there cannot be a "perfect experiment" in practice due to the existence of trade-offs, such as between certain components of internal and external validity. We show how researchers navigate these trade-offs and identify the design features most valued by the profession. Finally, we document systematic improvements in quality over time and heterogeneity in design features between subfields and experiment types.

Keywords: Experimental design; experimental economics; methodology; meta-science

JEL Classification: B41; C10; C90; C91; C93

Contact: Daniel Evans: University of Bonn, daniel.evans@uni-bonn.de. Luca Henkel: Erasmus University Rotterdam, ERIM, CESifo, JILAE, IZA@LISER, Tinbergen Institute; henkel@ese.eur.nl. Brian Jabarian: Carnegie Mellon University, CDR, CAAI, CESifo, J-PAL, IGL, JILAE, jabarian@cmu.edu. John A. List: University of Chicago, Australian National University, NBER, JILAE, jlist@uchicago.edu.

Acknowledgments: We sincerely thank Nour Abdelbaki, Ola Aboukhsaiwan, Franco Albino, Moisés Pedrozo, Robert Bernhardt, Evrim Belli, Drew Burd, Yaming Cao, Charles Changhao, Dyuti Chatterjee, Elodie Corvaisier, Charlotte Eisevogel, Zilan Erol, Faith Fatchen, Tim First, Argun Hild, Jiwon Hwang, Wenjie Jiang, Paul Kirch, Anna Mogilevskaja, Christina Petrik, Fidel Petros, Malte Rattenborg, Patrick Schneider, Max Spohn, Müge Süer, Marcos Sugastti, Greg Sun, Marina Talantceva, Cole Vandenberg, Siddarth Venkatesh, and Juliana Velasquez for contributing to the evaluation data of this paper. We are grateful to Thomas Dohmen, Robert Dur, Chris Roth, Peter Wakker, and Florian Zimmermann for valuable comments. Mateo Rubio provided excellent research assistance.

Funding: Daniel Evans is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/2 - 390838866

1 Introduction

Experiments are an essential method for generating scientific knowledge. They allow researchers to control who is exposed to which intervention, thereby generating unique insights into causal relationships. However, designing and conducting economic experiments requires researchers to manage many trade-offs, for instance between controlled and realistic choice environments, while respecting statistical, practical, and ethical constraints. Researchers often struggle to effectively navigate these trade-offs, especially early in their career. This multidimensionality also complicates the evaluation of experimental results for researchers, referees, and editors. Together, these features create significant barriers to generating and evaluating experimental evidence.

This paper introduces the Experimental Design Assessment Tool (EDAT), a framework and assessment tool designed to help researchers navigate the multidimensionality of experiments. The framework comprises five dimensions of experimental quality that cover the core aspects of experimental designs. Each dimension is further divided into a grand total of 30 components that capture specific design features, while a rubric maps each feature to a component-specific quality score. The framework therefore allows us to assess the quality of an experimental design across several dimensions at once. To assist with the application of our framework, we develop a survey-based assessment tool that generates grades for experimental designs. We then apply the tool to published experiments in economics in order to validate EDAT and study how researchers make design decisions in practice.

Our paper contributes to the practice and study of experiments in three ways. First, researchers can use EDAT to improve their designs prior to conducting an experiment. Second, researchers can also apply the assessment tool and framework to evaluate experiments that have already been conducted. In its original form, EDAT provides a systematic method for making the evaluation of experiments more transparent and efficient. It is also purposefully designed to be modular such that researchers can readily replace, extend, or modify any component to suit their needs and preferences. Third, we provide empirical evidence on the current state of experimental designs in economics. The main questions we investigate include whether trade-offs are inherent in experimental designs and, if so, how researchers resolve them and how such decisions are valued by the profession.

The framework divides experimental designs into five core dimensions: *Setup*, *Recovery*, *Estimation*, *Interpretation*, and *Generalizability & Scaling*. Each component within a dimension captures an important design aspect and is constructed such that it can be assigned a quality score. Importantly, the scoring only assesses the experimental design conditional on the research question; it does not assess the question's importance. As an illustrative example, the *Recovery* dimension assesses whether an experiment satisfies the assumptions required to identify a treatment effect. An important component within this dimension is attrition - the threat that participants prematurely drop out of the experiment. For this com-

ponent, quality scoring works as follows: an experiment receives the best score if there is (almost) no attrition. The second-best score is assigned if there is attrition, but statistical tests indicate it occurred randomly; third best is if those tests indicate non-random attrition. Lastly, the lowest score is assigned when no attrition tests are conducted.

Based on the framework, the assessment tool is a survey-style method that guides an evaluator step by step through an experimental design. For every component, evaluators receive background on the relevant concepts and answer targeted questions. The tool then automatically computes the quality scores from their inputs. Thus, the tool can be used to evaluate experiments by anyone with at least some – even if limited – knowledge of experiments.

To validate EDAT and to assess the current state of experiments in economics, we apply the tool to published experimental papers. To generate the dataset, we selected all experiments published in 2000, 2001, and 2020 in eight journals: the top-five economics journals, *Experimental Economics*, *Games and Economic Behavior*, and the *Economic Journal*. This resulted in a sample of 205 papers. We then hired 32 evaluators who applied the tool to the papers. Evaluators were mostly PhD students with research interests in experimental economics. Each paper was evaluated by up to four evaluators. We then merged the resulting data with scientific metadata: a paper’s citation count, publication outlet, the number of times it got included in a university class’s syllabus, its Altmetric score, which aggregates several engagement metrics such as media mentions, and the number of times it was mentioned in policy documents.

We start our empirical analysis by investigating whether the scores produced by EDAT are valid measures of design quality. We use two key metrics to assess validity: whether the scores are reliable and whether they predict scientific outcomes associated with design quality. We assess reliability by examining whether independent evaluators assign similar grades to the same paper. Using Krippendorff’s α as the standard approach in the literature to measure such reliability, we observe $\alpha = 0.81$ when two evaluators assess a paper, and $\alpha = 0.80$ when four evaluators assess a paper. Such levels are commonly considered in the literature to indicate a reliable measure with an almost perfect strength of agreement. As an additional measure of reliability, we calculate the proportion of component-level grades on which evaluators agree exactly. In total, 77% (61%) of grades match exactly when the paper is assessed by two (four) evaluators, and 15% (22%) of grades differ by one grade level. Together, these measures indicate a high level of reliability.

To assess predictivity as the second metric of validity, we regress scientific outcomes on our quality scores. We find that they are strongly predictive of all five outcomes examined: papers with higher EDAT quality scores receive more citations and policy mentions, have higher Altmetric scores, and are more likely to be published in a top-five journal and included in a university syllabus. These findings are robust across a range of alternative specifications and coding choices, such as the inclusion of controls for JEL code, year of

publication, and the number of authors on a paper, different weightings of quality scores, different regression models, and alternative resolution criteria for handling disagreement between evaluators. These results indicate that our quality scores are sufficiently reliable and predictive to allow for a meaningful investigation of the state of experimental design.

We begin by examining whether experimental designs involve trade-offs and, if so, how researchers manage this multidimensionality. We find strong evidence for the existence of trade-offs: while quality scores across many components are positively correlated with each other, negative correlations occur nearly as often. This implies that there is no clear quality sorting among published papers: designing and conducting an experiment that excels in some components necessarily means accepting lower quality in others. As a consequence, the existence of trade-offs means there is no “perfect experiment” that dominates other experiments in every dimension among experiments published in top journals.

We next characterize several notable trade-offs and complementarities across components. First, we find evidence for a trade-off between certain aspects of internal and external validity: higher naturalness of the experimental setting is associated with lower scores on components capturing the assumptions required to identify an internally valid average treatment effect. For instance, naturalness is associated with a higher incidence of spillovers, attrition, and hidden treatments. Importantly, these threats to identification appear to co-occur, as scores in these dimensions are positively correlated with each other. Moreover, experiments that have no selection by design tend to have larger problems with attrition and compliance. At the same time, researchers appear to adjust by employing more powerful statistical tests to detect and account for these threats to identification. In terms of complementarities, we find a positive association between the quality of an experiment’s *Setup* and its *Estimation*. For instance, the quality of its preregistration and its compliance with ethical duties are positively correlated with the likelihood of authors conducting a power analysis and adjusting for multiple hypothesis testing.

Given these trade-offs, how does the economics profession value the different ways in which researchers address them? To answer this question, we compare the predictive power of our five design dimensions for the scientific outcomes. The *Generalizability & Scaling* dimension emerges as the strongest predictor of a paper’s scientific outcomes, including citation count, number of mentions in policy documents, and likelihood of being published in a top-five journal. The association is sizable: an increase in the average *Generalizability & Scaling* score from the median to the 75th percentile is associated with a 12.59 p.p. higher top-five publication likelihood, a 44% higher citation count, and 38% more mentions in policy documents. In contrast, the *Recovery* dimension is not positively and statistically significantly associated with any of the examined outcomes. One potential interpretation of this result is that a sufficiently high quality in *Recovery* is a necessary condition for publication in a top journal. As a consequence, published papers either sufficiently mitigate threats to identification through ex ante design choices or test for violations using powerful statistical

tests ex post. In sum, our analysis reveals that published experiments appear to differ less in internal validity than in how they address external validity, specifically the *Generalizability & Scaling* dimension.

Lastly, we investigate how quality scores differ across types of experiments, research fields, and time. We find significant differences between laboratory/artefactual field experiments, which feature experimenter-constructed environments, and framed field experiments/natural field experiments, which feature natural environments. The former achieve higher grades in the *Recovery* dimension for components such as compliance, hidden treatments, or extent of attrition, while the latter achieve higher grades in the *Generalizability & Scaling* dimension. Likewise, we find systematic differences across research fields, with papers in the JEL categories labor and development economics having higher scores in the *Estimation* and *Generalizability & Scaling* dimensions than papers in the quantitative methods and microeconomics categories. Finally, we document that design quality has increased across all dimensions over time. In particular, we find that the average overall grade among papers published in top-five journals in 2000 and 2001 is lower than the average overall grade of a paper published in one of the field journals of our sample in 2020.

Related literature. Our paper contributes to three strands of research. First, we contribute to a literature on experimental methodology in economics. Important contributions such as Deaton and Cartwright (2018), Imbens (2018), Peters et al. (2018), Niederle (2025), Snowberg and Yariv (2025), and Masselus et al. (2026) as well as the handbook chapters by Kagel and Roth (2016) and Banerjee and Duflo (2017) provide helpful discussions and extensive guidance on designing experiments. We directly build on this literature by using its insights as input for the framework. These guides primarily take a deep dive into a single dimension or a couple of dimensions and/or they do not contain explicit quality scoring. We complement this literature by providing an approach that comprehensively covers the main dimensions of experiments in a single framework, which we then couple with an explicit assessment method to score the individual components. These two features allow our approach to be used to generate quality scores of experiments, which researchers can use to improve experimental designs ex ante and to evaluate completed experiments ex post.

Second, we contribute to a literature that investigates evaluation processes in economics. For instance, Laband and Piette (1994), Cherkashin et al. (2009), and Hofmeister and Krapf (2011) study editorial decision-making and how it depends on paper characteristics. Welch (2014) and Card and DellaVigna (2020) investigate agreement among referees in top finance and econ journals, while Forscher et al. (2019) investigates correlations among evaluators for grant proposals. Generally, these papers find low inter-evaluator correlations. Importantly, referees at journals and evaluators for grants evaluate papers on many different dimensions in addition to the execution of the presented results, for instance a paper's fit with the journal and the importance of the paper's research question. In our project, we focus

on evaluating a particular aspect of research papers, the experimental design. We show that systematic quality assessment scoring of experimental designs yields high inter-evaluator correlations, indicating that experimental design quality can be reliably assessed.

Third, we contribute to a literature that studies trends in economics, and specifically in experimental economics. Fréchette et al. (2022) and Reuben et al. (2022) investigate trends in experimental economics empirically. Akerlof (2020) discusses, and Andre and Falk (2025) investigates empirically normative views of researchers on which aspects of research should be valued. We characterize the state of designs in economics and how design decisions are valued.

2 Assessing experiments with EDAT: framework and assessment tool

In this section, we introduce our framework for evaluating the quality of experimental designs. The framework is based on List (2025) and divides the design aspect of an experiment into five core dimensions: (i) *Setup*, (ii) *Recovery*, (iii) *Estimation*, (iv) *Interpretation*, and (v) *Generalizability & Scaling*. Each dimension is further divided into a grand total of 30 components. Each component captures an important design aspect and is constructed such that it can be assigned a quality score. Next, we explain in detail the framework’s scope and approach before describing the individual components and the scientific reasoning behind their construction.

2.1 Approach of the framework

Scope. In order to map the core aspects of experimental designs in a unified framework, we first need to specify its scope. We focus on experiments that explicitly assign subjects to at least two conditions to estimate treatment effects. The assignment must be done either randomly, capturing between- and mixed-subject experiments, or by having subjects experience all conditions sequentially, capturing within-subject designs. This focus excludes designs and papers that (i) use observational data, (ii) exploit naturally occurring variation (natural experiments), and (iii) are surveys that estimate distributions (e.g., the distribution of risk aversion in a population) instead of treatment effects. We choose this focus on treatment effect estimation because it then enables us to construct in particular the *Recovery* and *Estimation* dimensions in a straightforward and generalizable way. We discuss this and other limitations in the coverage of the framework further in Section 2.2.6.

Coverage. For experiments within our scope, we organize design features into five dimensions, which are further divided into 30 components. Figure 1 provides a graphical overview

Figure 1: Overview of the dimensions and components of the framework



of all dimensions and components. The five dimensions are *Setup*, *Recovery*, *Estimation*, *Interpretation*, and *Generalizability & Scaling*. The *Setup* dimension contains components capturing ethical responsibilities and other considerations before running an experiment. *Recovery* concerns how well an experiment tests for and addresses threats to the identification of an average treatment effect. This dimension can thus be understood as capturing internal validity. *Estimation* covers statistical power, multiple hypothesis testing, and the design of treatment cells. *Interpretation* covers the investigation of mediators and moderators of treatment effects. Lastly, *Generalizability & Scaling* focuses on the extent to which the treatment effect may generalize from the experimental setting to target settings of interest, as well as how well it can be replicated. As such, this dimension can be understood as capturing external validity.

Scoring. We designed each component to be scorable on an ordinal, four-point scale where **Gold** = 4, **Silver** = 3, **Bronze** = 2, and **Insufficient** = 1. We choose four points as a compromise between differentiation and complexity. A high number of scale points allows fine distinctions of quality, capturing even small differences within design components. At the same time, they increase the complexity of the assessment method, and for some components a high degree of differentiation may simply not exist. The four-point scale allowed us to create sufficient differentiation across all 30 components while maintaining a clear exposition and structure for maximum interpretability.

The grading scheme was designed around two key principles. First, each ordering was developed so that grades are interpreted *ceteris paribus*: we consider a grade of **Gold** to be better than **Silver**, holding all other grades and features of the experiment constant. That is, the grade of a specific component should only be interpreted as a measure of quality for this component, and not in itself a judgment about the quality of the overall experiment. Indeed, a higher grade on one component may be suboptimal for the overall quality of the experiment when it implies receiving worse grades on other component(s) that may be more important to the research question. The reason for this principle is that it allows researchers to draw their own conclusions and judgments about the quality of an experimental design given the individual component grades. That is, the framework does not suggest or imply a specific weighting or aggregation of individual grades. Whether achieving high grades in some components is more desirable than in others may depend on the setting and context the experiment is conducted in, as well as the research question the experiment is aiming to answer. Moreover, this principle allows us to assess *empirically* how grades are related to each other across components, as we discuss in more detail in Section 4. The second key principle is that each ordering was developed to be conditional on the research question. As explained further in Section 2.2.6, we do not attempt to score the relevance and importance of the research question or how well-suited the design is for answering this question.

The grading scheme is formulated for experiments that have already been conducted. However, this does not imply that the framework cannot be applied before an experiment is conducted. Researchers may apply the framework using expected outcomes or use it to explore alternative scenarios. For instance, the framework can help researchers plan for different attrition scenarios and identify standards that may improve the ethical conduct of an experiment. Thus, the grading scheme functions not only as an evaluation tool but also as a guide for improving experimental designs.

2.2 Components of the framework

Next, we cover and justify the components and the associated scoring in detail. Appendix Table A.1 provides an overview of all components and their scoring.

2.2.1 Setup

The *Setup* dimension covers the duties and ethical factors to consider before running an experiment: institutional review board approval, preregistration, and avoiding unnecessary deception and harm to subjects.

Harm to subjects and others. An experiment is potentially harmful if it exposes subjects, bystanders, and/or proctors to risks of physical, financial, social, and other types of harm that exceed those of ordinary life. We follow the principles established in the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979), namely that research should “do [no] harm” and otherwise “maximize possible benefits and minimize possible harms.” Since it is not always possible to eliminate excess risks of harm, our grades on this component consider whether (i) the risks imposed may lead to improvements in the quality of evidence obtained on the research question and (ii) the net benefit for affected parties is still positive. Accordingly, **Gold** is given in the absence of risk of extraordinary harm, **Silver** when the risk is beneficial in expectation and helps answer the research question, **Bronze** when the risk is not beneficial in expectation but helps answer the research question, and **Insufficient** when the risk is not helpful in answering the research question.

Deception. Deception refers to providing subjects with unequivocally false information, e.g., telling subjects that their negotiation partners are fellow participants when they are actually experimental confederates.¹ Our grades follow the norm against deception in economics (Charness et al., 2022): **Gold** is given when no deception is employed, **Silver** when such deception is used without the risk of harm, **Bronze** when the risk is minimal, and **Insufficient** if the risk is more than minimal.

Ethical duties. Before running an experiment, researchers are often expected to obtain (i) ethical approval from an institutional review board (IRB) and (ii) ex ante informed consent from subjects. Both practices are endorsed and sometimes required by the Common Rule (2018). Our framework endorses these practices as valuable tools for ensuring ethical standards are met and protecting the welfare and autonomy of subjects. In lieu of obtaining ex ante consent, an ex post briefing about the experiment can partially remediate the violation of autonomy caused by enrolling subjects in an experiment without informing them. Accordingly, **Gold** is given when both IRB approval and informed consent are obtained, **Silver** if IRB approval is obtained and an ex post briefing is provided, **Bronze** if IRB approval

¹We distinguish deception from incomplete disclosure, which is the practice of withholding potentially relevant information from subjects without making any clearly false statements. Since disclosing information about an experiment can harm the integrity of its findings (see, e.g., Levitt and List (2007) and De Quidt et al. (2018)), some degree of incomplete disclosure is arguably unavoidable in human subjects experimentation. We also distinguish deception from the practice of providing true but misleading information to subjects.

is obtained but neither informed consent is obtained nor an ex post briefing is provided (or if consent is obtained but IRB approval is not), and **Insufficient** if none is obtained.²

Preregistration. Another common experimental duty is preregistration, i.e., a commitment by authors to the details of their experiment before running it in order to limit their scope for *p*-hacking and other questionable research practices (QRPs). Preregistration is increasingly common and endorsed by journal policies (Imai et al., 2025). However, its effectiveness may depend on whether the authors register sufficient details to credibly tie their own hands (Abrams et al., 2020; Brodeur et al., 2024). Our framework therefore awards higher grades to papers with more detailed preregistrations and for which the experiment is implemented as preregistered. In particular, we check if (i) the expected sample size, (ii) treatment and outcome variables, and (iii) a hypothesis are preregistered and similar to what is reported in the paper. Grades of **Gold**, **Silver**, **Bronze**, and **Insufficient** are given to experiments that fulfill the criteria for all three, two, one, or none of these pieces of information, respectively.

2.2.2 Recovery

The *Recovery* dimension covers the core assumptions needed to recover an internally valid treatment effect: SUTVA, independence of treatment assignment, observability, and compliance.

Hidden treatment versions. One assumption required to recover a valid average treatment effect (ATE) is the stable unit treatment value assumption (SUTVA). Under SUTVA, it is assumed that the treatment is uniformly applied to all subjects, i.e., hidden treatment versions (HTVs) are ruled out. Treatments at higher risk of HTVs include those that (i) are inherently difficult to standardize between subjects, e.g., one-on-one therapy sessions, (ii) are delivered manually and separately by multiple proctors, (iii) are carried out in multiple sessions or waves, and/or (iv) manipulate multiple aspects of the environment at once. **Gold** is given to experiments for which the risk of HTVs is low (≤ 1 of these and other factors apply). When the risk is higher, authors can take preventive measures in the design stage, such as training proctors, giving them written protocols, recruiting balanced groups of subjects into each session, and/or ensuring balance between secondary features manipulated by the treatment (**Silver**). Alternatively, authors can take ex post steps like controlling for session fixed effects, clustering standard errors, and/or testing for between-session balance (**Bronze**). Designs that contain no such steps receive **Insufficient**.

Spillovers (threats to identification). A spillover occurs when the treatment status of treated units affects the outcomes of untreated units. Since the existence of spillovers is

²Capozza et al. (2026) show that the classification partially predicts how citizens and researchers judge the acceptability of ethical practices in experiments.

a violation of SUTVA, they are usually treated as threats to identification. Experiments at higher risk of spillovers include those in which (i) treated and untreated units can interact with each other before outcomes are measured, (ii) subjects work, study, or live with each other in the same natural environment, (iii) there is a considerable delay between treatment administration and outcome measurement, and/or (iv) the size of the treatment effect plausibly depends on how many others are treated via, e.g., externalities or general equilibrium effects. **Gold** is given to experiments for which the risk of spillovers is low (≤ 1 of these and other factors apply). When the risk is higher, authors who adjust their randomization procedures (e.g., cluster randomization) or econometric specifications (e.g., linear-in-means model) receive **Silver**. Authors who instead present descriptive or suggestive evidence about the severity of spillovers receive **Bronze**. Designs that contain no such steps receive **Insufficient**.

Spillovers (objects of interest). Spillovers can be objects of interest (instead of threats to identification) to researchers who are interested in externalities, social interaction effects, general equilibrium effects, and other types of interactions. Randomized saturation designs are the preferred choice in these cases. To maximize statistical power, authors should allocate groups to the minimum number of saturation levels sufficient to test their hypothesis (**Gold**). For linear, U-shaped, and S-shaped hypotheses, at least 2, 3, and 4 levels are required, respectively. Designs with other levels (**Silver**) sacrifice power. Alternative identification methods that do not randomize, like linear-in-means models, are given **Bronze**. If the researchers are interested in spillovers but spillovers are implausible, the experiment receives **Insufficient**.

Randomization method. Another assumption required for a valid ATE is the statistical independence of the assignment mechanism from subjects' potential outcomes, which is achieved in experiments via explicit randomization. One sign of a successful randomization is balance in baseline variables between the treatment and control groups. Stratified randomization (including matched pairs designs) is optimal for this purpose (**Gold**), as it effectively ensures balance in certain baseline variables and enhances statistical power. Both completely randomized experiments (**Silver**) and unit-by-unit randomization (e.g., Bernoulli trials, **Bronze**) result in balanced baseline variables in expectation, but the latter method is less likely to do so and also allows for imbalanced sample sizes between treatment and control groups.

Randomization control. Given the threats to identification that can emerge from the mis-implementation of randomization plans and the imperfect administration of assigned treatments, researchers ideally control these processes directly (**Gold**). In reality, they must often delegate control to external parties. If these processes are delegated or shared, (i) researcher oversight and (ii) within-paper documentation of how they were implemented

can compensate for the lack of direct control. By oversight, we mean the researchers' ability to observe and inspect the implementation as it is being conducted by the external parties, with the potential to request corrections and other changes. **Silver** is given if both oversight and documentation are provided, **Bronze** if one but not both is provided, and **Insufficient** if neither is provided.

Testing for compromised randomization. Under genuine random assignment, the number of significant differences in baseline variables between treatment and control groups is expected to fall within a range around $c \times \alpha$, where c is the number of baseline variables tested and α is the significance level. Researchers should conduct balance tests to check for unusual deviations from this range. Additionally, the distribution of p -values across the individual tests should be approximately uniform, which researchers can test for to gain additional evidence on the success of the randomization (**Gold**). Instead of conducting balance checks (**Silver**), researchers can run sample ratio mismatch tests or otherwise conduct (sometimes qualitative) checks for compromised randomization (**Bronze**), like surveys to confirm treatment assignment. **Insufficient** is given if no tests or checks are reported.

Accounting for compromised randomization. Under genuine random assignment, the number of significant differences should fall within the expected range (**Gold**). When this is not the case, authors will ideally be aware of the nature of the compromise, i.e., they will have knowledge of (i) subjects' assigned and realized treatment statuses and (ii) the non-random decision rule that was used to determine their realized statuses. When these conditions are satisfied, econometric methods, including difference-in-differences, propensity score matching, controlling for imbalanced variables, or estimating treatment effects on a balanced subsample of subjects, can address the compromise (**Silver**). When they are not, the above methods or modeling, e.g., a permutation test or decision theoretic approach as in Heckman and Karapakula (2021), can still be used but with additional assumptions and caveats (**Bronze**). **Insufficient** is given if no steps have been taken to account for compromised randomization despite failed balance checks and/or insufficient evidence on potential compromises.

Testing for attrition. Another assumption required for a valid ATE is observability, i.e., that the main outcomes are observed by the researchers for all subjects after treatment assignment³; violations of this are called attrition. If substantial attrition occurs (defined below as $< 95\%$ observability), researchers should investigate whether it occurred randomly or non-randomly with respect to subjects' potential outcomes. The most plausible way of testing this is to check whether attrition is related to baseline outcomes, as suggested by, e.g., Ghanem et al. (2026) (**Gold**). Researchers can alternatively run selective attrition or determinants of attrition tests to check whether any baseline covariates relate to attrition

³Attrition that occurs before this step may affect representativeness but does not threaten internal validity.

(**Silver**). Less direct approaches include the rate of attrition test and informal checks, such as running a survey among attritors to inquire about their reasons for attriting (**Bronze**). **Insufficient** is given if no checks for attrition are employed.

Nature of attrition. The most effective way of preventing attrition-related threats to identification is to minimize attrition altogether, ideally maintaining at least 95% observability for each condition (**Gold**). Short of that, random attrition according to statistical tests is preferable (**Silver**) to non-random attrition (**Bronze**). Finally, no reporting on attrition leaves the researchers and readers unable to fully assess the internal validity of the results (**Insufficient**).

Accounting for non-random attrition. In the presence of non-random attrition, one correction approach relies on instrumental variable (IV) estimation. Ideally, the IV will be strongly predictive of a subject's response probability but unrelated to the main outcome variable. The most plausible IV is a cross-randomized treatment that discourages attrition by providing, e.g., enhanced reminders or incentives to respond to endline surveys (**Gold**). Short of that, observational IVs, inverse probability weighting, and similar weighting methods are alternative options (**Silver**). A final approach is to construct bounds, e.g., Horowitz and Manski (2000) or Lee (2009) bounds, which recover ranges of point estimates under varying assumptions (**Bronze**). **Insufficient** is given if no correction method is applied despite the presence of non-random attrition.

Balanced panel assumption. The balanced panel assumption (BPA) is an extension of the observability assumption to panel data. In particular, the BPA requires that observability holds in every period. Similar methods of testing and accounting for non-random attrition to those documented above are available with panel data, and the grading thus follows the same structure.

Compliance. Another assumption required for a valid ATE is perfect compliance, i.e., that all subjects comply with and take up their assigned treatment. Ideally, at least 95% of subjects would comply (**Gold**). Experiments at higher risk of noncompliance include those in which (i) take-up occurs away from experimenter oversight, (ii) take-up requires substantial additional effort from subjects, (iii) take-up occurs with some delay after treatment assignment, (iv) the treatment is potentially accessible by other means, and/or (v) participants are aware of treatments that other participants receive. **Gold** is also given to experiments with a low risk of noncompliance (≤ 1 of these and other factors apply). In experiments with compliance issues, authors can estimate the intent-to-treat (ITT) effect instead of the ATE (**Silver**), or the local average treatment effect (LATE) and bounds on the ATE with additional assumptions. Estimating LATE requires *exclusion* (treatment assignment cannot affect the outcome except through treatment receipt) and *monotonicity* (assignment can only shift

the probability of treatment receipt positively). Bounds only require exclusion. If the authors wish to estimate LATE or bounds, they should acknowledge the required assumptions (**Silver**). Failure to do so earns **Bronze**, while failure to take any steps towards addressing noncompliance earns **Insufficient**.

Within-subject design. The two assumptions required to recover an ATE with within-subject designs are causal transience (treatments only have an effect in the contemporary period that dissipates before the next treatment begins) and temporal stability (there are no time-varying factors that affect participant outcomes, such as fatigue or period-specific shocks). Stealth within-subject designs are the most compelling choice, as experiments that go unnoticed are more likely to meet the challenging causal transience assumption (**Gold**). When stealth designs are infeasible, authors can instead randomize subjects to different orders of the treatment conditions in order to mitigate threats to the two assumptions (**Silver**). Short of that, authors should at least disclose and discuss the assumptions in their paper (**Bronze**); otherwise **Insufficient** is given.

Surrogacy. In order to recover an ATE while using proxies (surrogates) for the main outcome(s), the proxies should fully capture the causal link between the treatment variable and the actual outcome of interest. Since this is untestable and unlikely to hold exactly in practice, authors will preferably not use proxy variables (**Gold**). If they do, they should demonstrate that their proxies correlate with or are predictive of the outcome variables in observational data (or cite another work that does, **Silver**). Short of that, the authors should offer another justification for the validity of their proxies, e.g., based on theory or informal argumentation (**Bronze**). Otherwise, **Insufficient** is given.

2.2.3 Estimation

The *Estimation* dimension covers best practices for estimating the treatment effect: power analysis, multiple hypothesis testing, and the allocation of subjects to treatment cells.

Power analysis. Statistical power calculations can help authors lower the risk of type II errors by designing adequately powered experiments; for this purpose, they must be done ex ante with planned and estimated values (**Gold**). They can also assist readers with interpreting the credibility of the results. This is again more useful if done ex ante, as ex post calculations using observed effect sizes tend to overestimate power for significant results, especially false positives (Hoenig and Heisey, 2001; Gelman, 2019) (**Silver**). Even if power calculations are not made, authors can discuss the role of statistical power in their experiment, e.g., by noting that their abnormally large sample size mitigates concerns about power or by offering caveats about their estimation of interaction terms due to a small sample size (**Bronze**). **Insufficient** is given in the absence of reported power calculations or discussions.

Design of treatment cells. Under certain conditions,⁴ authors can reliably construct treatment cells and allocate subjects between them in order to maximize statistical power. Our grade considers two main criteria. The first is whether the number of treatment cells is parsimonious given the form of the hypothesis of interest. For linear, U-shaped, and S-shaped hypotheses, two, three, and four treatment cells are the minimum numbers of cells necessary to test the hypotheses, respectively. The second is whether subjects are allocated optimally between the cells. In the case of a linear hypothesis and two cells, the optimal allocation is 50% in each cell (50/50).⁵ For a U-shaped hypothesis and three cells, it is 25% in the cells with the lowest and highest treatment intensities and 50% in the cell with medium intensity (25/50/25). For an S-shaped hypothesis, it is 17/33/33/17. These allocations maximize power for a fixed sample size or a fixed budget when per-participant costs are equal across treatment cells. Experiments that meet both criteria receive **Gold**; those that only meet the first receive **Silver**. Experiments that fail the first criterion by having more cells than necessary receive **Bronze**, while those that have insufficiently many to test the hypothesis receive **Insufficient**.

Multiple hypothesis testing. When authors run multiple hypothesis tests (MHTs) from a family of hypotheses without adjusting the significance threshold α , the probability of observing at least one false positive rises with the number of tests. Our grades on this component therefore consider whether the authors control the family-wise error rate (FWER) via MHT adjustments. The Bonferroni correction controls the FWER but does not adjust for dependence between tests, which can result in overly conservative FWERs (**Bronze**). The Holm stepdown correction procedure alleviates this conservatism by iteratively adapting the significance threshold used (**Silver**). Methods that directly account for between-test dependence, including Romano and Wolf (2005, 2016) and List et al. (2019), mitigate the problem most effectively (**Gold**). **Insufficient** is given if no correction for MHT is applied.

2.2.4 Interpretation

The *Interpretation* dimension covers the methods used to expand the interpretation of the treatment effect using moderators and mediators.

Mediation. The most direct and credible approach to gathering evidence about a mediator is to randomize or intervene on it. This can be done by, e.g., introducing treatment arms that include or exclude certain channels alongside the full treatment arm or by running sep-

⁴The first condition is that the outcome variable is continuous. The second is that outcome variances are equal across treatment cells. The third is that, regardless of the number of treatment cells, the intensity of the treatment variable must be interpretable on a continuous scale. Examples of treatments like this include monetary payments and hours of training (both on $[0, \infty)$ scales). Our grades do not apply if the outcome and/or treatment intensity are ordinal, categorical, or binary in nature.

⁵When grading papers, we check whether allocations are approximately, rather than exactly, optimal, e.g., that each cell contains between 40% and 60% of subjects in the 50/50 case.

arate mediation experiments (**Gold**). Alternatively, authors can measure mediator variables of interest and attempt to quantitatively decompose the main treatment effect of interest via, e.g., structural modeling or controlling for the mediator in a regression alongside the treatment (**Silver**). Interpreting these analyses causally relies on additional assumptions. A final approach is to gather indirect, qualitative, or correlational evidence about a mediator variable, e.g., showing correlations between the outcome, the treatment variable, and the mediator, or surveying subjects to inquire about how a treatment may have affected them (**Bronze**). **Insufficient** is given if mediators are reported or discussed without evidence.

Moderation with pre-treatment covariates. The most direct and credible approach to gathering evidence about a moderator is also to randomize or intervene on it. This can be done by introducing treatment arms that cross-randomize the treatment with different values of the moderator or by running a separate moderation experiment (**Gold**). Alternatively, methods like causal forests exist to assist authors with estimating conditional average treatment effects (CATEs) (Athey and Imbens, 2016; Wager and Athey, 2018). These approaches make it feasible to leverage information on several covariates simultaneously and do not rely on particular functional form assumptions (**Silver**). Finally, authors can use classical techniques like linear regression or regressions on different subsets of the data to estimate CATEs (**Bronze**). **Insufficient** is given if moderators are reported or discussed without evidence.

Moderation with extensive and intensive margins. Decomposing treatment effects into extensive and intensive margin effects provides additional evidence that readers and policy-makers may find useful. This is only possible when the outcome variable is bounded from below at 0 and 0 represents non-participation, as with, e.g., hours worked per week in a sample of workers. Tobit models can be used to provide this decomposition, but authors should acknowledge the stringent assumptions required to use them (**Gold**). An alternative approach is to use baseline outcome data and difference-in-differences to measure the intensive margin effect among subjects with positive outcomes both before and after treatment. Other approaches include estimating bounds on the extensive and intensive margin effects and comparing pre- and post-treatment means of the outcome conditional on participation, which rarely identifies the intensive margin effect. If the baseline-data approach is used and the assumptions are discussed, **Silver** is given; if bounds are estimated, or if the baseline-data approach is used without a discussion of assumptions, **Bronze** is given. **Insufficient** is given if margins are reported or discussed without evidence.

2.2.5 Generalizability & Scaling

The *Generalizability & Scaling* dimension covers the topic of whether the treatment effect generalizes and is scalable to key settings of interest, assessed using the four SANS criteria (List, 2020).

Selection (SANS). The first criterion is *selection* into the experiment, which limits the generalizability of experimental findings because self-selecting subjects may not be representative of the population. As such, it is preferable when selection into the experiment is not possible, i.e., when the experiment is implemented stealthily and without scrutiny in the subjects' natural environment (**Gold**). Authors can also earn **Gold** by demonstrating that their sample is statistically representative of the population of interest. If the sample is unrepresentative, authors will ideally report on which variables the sample differs from the population (**Silver**). If no information on representativeness is disclosed, authors should acknowledge potential selection issues as a caveat (**Bronze**); otherwise **Insufficient** is given.

Attrition and compliance (SANS). The second SANS criterion is *attrition and compliance*. One common reason experimental findings fail to generalize from the subject setting to the target setting is that attrition and compliance rates often change between these settings. As such, it is preferable that the experiment is already conducted with subjects from the target population in their natural environment, so that attrition and compliance rates will not differ (**Gold**). Short of that, authors should explicitly report attrition and compliance rates from the subject setting and discuss how they may differ in the target setting. If both are discussed, **Silver** is given; if one is discussed, **Bronze** is given; if neither is discussed, **Insufficient** is given.

Naturalness (SANS). The third SANS criterion is *naturalness*. Experimental findings can also fail to generalize due to experimenter scrutiny and construct validity issues in the subject setting that make it unrepresentative of the target setting. These concerns are mitigated or eliminated when the experiment is run covertly with target subjects in their natural environment (**Gold**), as experimenter scrutiny is minimal and outcomes can be measured directly rather than via artificial constructs. Other options include running the experiment overtly in the natural environment (**Silver**) or conducting it in an artificial setting. If the authors choose an artificial setting, they receive a grade of **Bronze** if they acknowledge experimenter scrutiny and construct validity issues as potential caveats, and **Insufficient** otherwise.

Scalability (SANS). The final SANS criterion is *scalability*. This criterion is most relevant when the authors are testing an intervention that is intended to improve policy-relevant outcomes. It intends to answer the following question: "how strong is the evidence provided by the experiment that the intervention would be similarly effective if scaled up to the natural setting of interest and/or to a significantly expanded number of treatment recipients?" We identify seven main features of experiments that enhance their scalability: (i) selection into the experiment was not possible or the authors demonstrate sample representativeness; (ii) the experiment was conducted with subjects from the target population of interest; (iii) the experiment was conducted in the target natural environment of interest; (iv) the interven-

tion does not rely on access to skilled personnel or scarce resources; (v) the intervention does not require many staff or multiple days to implement; and, when the number of treatment recipients is significantly increased, neither (vi) negative general equilibrium effects nor (vii) negative social spillovers are likely to occur. **Gold** is given if at least 5 criteria are met, **Silver** if 3–4 are met, **Bronze** if fewer than 3 are met and this is mentioned as a caveat. Otherwise, **Insufficient** is given.

Replicability. The replicability of an experiment is essential to its credibility and generalizability. Authors can provide strong evidence on replicability by conducting a replication of the main experiment and presenting replication results within the same paper (**Gold**). This could be an exact, disparate-population, or conceptual replication. Short of that, an experiment receives a **Silver** if it is sufficiently accessible for others to conduct a replication of it. The documentation requirement is satisfied if the authors disclose sufficient information about the subjects, the experimental environment, and the nature of the treatment for someone with appropriate access to re-implement it. The second requirement is satisfied as long as none of the following features are present: treatments or settings that (i) require special permission to randomize or (ii) are expensive or require special technologies; (iii) administrative or proprietary data; and research questions that (iv) require multi-year experimental timeframes or (v) can only be answered with participants from specific regions or cultures. Experiments that meet the first requirement, but not the second, receive **Bronze**.⁶ Those that fail the first requirement receive **Insufficient**.

Availability of materials. Releasing the code and data necessary to reproduce experimental results enhances the credibility and reproducibility of an experiment. Similarly, releasing (as relevant) the materials given to experimental subjects, the administration protocols used by proctors, and documentation of the experimental procedures enhances readers' understanding of how it was conducted. Ideally, these items (code, data, materials) are available (**Gold**). If one or more of these items are missing due to confidentiality reasons (e.g., an agreement with an external party restricting disclosure) or are available upon request, the authors should disclose that information (**Silver**); otherwise, **Bronze** is given. If all items are missing, **Insufficient** is given.

2.2.6 Dimensions not currently covered

Our aim in developing this framework was to be as comprehensive as possible in characterizing measurable and relevant features of experimental designs while keeping the overall length manageable. Naturally, this means not all conceivable components of experiments are included, as there is a trade-off between compactness and coverage. Therefore, we build the

⁶If the documentation requirement fails because the authors have a confidentiality agreement with an external party that restricts disclosure of the missing information, **Bronze** is still granted.

framework modularly, which allows researchers to readily adapt the framework to their purposes, excluding or expanding on dimensions and components as they see fit. As such, the framework can be seen as a “living document” in that future versions may feature new components as well as modifications to existing ones. Additional possible components can be separated into those that are potentially worth adding in a future version and those that we exclude deliberately.

Components that we conceive to be worth adding in the future include:

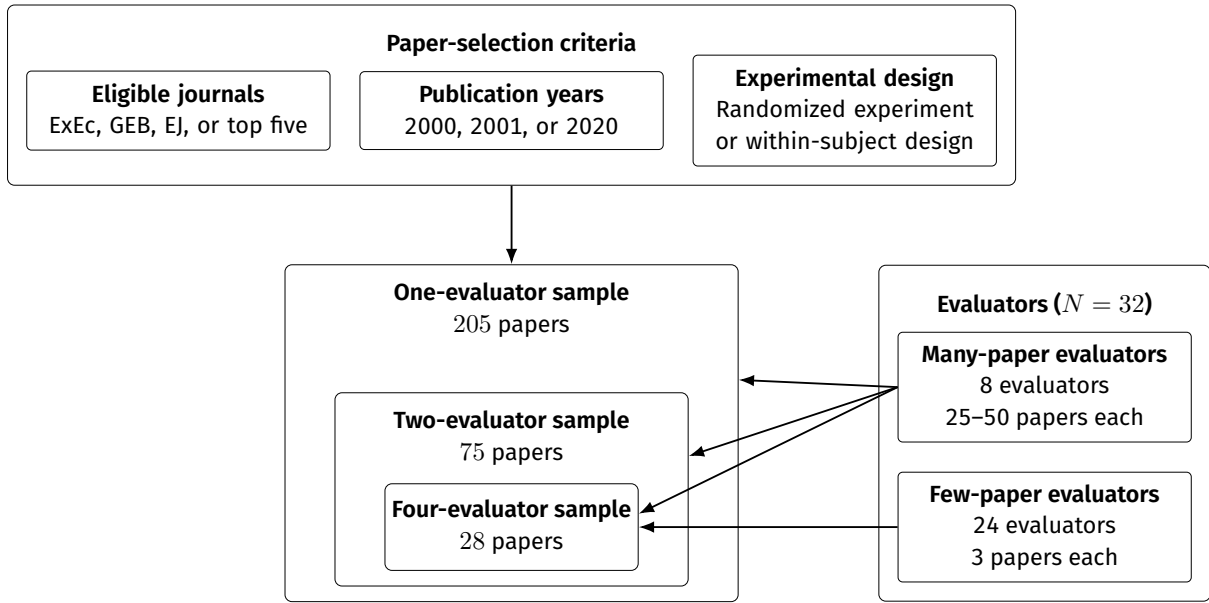
- The construct validity of certain commonly elicited constructs, such as risk preferences, which can be elicited, e.g., using either self-assessment or incentivized lottery choices.
- The theory relevance of an experiment as a counterpart to policy relevance.
- The adequacy with which certain concerns about experiments, such as experimenter demand effects and the strength of incentives, are dealt with in a particular design.

At the same time, there are also components that we deliberately excluded. These include the relevance of the research question and whether the design is appropriate to test it. While obviously crucial for the overall assessment of an experimental *paper*, we exclude these components because we restrict our framework to the evaluation of experimental *designs*. Furthermore, we assess that grading them would require a subjective, value-driven assessment that is out of step with our current approach. Moreover, these judgments are potentially highly paper-specific, which would likely make it infeasible to find a scoring system that is widely applicable and flexible enough for the circumstances of each individual paper. Indeed, arguments have been made that such questions cannot be objectively answered (Weber, 1919) and that there is a plurality of opinions in the profession about which questions are “worth asking” (Andre and Falk, 2025).

2.3 Assessment tool

Given the breadth of the framework and our goal of providing a scalable method of generating and interpreting experimental evidence, we developed a survey tool that a researcher can use to apply the framework. The tool guides its user step by step through an experimental design, systematically covering all of the framework’s components. For every component, evaluators receive background on the relevant concepts and answer targeted questions. After each block of questions, the survey automatically generates a grade for a component and presents it to the user. Users are simultaneously given the definitions and criteria needed to earn each component grade. With this information, users can manually rectify any coding mistakes they made by suggesting revisions to the generated grade for each component. They are given another opportunity to do this at the end of the survey. For an example, see Appendix Section C. There, we show the questions and flow that lead to the grade for the *Nature of attrition* component.

Figure 2: Mapping between papers and evaluators



Notes: This figure summarizes the structure of the dataset. Papers enter the sample if they meet our paper-selection criteria. These criteria yield 205 papers, each of which was evaluated at least once (*One-evaluator sample*). Nested subsamples were evaluated by two (*Two-evaluator sample*) or four evaluators (*Four-evaluator sample*). Evaluations were performed by *Many-paper evaluators*, who each graded 25–50 papers, and *Few-paper evaluators*, who each graded 3 papers. Arrows indicate which evaluator group graded which subsample.

To make applying the tool as efficient and standardized as possible, the tool has two key components. First, the tool contains follow-up questions and conditional statements, which appear based on the user input to the initial questions. This simplifies going through the framework, as for instance some components depend on earlier input and might be skipped. For instance, the component *Testing for attrition* is only relevant if for the component *Nature of attrition* a lack of full observability is given as an answer. Second, each question in the survey comes with a “More information” button. Clicking the button reveals extensive practical and conceptual guidance for the user to answer the respective question. For instance, the guidance contains definitions of the core concepts of the question as well as examples to help understanding. Importantly, this feature, together with the fact that we spent extensive efforts standardizing the initial questions, helps ensure that the evaluations resulting from the tool are as objective and standardized as possible. This will ideally minimize measurement error and maximize the predictive potential of our grades. For use more broadly, these features also make it feasible for students and researchers with sufficient background knowledge (but not necessarily much experience with experiments) to generate systematic quality evaluations.

3 Applying EDAT to published experiments

To assess the performance of EDAT, we assembled a sample of 205 experimental papers and had 32 researchers evaluate them using the assessment tool. Figure 2 provides an overview of the paper sample, the evaluators, and the mapping of evaluators to papers.

3.1 Sample of papers

To assemble our sample of papers, we first collected information on experimental papers published in economics journals. We selected papers using the following three criteria:

1. The paper must have been published in one of two top experimental field journals (*Experimental Economics/ExEc* or *Games and Economic Behavior/GEB*), in a particular general interest journal (*Economic Journal/EJ*), or in a top-five journal, namely the *American Economic Review (AER)*, *Econometrica (ECMA)*, the *Journal of Political Economy (JPE)*, the *Quarterly Journal of Economics (QJE)*, or the *Review of Economic Studies (REStud)*. This allows us to investigate whether different journals prioritize different components of quality.
2. The paper must have been published in 2000, 2001, or 2020. The year 2020 is recent enough to characterize contemporary practices in experimental economics while allowing sufficient time for meaningful outcomes, such as citation counts, to have materialized. We chose 2000 and 2001 to assess time trends.
3. The paper must have at least one experiment that fulfills our framework criteria: explicitly randomizing participants into different conditions (between- and mixed-subject designs) or allowing them to experience all conditions (within-subject designs) with the goal of estimating treatment effect(s) (see Section 2).

Applying these criteria yielded a final sample of 205 papers, including 65 papers from *ExEc*, 47 papers from *GEB*, 24 papers from *EJ*, 25 papers from *AER*, 16 papers from *QJE*, 13 papers from *JPE*, 8 papers from *ECMA*, and 7 papers from *REStud*. See Appendix Table D.1 for the list of papers coded.⁷

3.2 Evaluation procedure

We recruited two groups of evaluators to apply the survey assessment tool to the sampled papers. The groups differed in both the number of evaluators and the number of papers assigned to each evaluator. Figure 2 maps evaluators to papers. We describe the two groups and the evaluation procedure below.

⁷Among the 205 papers are six which are co-authored by one of the authors of this paper. Our results stay the same if we exclude those papers.

Many-paper evaluators. To code the full sample, we hired 8 *many-paper evaluators*. Given the judgmental nature of the task and the number of components to grade, we recruited PhD students with some experience or interest in experimental economics. Five evaluators were based at the University of Bonn, two at the University of Chicago, and one at Erasmus University Rotterdam. Each evaluator graded between 25 and 50 papers.

Before beginning the main coding task, evaluators attended training sessions to familiarize themselves with the framework and survey tool. Each evaluator was asked to read a set of chapters from List (2025) before the orientation session. During the orientation, we presented an overview of the framework and gave evaluators an opportunity to ask questions. Over the following weeks, each evaluator coded 3 training papers using the survey tool. In a second training session, each evaluator received both general and individualized feedback on the quality of their coding.

Each evaluator received a flat payment for coding their assigned papers. Payment did not vary with either the number of papers completed or the speed of completion, because we wanted evaluators to take the time needed to assess each paper carefully. Evaluators knew that we would analyze their responses for reliability and that payment was contingent on reasonable performance. All evaluators received the full payment.

Once evaluators became familiar with the survey and framework, we estimated that reading and evaluating a paper would take approximately 3 hours. The median survey duration was 3.7 hours⁸ and efficiency gains were substantial: median survey durations declined by a median of 21% between the first and second halves of each evaluator’s set of papers. We demonstrate later that these gains did not come at the expense of reliability.

We intentionally assigned 75 papers to two independent evaluators, which allows us to calculate inter-evaluator reliability and agreement in this subsample. In an additional coding exercise, we asked six evaluators to act as a “third evaluator” and resolve disagreements in the twice-coded papers. For each paper, the third evaluator received a sheet summarizing the components on which the initial two evaluators disagreed. Each third evaluator reviewed 12 or 13 papers, determined final grades for disputed components, and provided a short justification for each decision. We also asked them to review any manually suggested revisions and comments left by the initial evaluators. We use these third-evaluator-resolved grades in all analyses after the reliability section. For two main analyses (Table 1 and Figure 5), we additionally conduct robustness checks using the average component grade across the two evaluators rather than the third-evaluator-resolved grade; these are reported in Appendix Table B.2 and Figure B.9.

Few-paper evaluators. To assess the usability of the survey tool without training and its reliability when more than two evaluators provide grades, we recruited 24 *Few-paper eval-*

⁸It is unclear whether this is an over- or under-estimate of the actual time taken per paper. The 3.7 hours includes time that the evaluators spent on unrelated activities after having initially opened the survey window. However, it excludes the time taken to read papers before starting a new survey session.

uators, each of whom coded 3 papers (Figure 2). These evaluators were Predocs (7), PhD students (13), and Postdocs (4) from 17 different institutions. Because papers in economics are typically reviewed by between two and four referees (Card and DellaVigna, 2020), we assigned papers to *Few-paper evaluators* to obtain 28 papers with four independent evaluations (Figure 2). *Few-paper evaluators* did not participate in a training session, receive feedback, or receive payment for applying the tool.

3.3 Variables

This section introduces the non-grade variables used in the analyses in Section 4. Table A.2 reports summary statistics for all variables.

Proxies for impact and quality. We collect several variables that measure, or serve as proxies for, a paper’s impact and quality:

- **Citation count:** We collected citation counts from Google Scholar for all papers on January 21, 2026. To address the long right tail of the citation distribution, we use $\log(\text{citations})$ in our analyses. Only one paper had zero citations and is therefore dropped from these analyses.
- **Publication outlet:** All coded papers were published in one of eight economics journals. For interpretability, we classify each paper’s journal as top five (1) or non-top-five (0).
- **Syllabus inclusion:** We collected these counts from Open Syllabus on March 2, 2026. Open Syllabus tracks the number of times each paper appears in a sample of millions of university syllabi (Open Syllabus, 2026). For interpretability, we classify papers according to whether they appear in at least one syllabus (1) or none (0).
- **Altmetric score:** We collected these scores from Altmetric on May 10, 2026. In addition to citations, Altmetric tracks alternative measures of a paper’s impact and engagement, including mentions in news media, blog posts, and social media (Altmetric, 2024). It compiles these metrics into the score recorded in this variable. To address the long right tail of the Altmetric-score distribution, we use $\log(1 + \text{Altmetric scores})$ in our analyses. We add 1 before taking the logarithm because 12 papers have an Altmetric score of zero.
- **Policy mentions:** We collected these counts from Altmetric on May 10, 2026. The number of times a paper is cited in a policy document is one of the variables Altmetric tracks (Altmetric, 2024). To address the long right tail of the policy-mention distribution, we use $\log(1 + \text{Policy mentions})$ in our analyses. We add 1 before taking the logarithm because 65 papers have zero policy mentions.

Covariates and moderators.

- **Number of authors:** we count the number of co-authors in the published version of the paper. This is used as a control variable to proxy for author team resources and manpower.
- **JEL code:** we record the paper’s JEL codes from the published version (if available) or from an older, working paper version if not. We could not find JEL codes for 8 papers. In these cases, we assigned JEL codes based on the topic of the paper.
- **Year of publication:** we record the paper’s year of publication. 32 papers were published in 2000, 33 were published in 2001, and 140 were published in 2020.

4 Empirical evidence from applying EDAT

Applying EDAT to our sample of published papers reveals substantial heterogeneity in grades, both within and across dimensions. Appendix Figure A.1 displays density plots of the grades, and Appendix Table A.2 reports the corresponding summary statistics. Given the extensive variation in grades, we start our analysis by assessing whether they are valid measures of design quality. We then use the grades to explore how researchers make design decisions.

4.1 Assessing the validity of EDAT

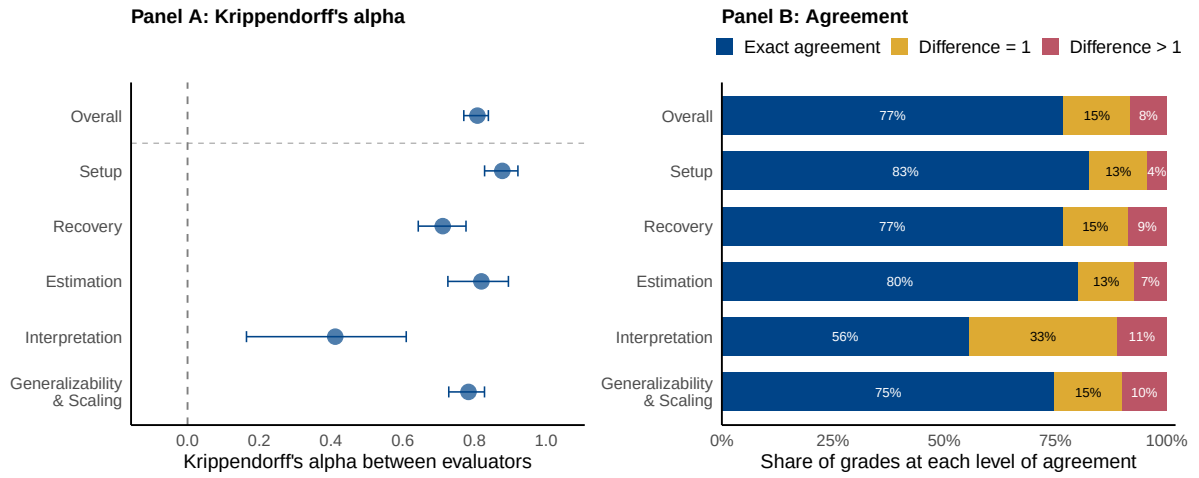
To assess the validity of EDAT grades, we focus on two questions: first, whether EDAT produces *reliable* quality scores; second, whether these scores predict outcomes linked to quality. Positive answers to both questions would indicate that EDAT scores are reliable and relevant.

4.1.1 Reliability

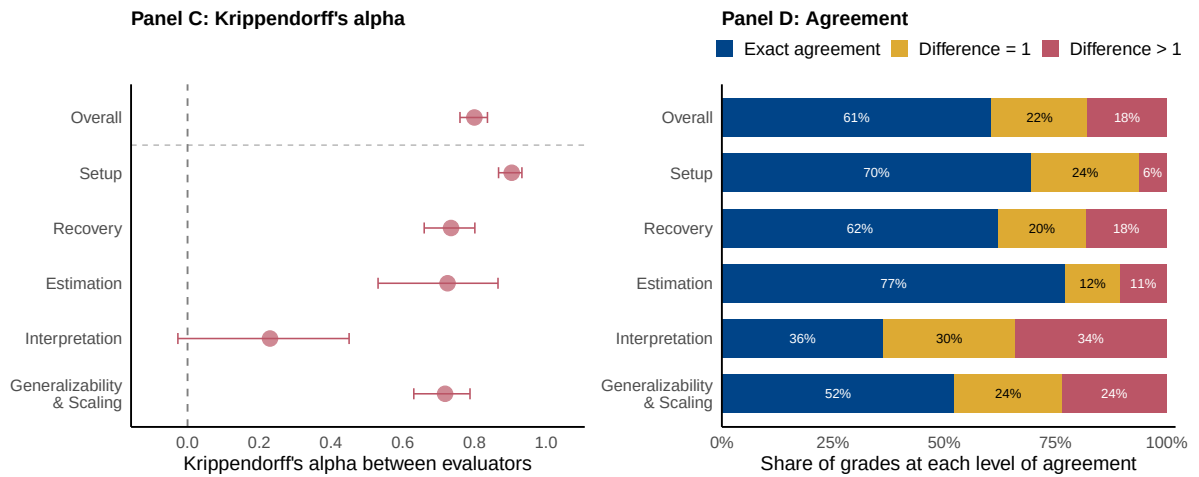
Measuring reliability. We assess reliability by testing whether evaluators agree on quality scores. High agreement indicates that the assessment is relatively objective, whereas low agreement suggests that the underlying concept, the assessment guidelines, or both involve substantial subjective judgment. Following the measurement-validity literature, we use two measures. The first is Krippendorff’s α (Krippendorff, 2011), a continuous scale where $\alpha = 1$ means that evaluators’ scores perfectly agree, $\alpha = 0$ means that agreement is no better than expected by chance, and $\alpha < 0$ indicates systematic disagreement in evaluators’ scores. The second distinguishes between exact agreement, disagreement by one grade level, and disagreement by more than one grade level.

Figure 3: Inter-evaluator reliability across samples and dimensions

Two-evaluator sample



Four-evaluator sample



Notes: This figure reports the inter-evaluator reliability of EDAT scores, i.e., the extent to which independent evaluators assign the same grade when scoring the same paper. Each evaluator grades the applicable components of a paper on a four-point scale from 1 (Insufficient) to 4 (Gold). Panels A and B show reliability in the *Two-evaluator sample*: the 75 papers each graded by two *Many-paper evaluators*. Panels C and D show reliability in the *Four-evaluator sample*: the 28 papers each graded by four evaluators (two *Many-paper evaluators* and two *Few-paper evaluators*). Reliability in Panels A and C is assessed using Krippendorff's α , computed over the component grades of all evaluators of a paper. The bars in these panels are 95% confidence intervals obtained via bootstrapping clustered at the paper level. In Panels B and D, reliability is assessed using agreement shares, which report the fraction of graded components on which the evaluators' grades coincide exactly, span one grade level, or span more than one grade level.

Results on reliability. EDAT scores display high agreement across both evaluator samples. In the *Two-evaluator sample* (Panels A and B of Figure 3), the overall α is 0.81 when pooling across dimensions. The exact-agreement statistics tell a similar story: evaluators assign the same grade in 77% of cases, disagree by one grade level in 15%, and disagree by more than one grade level in only 8%. Reliability is broadly similar across dimensions, with the exception of the *Interpretation* dimension.⁹ In Panels C and D, we analyze the subsample

⁹A closer examination of the individual components of *Interpretation* reveals two specific sources of subjectivity that contribute to its lower reliability and are not relevant to other dimensions. First, we occasionally

of papers that received four evaluations per paper. We find a very similar overall α of 0.80, as well as similarly high values among the five dimensions individually: none of the pairwise differences in α between the *Two-evaluator* and *Four-evaluator* samples is statistically significant at conventional levels ($p > 0.10$ in all cases). With respect to agreement shares, we find that all four evaluators agree on the grade in 61% of cases, while in 22% of cases the disagreement among all four spans one grade. These agreement shares are significantly lower both overall and within individual dimensions than those in the *Two-evaluator* sample (all pairwise differences have $p < 0.05$ except for *Estimation*). Lower agreement shares are expected because agreement among all four evaluators constitutes a stricter criterion than agreement between two evaluators.

These reliability levels meet conventional benchmarks. In content analysis, measures that achieve $\alpha \geq 0.80$ on categorical data such as ours are considered “reliable” (Krippendorff, 2004) and constitute an “almost perfect” strength of agreement (Landis and Koch, 1977). Thus, our scoring procedure reliably implements the framework. For context, Card and DellaVigna (2020) and Welch (2014) report inter-referee correlations of 0.15 to 0.30 at leading economics journals.¹⁰ This comparison should not be interpreted as a benchmark: referee recommendations reflect factors beyond experimental design quality, including the importance of the research question and the paper’s fit with the journal. Instead, it illustrates that focusing on the subdimension of experimental design quality can yield substantially higher agreement, even when the number of evaluators rises from two to four.

Robustness of reliability analysis. So far, we have compared the *Two-evaluator* paper sample with the *Four-evaluator* sample. However, this comparison changes both the sample of papers and the composition of evaluators. To isolate differences between evaluator groups, in Appendix Figures B.1 (for Krippendorff’s α) and B.2 (for the level of agreement) we directly compare the reliability of *Many-paper evaluators* and *Few-paper evaluators* on the same subset of 28 papers that were evaluated by both groups. We find that both reliability measures are similar across the two evaluator samples. For instance, while *Many-paper evaluators* agree with each other in 79% of cases for papers in the *Four-evaluator* sample (Appendix Figure B.2, Panel C), *Few-paper evaluators* agree in 76% of cases (difference in agreement shares, $p > 0.10$; Appendix Figure B.2, Panel D).

In addition, our analysis so far has excluded instances of missing data, i.e., cases where at least one evaluator classified a component as not applicable. As a robustness check, we

observe disagreement about whether a specific secondary variable studied in the paper was a mediator or moderator for the main relationship of interest. Second, papers often discuss several potential mediators and moderators and provide varying degrees of evidence for each. As such, evaluators sometimes correctly answered questions about the analyses of mediators and moderators but had different secondary variables in mind.

¹⁰Card and DellaVigna (2020) and Welch (2014) use Pearson correlation coefficients (as well as Spearman coefficients and agreement measures) to assess the inter-referee correlations. Appendix Figure B.4 displays Pearson correlations for our samples, and yields results very similar to those from our main analysis using Krippendorff’s α .

recalculate the agreement shares after introducing a fourth category, “Graded/NA”, i.e., components that at least one evaluator classified as not applicable while at least one other classified them as applicable. Cases in which all evaluators classify a component as not applicable are counted as exact agreement. With the exception of the interpretation dimension, we find low rates of disagreement over whether a component is applicable or not; see Appendix Figure B.3. For instance, in the *Two-evaluator* sample, evaluators disagree overall in 7% of cases on whether a component is applicable or not. For the interpretation dimension, we find higher shares of disagreement, especially in the *Four-evaluator* sample. These higher disagreement shares are mainly driven by disagreement over whether variables are treated as mediators or moderators in a given paper.

Heterogeneity in reliability. Does reliability decrease as paper complexity increases? To assess this, we calculate each paper’s number of pages and its readability, measured by the Flesch–Kincaid grade level. In univariate regressions of reliability on these measures, we find that longer and less readable papers are associated with lower agreement; see Appendix Table B.1. However, these associations appear to reflect differences across fields, as they disappear once we control for JEL codes.

Next, we investigate the role of evaluator experience. As our *Many-paper evaluators* each evaluated at least 25 papers, we can examine whether reliability increases with evaluator experience. Specifically, we split the papers evaluated by the *Many-paper evaluators* according to whether each was among the first 13 or last 12 papers evaluated. We find that α increases from 0.79 to 0.83, while the share of exact agreement increases from 75% to 79%. As these increases appear to be quantitatively quite small and are not statistically significant ($p > 0.10$), experience does not seem to play a major role in driving reliability. This is consistent with our findings that reliability does not differ much between *Many-paper evaluators* and *Few-paper evaluators*.

To summarize, we find high values of α and high agreement shares across dimensions and samples, with the exception of the *Interpretation* dimension. The high levels are remarkable insofar as our evaluators are early-career researchers and, in the case of the *Few-paper evaluator* sample, had received neither training in the evaluation tool nor extensive exposure to it. Nevertheless, evaluators agree on quality scores, indicating the reliability of the quality scores of our assessment tool.

Result 1. *Quality scores produced by EDAT display a high level of reliability between independent evaluators.*

4.1.2 Predictive power

Measuring quality proxies. Reliability alone does not establish that EDAT scores capture meaningful variation in design quality. Indeed, while our results indicate that evaluators

Table 1: Predicting quality proxies using EDAT

	Top five (0/1)		log(Citations)		Syllabus (0/1)		log(1 + Altmetric score)		log(1 + Policy mentions)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Average grade	0.58*** (0.11)	0.44*** (0.15)	0.74* (0.38)	1.55*** (0.43)	0.36*** (0.12)	0.40*** (0.14)	2.57*** (0.39)	1.60*** (0.44)	1.45*** (0.38)	0.83** (0.39)
Mean dep. var.	0.34	0.34	4.37	4.37	0.41	0.41	2.39	2.39	1.23	1.23
Mean indep. var.	2.62	2.62	2.62	2.62	2.62	2.62	2.65	2.65	2.65	2.65
R^2	0.11	0.42	0.02	0.52	0.04	0.39	0.21	0.46	0.09	0.56
N	205	205	204	204	205	205	181	181	181	181
JEL code FE		✓		✓		✓		✓		✓
Year FE		✓		✓		✓		✓		✓
log(No. authors)		✓		✓		✓		✓		✓

Notes: This table reports OLS estimates of the relationship between proxies for scientific quality and EDAT scores. The dependent variable in columns (1-2) is an indicator for whether the paper was published in a top-five instead of a non-top-five journal; in columns (3-4), it is the logarithm of the paper's total citations; in columns (5-6), it is an indicator for whether the paper was included in a course syllabus; in columns (7-8), it is the logarithm of one plus the paper's Altmetric score; and in columns (9-10), it is the logarithm of one plus the number of times the paper was mentioned in a policy document. The independent variable is a paper's average EDAT grade across all components. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Controls are log(Number of authors), publication-year fixed effects, and JEL-code fixed effects. Robust standard errors in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

agree on scores, these scores might not capture genuine underlying dimensions of quality but features that are unrelated to quality. We therefore examine whether the scores *predict* outcomes that the economics profession regards as proxies for quality: (i) publication in a top-five rather than a non-top-five journal; (ii) the logarithm of total citations; (iii) inclusion on a course syllabus; (iv) the logarithm of one plus the paper's Altmetric score, which captures attention from the research community and the public; and (v) the logarithm of one plus the number of times a paper was mentioned in a policy document.

These proxies do not perfectly capture a paper's quality or uniquely reveal experimental design quality. Other factors, such as the relevance and importance of the research question, may be stronger predictors of these outcomes. Nevertheless, because design quality is an important component of overall paper quality, EDAT scores should be significantly associated with these outcomes if they capture meaningful variation in quality.

To test these predictions, we use a paper's average grade across all components. This average should not be interpreted as a normative measure of design quality: such an interpretation would require component grades to be comparable and equally important determinants of design quality, an assumption that is too strong and was not a goal of our setup. Instead, the average grade provides a first approximation to an underlying quality dimension, allowing us to examine whether it predicts scientific outcomes across papers.¹¹ The analysis is therefore aggregate rather than a normative statement about individual papers. We also use alternative weighting schemes to assess the sensitivity of the results to different ways of aggregating grades.

¹¹One caveat is that evaluators were not blinded to author identities or publication outlets. As such, we cannot rule out that halo effects associated with these features partly drive correlations between evaluators' scores and their associations with scientific outcomes. However, since we asked evaluators to answer questions based on clear criteria rather than directly eliciting perceived grades, the severity of this issue should be minimal. We did not uncover any evidence of this issue during the initial training sessions or in any subsequent reviews of the final data.

Results on predictive power. Table 1 shows the results. We find that a paper’s average grade significantly predicts all of our outcomes. That is, a higher average grade is associated with a higher likelihood that the paper is published in a top-five journal and mentioned in course syllabi, as well as with higher citation counts, Altmetric scores, and mentions in policy documents. This result holds both in univariate regressions and in specifications that include publication-year and JEL-code fixed effects and control for the number of authors. We include these variables to control for the fact that papers published in 2000 and 2001 had substantially more time for these outcomes to materialize¹² and to account for field-specific citation norms and authorship effects. The associations are also quantitatively large: an increase in the average grade from the median of 2.57 to the 75th percentile of 2.79 is associated with a 13.15-percentage-point increase in the likelihood of the paper being published in a top-five journal, an 18% higher citation count, and a 44% increase in policy mentions.

Robustness of predictive power analysis. When we use a paper’s average grade, each component receives the same weight. Because some dimensions contain more components than others, they mechanically receive greater weight. As a robustness check, we first calculate the average grade within each dimension and then average these dimension-level grades. This gives each dimension the same weight. We find very similar results; for details, see Appendix Table B.3. In addition, following common practice in the literature, our main specification estimates linear probability models for binary outcomes and applies a logarithmic transformation to count outcomes. However, these modeling choices may introduce bias when the outcome contains many zeros (Chen and Roth, 2024). In Appendix Table B.4, we thus repeat our analysis with negative binomial models for the count outcomes and logit models for the binary outcomes. We again find that a paper’s average grade strongly and significantly predicts all five outcomes. In a third robustness check, to exclude the possibility that our results are driven by the use of grades determined after resolution by a third evaluator (for the subsample of papers with two evaluations), we redo this analysis in Appendix Table B.2 using averaged grades instead. The results remain very similar. In a final check, since experiment type may be associated with scientific outcomes in ways that are not perfectly captured by EDAT scores, we test in Appendix Table B.6 whether our results on predictive power are robust to the inclusion of experiment-type fixed effects on top of the other main controls. The average grade remains statistically significantly associated with all scientific outcomes except for policy mentions.

¹²To further address this issue, in Appendix Table B.5 we run a robustness check where we use the average grade to predict the number of mentions recorded for each paper during the week and one, three, six, and twelve months preceding our May 10, 2026 collection of Altmetric data. Since these mentions might be biased towards the papers that were published more recently — potentially the opposite confound from using cumulative scientific outcomes — we still include year fixed effects in the specifications with controls. We find that the estimated coefficient increases as the time window lengthens and becomes statistically significant at the 10% level for the three-month window. Thus, our main result is also robust to using fixed time windows.

Result 2. *EDAT produces quality scores that are highly predictive of quality proxies that the profession considers relevant.*

4.2 Assessing trade-offs

Having established the reliability and relevance of our quality scores, we turn to our central question: how do researchers navigate the multidimensionality of experimental designs?

4.2.1 Existence of trade-offs

We first examine how quality is related across individual components of experimental design by calculating pairwise correlations among all components. Three empirical patterns are possible. First, clear quality sorting would arise if some papers score higher than others on most or all components; it would yield predominantly positive correlations and no negative correlations. Second, component-level quality scores could be largely unrelated, yielding predominantly zero correlations. Third, trade-offs would arise if stronger performance on some components is associated with weaker performance on others. This pattern would imply both positive and negative correlations, as conducting an experiment that excels in some components necessarily means accepting lower quality in others.

Figure 4 presents the results, which strongly support the third case: the existence of trade-offs. As is evident from the figure, many components are positively correlated with each other, with the largest positive correlation reaching 0.92 (*Testing and Accounting for compromised randomization*; $p < 0.01$). The most negative correlation coefficient is -0.64 ($p < 0.01$). Similar shares of pairwise correlations are statistically significantly positive and negative (26% and 21%, respectively). Importantly, these patterns are not driven by between-journal differences. In Appendix Figure B.5, we replicate Figure 4 using only papers published in top-five journals, finding essentially the same pattern.

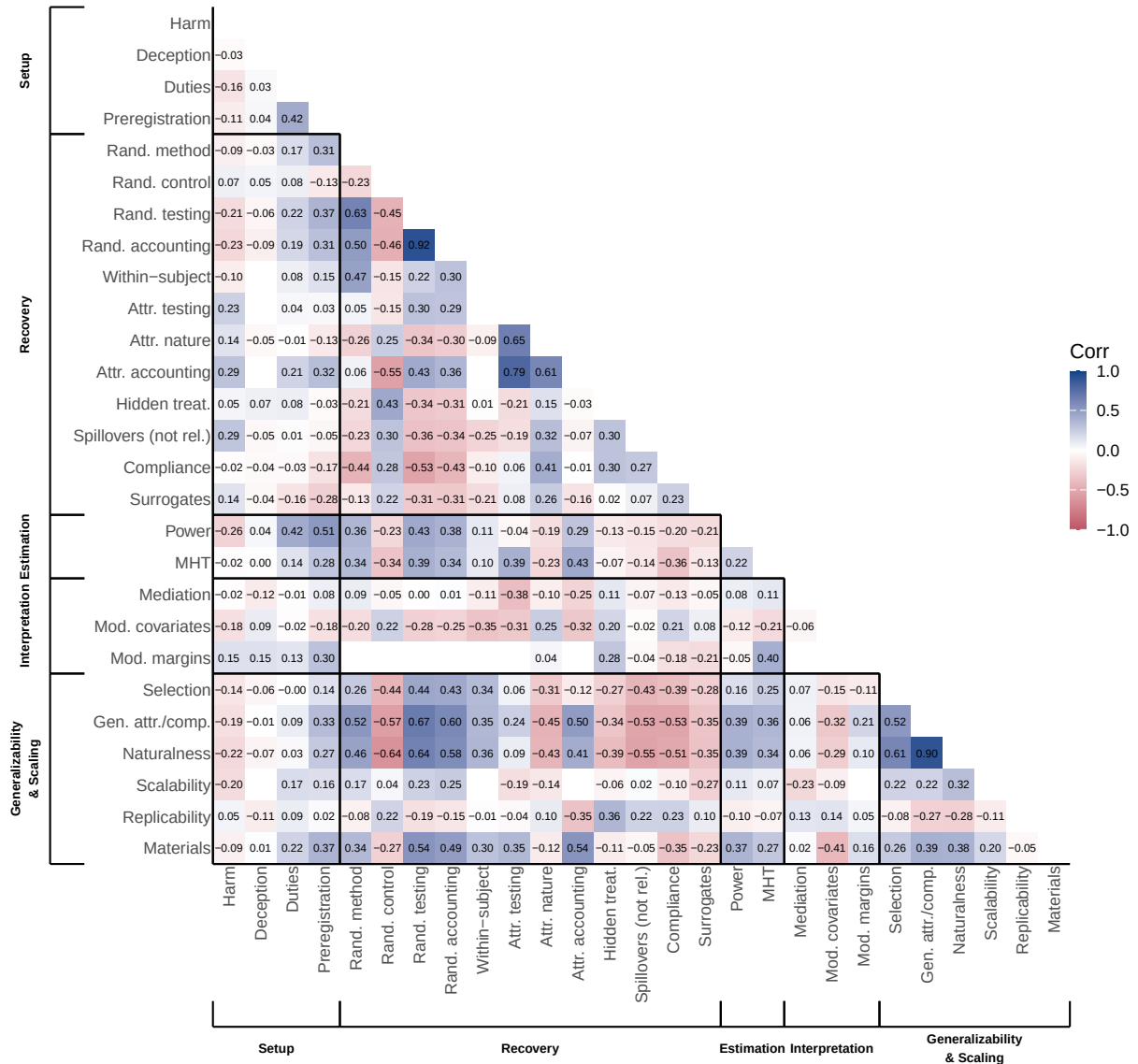
Result 3. *Among published papers, genuine trade-offs in experimental designs exist: higher quality scores in some components are associated with lower scores in others.*

4.2.2 Structure of trade-offs

The existence of trade-offs raises three questions: which dimensions and components are substitutes or complements, where do these patterns arise, and how does the profession value the ways researchers resolve trade-offs?

Severity of threats to identification versus generalizability. A key trade-off visible in our data is between certain components of the *Recovery* and *Generalizability & Scaling* dimensions. For instance, we observe that the *Naturalness* component—the extent to which the experiment is conducted in the setting of interest—is associated with the extent to which

Figure 4: Correlations between individual components



Notes: This figure displays the correlation matrix between component grades. Each cell represents the pairwise Pearson correlation between the grades of two components, using all papers for which both components were graded; the number in each cell is the correlation coefficient. Blue (red) cells indicate positive (negative) correlations. Cells with fewer than 15 paired observations are left blank, and two components (*Spillovers relevant to the research question* and *Design of treatment cells*) are omitted entirely for the same reason. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Full names for the abbreviated axis labels are given in Appendix Table A.1.

threats to identification are present. Appendix Figure B.6 shows this in detail: *Naturalness* is negatively correlated with measures of compliance, control over randomization, and the absence of spillovers, attrition, and hidden treatments, with correlations ranging from -0.39 to -0.64 (all $p < 0.01$). The figure also reveals that these threats to identification appear together, as all five threats are positively correlated with each other. For instance, the extent to which a design achieves compliance and the extent to which outcomes are observed are correlated with a coefficient of 0.41 ($p < 0.01$). Likewise, we find that *Selection*, another important component of the *Generalizability & Scaling* dimension, is negatively associated with

components related to threats to identification. That is, experiments that have no selection by design tend to have larger problems with attrition and compliance.

Addressing threats to identification versus generalizability. The *Recovery* dimension captures both the severity of threats to identification and the degree to which an experiment tests for and accounts for them. For example, the *Nature of attrition* component captures the extent to which attrition is a problem for a given experiment, while *Accounting for non-random attrition* captures how this threat is addressed. Experiments with higher grades in *Generalizability & Scaling* score worse on components capturing threats to identification, but better on several components relating to the testing and mitigation of these threats. Similarly, Appendix Figure B.7 shows that experiments facing more severe threats to identification tend to receive higher grades on testing and mitigation.

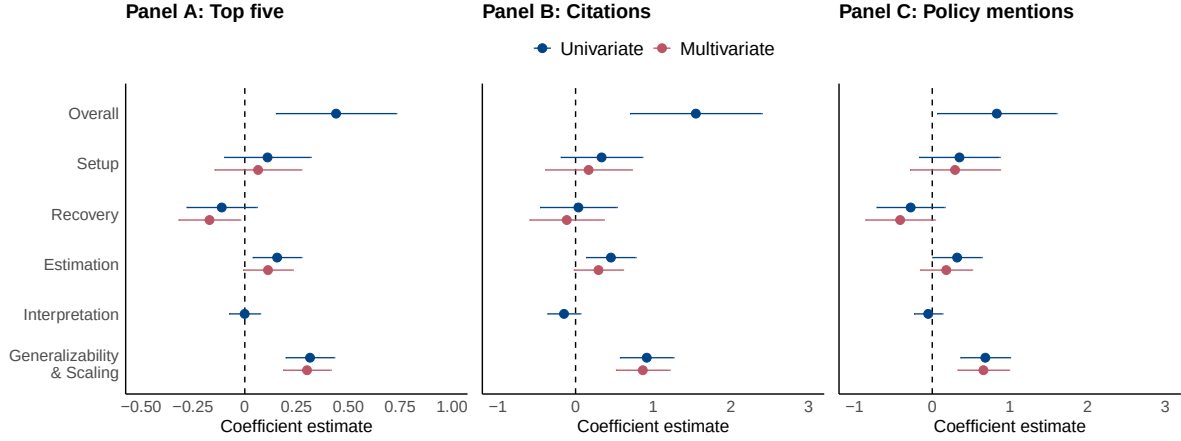
Complementarities between Setup and Estimation. A third interesting comparison concerns the *Setup* and *Estimation* dimensions. Here, we find several systematic positive correlations, indicating complementarities. In particular, the components *Preregistration* and *Ethical duties* are associated with higher grades for the *Power analysis* and *Multiple hypothesis testing* components.

Result 4. *Systematic trade-offs exist between certain components of Recovery and Generalizability & Scaling, while logical complementarities exist between subcomponents of Setup and Estimation.*

4.2.3 How are dimensions valued?

Given the existence of trade-offs, how does the profession value the ways in which researchers resolve them? Specifically, we investigate how differences in quality across dimensions are associated with our scientific outcome variables. In Figure 5, we investigate the association between quality scores within each dimension and three outcome variables—publication in a top-five journal (Panel A), citations (Panel B), and policy impact (Panel C). We estimate two sets of specifications: univariate specifications, in which we regress each outcome separately on each dimension score, and multivariate specifications, in which all dimension scores enter jointly. We find that the *Generalizability & Scaling* dimension is strongly associated with publication in a top-five journal, higher citation counts, and more mentions in policy documents (all $p < 0.01$). Here, an increase in the average *Generalizability & Scaling* grade from the median of 1.80 to the 75th percentile of 2.20 is associated with a 12.59 p.p. higher likelihood of the paper being published in a top-five journal, a 44% higher citation count, and an increase of 38% in policy mentions. One concern is that experiment type may be associated with scientific outcomes in ways that are not perfectly captured by EDAT scores. To address this, Appendix Figure B.10 repeats the analysis in Figure 5 with experiment-type fixed effects and again finds positive associations in the *General-*

Figure 5: Relating dimension grades to publication, citation, and policy-mention outcomes

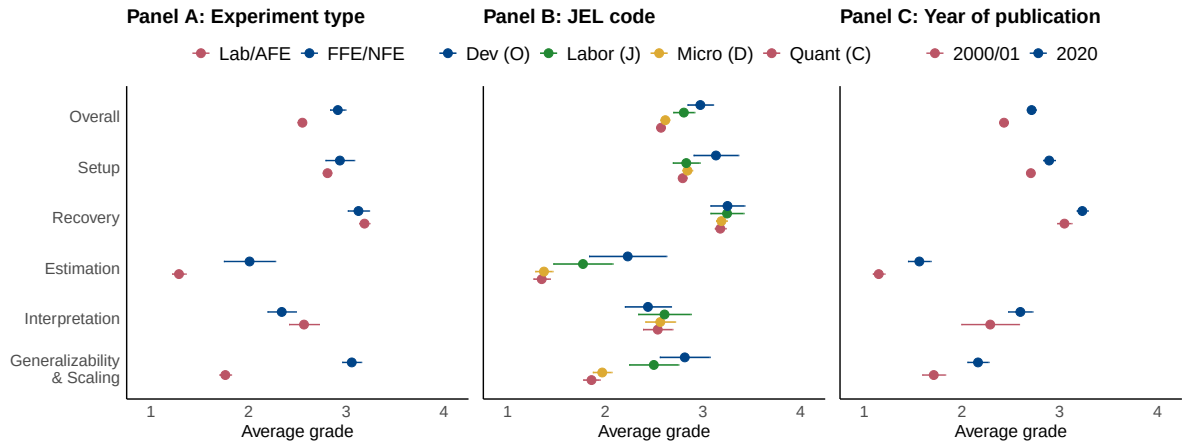


Notes: This figure displays coefficient plots summarizing the relationship between proxies for scientific quality and EDAT scores. Each point is a coefficient estimate from a regression of the outcome on the paper's average grade for a given dimension. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Panel A reports estimates from a linear probability model for publication in a top-five journal; Panels B and C are OLS regressions with dependent variables $\log(\text{Citations})$ and $\log(1 + \text{Policy mentions})$, respectively. *Univariate* estimates (blue) are from regressions that feature only one dimension grade at a time; *multivariate* estimates (red) are from a joint regression of the outcome on grades for *Setup*, *Recovery*, *Estimation*, and *Generalizability & Scaling*. *Interpretation* is excluded from the joint regression due to missing data. All specifications include controls for $\log(\text{Number of authors})$, publication-year fixed effects, and JEL-code fixed effects. Bars are 95% confidence intervals based on robust standard errors.

izability & Scaling dimension. These remain statistically significant for top-five publication and citations; for policy mentions, the association is significant at the 10% level only in the multivariate specification. Moreover, Appendix Figure B.11 estimates this coefficient separately for the Lab/AFE and FFE/NFE subgroups. The coefficients are positive for all three outcomes and statistically significant for top-five publication and citations in the Lab/AFE group, indicating that the predictive power of *Generalizability & Scaling* is not entirely driven by between-subgroup differences. Within FFE/NFE, only the citation coefficient in one specification is statistically significant. The point estimates are uniformly positive and broadly similar between subgroups, but estimates in the smaller FFE/NFE group are less precise.

The coefficients on *Estimation* scores are also positive and significantly associated with the outcomes ($p < 0.05$ for all in the univariate specifications), albeit smaller in magnitude: the coefficient on *Generalizability & Scaling* is significantly larger than the *Estimation* coefficient for top-five publication probability and citations (both $p < 0.05$), but only at the 10% level for policy mentions. Interestingly, *Recovery* scores are not significantly associated with outcomes ($p > 0.10$ in all univariate specifications). One potential interpretation is that *Recovery* is a necessary condition for publication: experiments that have severe identification problems may end up not being published. Moreover, authors likely design their experiments to mitigate the threats to identification, either in the design stage or in the statistical testing stage. For instance, as we have seen in the previous section, some papers face threats to identification such as attrition and lack of compliance, and respond by employing appropriate tests to measure and bound the consequences. Accordingly, our results indicate

Figure 6: Average dimension grades by experiment type, JEL code, and year



Notes: This figure displays heterogeneity in dimension grades by key paper characteristics. Each point shows the mean grade of a given dimension for that particular subsample of papers. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Panel A splits papers by experiment type, pooling laboratory and artefactual field experiments (Lab/AFE) and framed and natural field experiments (FFE/NFE). Panel B groups papers by JEL code: Dev (O) = Development; Labor (J) = Labor Economics; Micro (D) = Microeconomics; and Quant (C) = Quantitative Methods. These groups are not mutually exclusive, as papers can carry multiple JEL codes. Panel C compares papers published in 2000/01 with those published in 2020. Bars are 95% confidence intervals based on the standard error of the mean.

that published papers appear to differ less in internal validity than in how they address external validity, specifically the *Generalizability & Scaling* dimension.¹³

Result 5. *The strongest association between quality scores and scientific outcomes is in the Generalizability & Scaling dimension. In contrast, Recovery does not consistently predict outcomes, suggesting it is a necessary condition for publication.*

4.2.4 Dimensions across methods, fields and over time

Differences across experimental types. In Figure 6 Panel A, we compare grades across different types of experiments: (i) laboratory experiments (Lab), which involve convenience samples of participants who are aware of the experiment and are placed in an environment constructed by the experimenter; (ii) artefactual field experiments (AFE), which involve participants from the relevant target population but retain the laboratory features of an experimenter-constructed environment and participant awareness; (iii) framed field experiments (FFE), which involve participants from the target population in a natural market setting who remain aware of the experiment; and (iv) natural field experiments (NFE), which feature subjects of the target population in natural markets who are not aware of the experiment. Because NFEs account for only 3% of our sample, we pool them with FFEs.

¹³Appendix Figure B.8 repeats the analysis using logit and negative binomial models and yields similar results. To address the concern that our results are driven by the use of third-evaluator-resolved grades (for the subsample of papers with two evaluations), we also redo this analysis in Figure B.9 using averaged grades instead. The results remain extremely similar.

We find that Lab/AFE have higher grades in the *Recovery* dimension but not significantly so ($p > 0.10$), while FFE/NFE score higher particularly in the *Generalizability & Scaling* dimension ($p < 0.01$). The greater control over the features of the decision environment in a Lab/AFE through the use of constructed environments allows these types to better deal with threats to identification. As displayed in Appendix Figure B.12, Lab/AFEs achieve significantly higher grades for components such as compliance, hidden treatments, extent of attrition, or control over randomization ($p < 0.01$ for all). The drawback is lower control over the naturalness of the subject pool and decision. Here, FFE/NFEs achieve significantly higher grades when dealing with components such as selection, generalizability, and naturalness ($p < 0.01$ for all), as Appendix Figure B.13 shows. Our results thus add empirical evidence to the discussion of the advantages and disadvantages of different types of experiments. See also List (2026) for a formal treatment of the complementary advantages of the four experimental types.

Differences across fields. Panel B of Figure 6 shows how grades across dimensions vary by research field, as identified by JEL codes. Papers in JEL categories J (*Labor*) and O (*Development*) receive significantly higher scores in the *Estimation* and *Generalizability & Scaling* dimensions than those in C (*Quantitative Methods*) and D (*Microeconomics*) (all pairwise comparisons, $p < 0.05$). Part of the differences can be attributed to different compositions in the type of experiments across research fields, as, for instance, experiments in development are more likely to be FFEs or NFEs. Interestingly, however, a significant difference persists even when we control for the type of experiment (see Appendix Table B.7 for details).

Differences over time. In Figure 6 Panel C, we examine how quality scores have changed over time. We find significantly higher scores for papers published in 2020 across all dimensions ($p < 0.01$ for all dimensions except *Interpretation*, for which the difference is marginally significant at $p < 0.10$). These results suggest that experimental design quality has improved over time. Indeed, we find that the average overall grade among papers published in top-five journals in 2000 and 2001 (2.46) is lower than the average overall grade of a paper published in one of the field journals of our sample in 2020 (2.62) ($p < 0.01$). Part of this development may not be surprising: scores on several components depend on tests, methods, or practices that were not widely known or available in 2000 and 2001. For instance, the use of multiple hypothesis testing has increased substantially over time. While 1.5% of papers in 2000 and 2001 employ MHT, 19% do so in 2020 ($p < 0.01$). Similarly, preregistration did not exist in economics until 2013 (Imai et al., 2025). In line with this, whereas none of the papers in our 2000 and 2001 sample were preregistered, 16% of those published in 2020 were ($p < 0.01$). However, the fact that increases in grades appear across the entire spectrum of experimental designs suggests genuine improvements in quality.

Result 6. *We measure systematic improvements in quality over time and heterogeneity in design features between subfields and experiment types.*

5 Conclusion

The goal of this paper is to help researchers improve the design and evaluation of experiments. To do so, we developed two tools: First, a comprehensive framework that distills an experiment into 30 distinct components. Each captures an important aspect of experimental designs and contains a quality rubric. Second, an assessment tool that operationalizes the framework as a survey-type evaluation instrument. The tool walks a user through the framework with targeted questions and automatically generates quality scores based on the user's answers. Using data from 205 published experiments, we applied the framework and assessment tool to provide evidence on its validity, reliability and usefulness. Our results provide a comprehensive assessment of the current state of experiments in economics: how the profession makes design decisions and how they are valued.

References

- Abrams, Eliot, Jonathan Libgober, and John A. List (2020).** "Research Registries: Facts, Myths, and Possible Improvements." Working Paper 27250. National Bureau of Economic Research. [10]
- Akerlof, George A. (2020).** "Sins of Omission and the Practice of Economics." *Journal of Economic Literature* 58 (2): 405–18. [6]
- Altmetric (2024).** "The Changing Landscape of Journal Performance Measurement: How Publication Planners and Medical Affairs Communications Teams Can Leverage the New Impact Metrics That Matter for Journal Selection." White paper. Altmetric, a Digital Science company. [23]
- Andre, Peter, and Armin Falk (2025).** "What's Worth Knowing?: Economists' Opinions about Economics." *Working Paper*. [6, 19]
- Athey, Susan, and Guido Imbens (2016).** "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–60. [16]
- Banerjee, Abhijit V., and Esther Duflo (2017).** *Handbook of Economic Field Experiments*. Handbook of Economic Field Experiments. Amsterdam: North-Holland, an imprint of Elsevier. [5]
- Brodeur, Abel, Nikolai M. Cook, Jonathan S. Hartley, and Anthony Heyes (2024).** "Do Preregistration and Preanalysis Plans Reduce p-Hacking and Publication Bias? Evidence from 15,992 Test Statistics and Suggestions for Improvement." *Journal of Political Economy Microeconomics* 2 (3): 527–61. [10]
- Capozza, Francesco, Robert Dur, and Pedro Rey-Biel (2026).** "Measuring the Acceptability of Experimental Research in Economics." *Working Paper*. [10]
- Card, David, and Stefano DellaVigna (2020).** "What Do Editors Maximize? Evidence from Four Economics Journals." *Review of Economics and Statistics* 102 (1): 195–217. [5, 23, 26]

Charness, Gary, Anya Samek, and Jeroen Van de Ven (2022). “What is considered deception in experimental economics?” *Experimental Economics* 25 (2): 385–412. [9]

Chen, Jiafeng, and Jonathan Roth (2024). “Logs with Zeros? Some Problems and Solutions.” *The Quarterly Journal of Economics* 139 (2): 891–936. [29]

Cherkashin, Ivan, Svetlana Demidova, Susumu Imai, and Kala Krishna (2009). “The inside Scoop: Acceptance and Rejection at the Journal of International Economics.” *Journal of International Economics* 77 (1): 120–32. [5]

Common Rule (2018). “Common Rule.” *Code of Federal Regulations*, Vol. 45 § 46. U.S. Department of Health and Human Services. [9]

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth (2018). “Measuring and bounding experimenter demand.” *American Economic Review* 108 (11): 3266–302. [9]

Deaton, Angus, and Nancy Cartwright (2018). “Understanding and misunderstanding randomized controlled trials.” *Social Science & Medicine* 210: 2–21. [5]

Forscher, Patrick S., Markus Brauer, Flavio Azevedo, William Taylor Laimaka Cox, and Patricia G. Devine (2019). “How many reviewers are required to obtain reliable evaluations of NIH R01 grant proposals?” Preprint. PsyArXiv. [5]

Fréchette, Guillaume R., Kim Sarnoff, and Leeat Yariv (2022). “Experimental Economics: Past and Future.” *Annual Review of Economics* 14 (Volume 14, 2022): 777–94. [6]

Gelman, Andrew (2019). “Don’t calculate post-hoc power using observed estimate of effect size.” *Annals of surgery* 269 (1): e9–e10. [14]

Ghanem, Dalia, Sarojini Hirshleifer, and Karen Ortiz-Becerra (2026). “Testing Attrition Bias in Field Experiments.” *Journal of Human Resources* 61 (2): 690–731. [12, 41]

Heckman, James J, and Ganesh Karapakula (2021). “Using a satisficing model of experimenter decision-making to guide finite-sample inference for compromised experiments.” *Econometrics Journal* 24 (2): C1–C39. [12]

Hoenig, John M, and Dennis M Heisey (2001). “The abuse of power: the pervasive fallacy of power calculations for data analysis.” *American Statistician* 55 (1): 19–24. [14]

Hofmeister, Robert, and Matthias Krapf (2011). “How Do Editors Select Papers, and How Good Are They at Doing It?” *The B.E. Journal of Economic Analysis & Policy* 11 (1): Article 64. [5]

Horowitz, Joel L, and Charles F Manski (2000). “Nonparametric analysis of randomized experiments with missing covariate and outcome data.” *Journal of the American Statistical Association* 95 (449): 77–84. [13, 41]

Imai, Taisuke, Séverine Toussaert, Aurélien Baillon, Anna Dreber, Seda Ertaç, Magnus Johannesson, Levent Neyse, and Marie Claire Villeval (2025). “Pre-Registration and Pre-Analysis Plans in Experimental Economics.” *Working Paper*. [10, 35]

Imbens, Guido W. (2018). “Understanding and misunderstanding randomized controlled trials: A commentary on Deaton and Cartwright.” *Social Science & Medicine* 210: 50–52. [5]

Kagel, John H., and Alvin E. Roth, eds. (2016). *The Handbook of Experimental Economics: Volume 2*. Princeton: Princeton University Press. 776 pp. [5]

Krippendorff, Klaus (2004). *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Thousand Oaks, CA: Sage. [26]

Krippendorff, Klaus (2011). “Computing Krippendorff’s Alpha-Reliability.” *Working Paper*. [24]

Laband, David N., and Michael J. Piette (1994). “Favoritism versus Search for Good Papers: Empirical Evidence Regarding the Behavior of Journal Editors.” *Journal of Political Economy* 102 (1): 194–203. [5]

Landis, J. Richard, and Gary G. Koch (1977). “The Measurement of Observer Agreement for Categorical Data.” *Biometrics* 33 (1): 159–74. [26]

Lee, David S (2009). “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects.” *Review of Economic Studies* 76 (3): 1071–102. [13, 41]

Levitt, Steven D., and John A. List (2007). “What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?” *Journal of Economic Perspectives* 21 (2): 153–74. [9]

List, John (2026). “Don’t Give Up on Lab Experiments: Why the Field Still Needs the Lab.” *Working Paper*. [35]

List, John A. (2020). “Non est disputandum de generalizability? A glimpse into the external validity trial.” Working paper. National Bureau of Economic Research. [16]

List, John A. (2025). *Experimental Economics: Theory and Practice*. Chicago, IL: University of Chicago Press. [6, 22]

List, John A., Azeem M. Shaikh, and Yang Xu (2019). “Multiple hypothesis testing in experimental economics.” *Experimental Economics* 22 (4): 773–93. [15]

Masselus, Lise, Christina Petrik, and Jörg Ankel-Peters (2026). “Lost in the Design Space? Construct Validity in the Microfinance Literature.” Manuscript, March 2026. [5]

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979). “The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research.” Working paper. Washington, D.C.: U.S. Department of Health, Education, and Welfare. Accessed September 15, 2025. [9]

Niederle, Muriel (2025). “Experiments: Why, How, and A Users Guide for Producers as Well as Consumers.” *Working Paper*. [5]

Open Syllabus (2026). “What is Open Syllabus?” <https://blog.opensyllabus.org/about-os>. Open Syllabus Blog. Accessed July 14, 2026. [23]

Peters, Jörg, Jörg Langbein, and Gareth Roberts (2018). “Generalization in the Tropics – Development Policy, Randomized Controlled Trials, and External Validity.” *World Bank Research Observer* 33 (1): 34–64. [5]

Reuben, Ernesto, Sherry Xin Li, Sigrid Suetens, Andrej Svorenčík, Theodore Turocy, and Vasileios Kotsidis (2022). “Trends in the publication of experimental economics articles.” *Journal of the Economic Science Association* 8 (1–2): 1–15. [6]

Romano, Joseph P, and Michael Wolf (2005). “Exact and approximate stepdown methods for multiple hypothesis testing.” *Journal of the American Statistical Association* 100 (469): 94–108. [15, 41]

Romano, Joseph P, and Michael Wolf (2016). “Efficient computation of adjusted p-values for resampling-based stepdown multiple testing.” *Statistics & Probability Letters* 113: 38–40. [15]

Snowberg, Erik, and Leeat Yariv (2025). “Evaluating Experimental Designs.” In *Handbook of Experimental Methodology*. Erik Snowberg and Leeat Yariv, ed. North Holland: Elsevier. [5]

Wager, Stefan, and Susan Athey (2018). “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests.” *Journal of the American Statistical Association* 113 (523): 1228–42. [16]

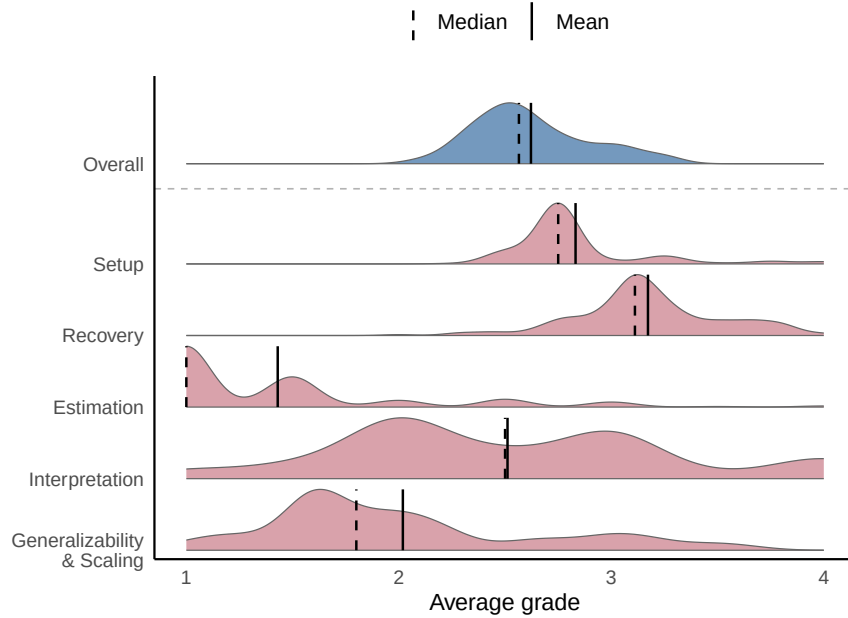
Weber, Max (1919). “Wissenschaft als Beruf.” In *Geistige Arbeit als Beruf*. München, and Leipzig: Duncker & Humblot. [19]

Welch, Ivo (2014). “Referee Recommendations.” *Review of Financial Studies* 27 (9): 2773–804. [5, 26]

Appendix

A Additional figures and tables

Figure A.1: Distribution of grades by dimension



Notes: This figure displays the distribution of average dimension grades across papers. Each row corresponds to one dimension, with the overall average grade on top. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Distributions are based on the 205 papers in the analysis sample, except the distribution for *Interpretation*, as we only have grades on this dimension for 160 papers. Dashed and solid vertical lines mark the median and mean of each distribution, respectively.

Table A.1: Component grading summary

Component	Abbreviation	Gold	Silver	Bronze	Insufficient	Ungraded if
Setup						
Harm to subjects & others	Harm	Risks of harm no greater than those of ordinary life	Additional risks to help answer research question, beneficial in expectation	Additional risks to help answer research question, not beneficial in expectation	Additional risks that do not help to answer research question	N/A; always graded
Deception	Deception	No outright deception	Outright deception used with almost no risk of harm	Outright deception used with minimal risk of harm	Outright deception used with greater than minimal risk of harm	N/A; always graded
Ethical duties	Duties	Both IRB approval and informed consent obtained	IRB approval obtained, informed consent not; participants briefed ex post	IRB approval obtained, informed consent not (or vice versa); no ex post briefing	Neither IRB approval nor informed consent obtained	N/A; always graded
Preregistration	Preregistration	Preregistered with adequate details that match the paper	Preregistered with minimal missing/non-matching details	Preregistered with substantial missing/non-matching details	Not preregistered, or all important details missing/non-matching	N/A; always graded
Recovery						
Hidden treatment versions (HTVs)	Hidden treat.	HTVs implausible given design	HTVs plausible, authors mitigate in design stage	HTVs plausible, authors address econometrically	HTVs plausible, authors do not mitigate or address	N/A; always graded
Spillovers (threats to identification)	Spillovers (not rel.)	Spillovers implausible given design	Spillovers plausible, authors mitigate in design stage or address econometrically	Spillovers plausible, authors address with descriptive evidence	Spillovers plausible, authors do not mitigate or address	Spillovers are objects of interest
Spillovers (objects of interest)	Spillovers (rel.)	Randomized saturation design, parsimonious number of saturation levels	Randomized saturation design, inefficiently many saturation levels	Alternative method of estimating spillover effects	Spillovers implausible, or insufficiently many saturation levels to test hypothesis	Spillovers are threats to identification
Randomization method	Rand. method	Method that leverages pre-treatment covariates (e.g., stratification, matched pairs)	Completely randomized experiment	Bernoulli trials, alternative method, or insufficient information	Not genuinely randomized	Within-subject design
Randomization control	Rand. control	Direct researcher control over randomization, treatment administration	Control (partly) delegated with sufficient oversight, documentation in paper	Control delegated with sufficient oversight or documentation, not both	Control delegated without sufficient oversight or documentation	Within-subject design
Testing for compromised randomization	Rand. testing	Balance testing, tests for uniformity of p-values	Balance testing only	Sample ratio mismatch test or other checks for compromised randomization	No tests or checks reported	Within-subject design
Accounting for compromised randomization	Rand. accounting	Balance testing shows successful randomization	Compromise adequately documented, addressed econometrically	Inadequately documented, addressed econometrically or with modeling	Successful randomization not demonstrated or compromise remains unaddressed	Within-subject design
Testing for attrition	Attr. testing	Test that leverages baseline outcomes, e.g., Ghanem et al. (2026)	Test that leverages covariates, e.g., selective attrition or determinants of attrition test	Other checks for attrition, e.g., rate of attrition test or survey among attritors	No checks for attrition	≥ 95% participant outcomes observed
Nature of attrition	Attr. nature	≥ 95% participant outcomes observed	Random attrition according to statistical tests	Non-random according to statistical tests	Statistical tests for attrition not run	N/A; always graded
Accounting for non-random attrition	Attr. accounting	Selection correction with randomized treatment IV	IPW or selection correction with observational IV	Bound estimation, e.g., Horowitz and Manski (2000) or Lee (2009) bounds	No methods of accounting for non-random attrition	≥ 95% participant outcomes observed or random attrition identical to attrition grades
Balanced panel assumption ^a	–	Identical to attrition grades	Identical to attrition grades	Identical to attrition grades	Identical to attrition grades	N/A; always graded
Compliance	Compliance	≥ 95% compliance or compliance issues implausible given design	ITT, LATE, or bounds on ATE estimated, authors state required assumptions	ITT, LATE, or bounds on ATE estimated, authors do not state required assumptions	No methods of accounting for noncompliance	N/A; always graded
Within-subject design	Within-subject	Stealth design, i.e., participants unaware of experiment	Overt or epochal design with random treatment ordering	Overt or epochal design, authors state required assumptions	Overt or epochal design, authors do not state required assumptions	Between-subject design
Surrogacy	Surrogates	No proxy variables used	Proxy used, correlation with variable of interest shown in observational data	Proxy used, other justification offered	Proxy used, no justification offered	N/A; always graded
Estimation						
Power analysis	Power	Ex ante power or MDE calculations	Ex post power or MDE calculations	Informal discussion of power	No power discussion or calculation	N/A; always graded
Design of treatment cells	Treatment cells	Participants allocated near-optimally to parsimonious number of cells	Participants allocated non-optimally to parsimonious number of cells	Inefficiently many cells given form of relationship of interest	Insufficiently many cells given form of relationship of interest	Within-subject design, non-continuous variables
Multiple hypothesis testing	MHT	Correction that accounts for dependence between tests, e.g., Romano and Wolf (2005)	Correction that iteratively adapts significance threshold, e.g., Holm stepdown	Bonferroni or other correction	No correction for MHT	N/A; always graded
Interpretation						
Mediation	Mediation	Mediators randomized or directly intervened on	Mediators measured, effects quantitatively decomposed into channels	Indirect, qualitative, or correlational evidence on mediators	Mediators identified by authors but no evidence	No mediators identified by authors
Moderation with pre-treatment covariates	Mod. covariates	Moderators randomized or intervened on	Evidence from causal forests or other ML methods	CATEs estimated with classical methods	Moderators identified by authors but no evidence	No moderators identified by authors
Moderation with extensive & intensive margins	Mod. margins	Tobit model, authors state required assumptions	Intensive margin estimated with baseline data, authors state required assumptions	Bounds on margins estimated, or other methods used without stating assumptions	Margins identified by authors but no evidence	No margins with pre-treatment variation identified
Generalizability & Scaling						
Selection (SANS)	Selection	Selection into experiment not possible or sample representativeness shown	Deviations in representativeness identified, documented	Representativeness unknown, acknowledged as caveat	Representativeness unknown, not acknowledged as caveat	N/A; always graded
Attrition & compliance (SANS)	Gen. attr./comp.	Experiment conducted with population of interest in target setting	Attrition & compliance rates reported, differences with target setting discussed	Rates reported or differences discussed, not both	Rates not reported, differences not discussed	N/A; always graded
Naturalness (SANS)	Naturalness	Experiment conducted covertly in natural environment	Overtly in natural environment	Overtly in artificial setting, acknowledged as caveat	Overtly in artificial setting, not acknowledged as caveat	N/A; always graded
Scalability (SANS)	Scalability	Meets ≥ 5 of 7 scaling criteria ^b	Meets 3–4 criteria	Meets < 3 criteria, acknowledged as caveat	Meets < 3 criteria, not acknowledged as caveat	No interventions targeted at policy outcomes
Replicability	Replicability	Replication study already conducted	Population, setting, treatment documented & easy for replicators to access	Population, setting, treatment documented, at least one not easy to access	Insufficient information about population, setting, or treatment to replicate	N/A; always graded
Availability of materials	Materials	Code, data, materials available in replication package, paper, or appendix	At least one available, others available on request or missing due to confidentiality	At least one available, others missing for unknown reasons	No code, data, or materials available	N/A; always graded

Notes: This table summarizes the information about component grading that is detailed in Section 2.2 in the main text. Each row corresponds to one component and is grouped by its dimension. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). The four grade columns give the criteria a component must meet to receive each grade, and the **Ungraded if** column gives the conditions under which a component is not graded. The **Abbreviation** column gives the short label used for each component on the axes of the component-grade correlation matrices in Figure 4 and Appendix Figure B.5.

a: The *Balanced panel assumption* is graded identically to the attrition components and has no distinct correlation-matrix label (marked “–”).

b: The seven scaling criteria are as follows: (i) Selection into experiment not possible or sample representativeness shown; (ii) Experiment conducted with participants from the population of interest; (iii) Experiment conducted in the natural environment of interest; (iv) Intervention does not rely on access to skilled personnel or scarce resources; (v) Intervention does not require many staff or multiple days to implement; (vi) Negative general equilibrium effects are unlikely; (vii) Negative social spillovers are unlikely.

Table A.2: Summary statistics

Variable	Mean	Median	Min	Max
Grades				
Overall	2.62	2.57	2.05	3.30
Setup	2.83	2.75	2.25	4.00
Recovery	3.17	3.11	2.00	4.00
Estimation	1.43	1.00	1.00	4.00
Interpretation	2.51	2.50	1.00	4.00
Generalizability & Scaling	2.02	1.80	1.00	3.67
Number of graded components	21.79	22.00	17.00	26.00
Manuscript metadata and impact				
Number of authors	2.69	3.00	1.00	7.00
Citation count	262.41	64.00	0.00	6705.00
Syllabus inclusion	0.41	0.00	0.00	1.00
Altmetric score	37.01	9.00	0.00	1422.00
Policy mentions	8.36	1.00	0.00	104.00
Experiment type				
Lab experiment	0.62			
Artefactual field experiment	0.18			
Framed field experiment	0.17			
Natural field experiment	0.03			
Experimental design				
Between-subject design	0.80			
Mixed-subject design	0.13			
Within-subject design	0.07			
Published in				
Top five	0.34			
<i>Experimental Economics</i>	0.32			
<i>Games and Economic Behavior</i>	0.23			
<i>Economic Journal</i>	0.12			
<i>American Economic Review</i>	0.12			
<i>Econometrica</i>	0.04			
<i>Journal of Political Economy</i>	0.06			
<i>Quarterly Journal of Economics</i>	0.08			
<i>Review of Economic Studies</i>	0.03			
Year of publication				
Published in 2000	0.16			
Published in 2001	0.16			
Published in 2020	0.68			
JEL codes				
A General Economics and Teaching	0.02			
B History of Economic Thought, Methodology, and Heterodox Approaches	0.01			
C Mathematical and Quantitative Methods	0.75			
D Microeconomics	0.68			
E Macroeconomics and Monetary Economics	0.03			
G Financial Economics	0.09			
H Public Economics	0.10			
I Health, Education, Welfare	0.10			
J Labor and Demographic Economics	0.14			
K Law and Economics	0.03			
L Industrial Organization	0.08			
M Business Administration and Business Economics	0.04			
N Economic History	0.00			
No JEL code	0.00			
O Economic Development, Innovation, Technological Change, and Growth	0.09			
P Economic Systems	0.01			
Q Agricultural and Natural Resource Economics	0.03			
R Urban, Rural, Regional, Real Estate, and Transportation Economics	0.01			
Z Other Special Topics	0.04			
N papers	205			

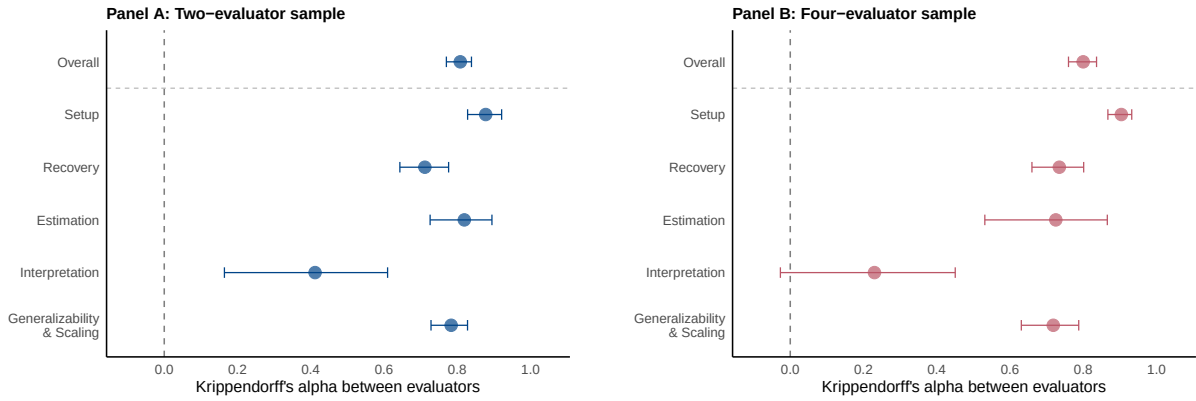
Notes: This table reports summary statistics for the 205 papers in the analysis sample. Each grade row corresponds to a paper's average grade for a given dimension, with the overall average grade on top. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Grade rows are based on all 205 papers, except the row for *Interpretation*, as we only have grades on this dimension for 160 papers. *Number of graded components* reports how many of a paper's components received a grade, as components that do not apply to a paper's design are left ungraded. The columns summarize the distribution of each variable across papers. Indicator variables are reported as shares, for which only the mean is informative. JEL-code groups are not mutually exclusive, as papers can carry multiple JEL codes.

B Additional results and robustness

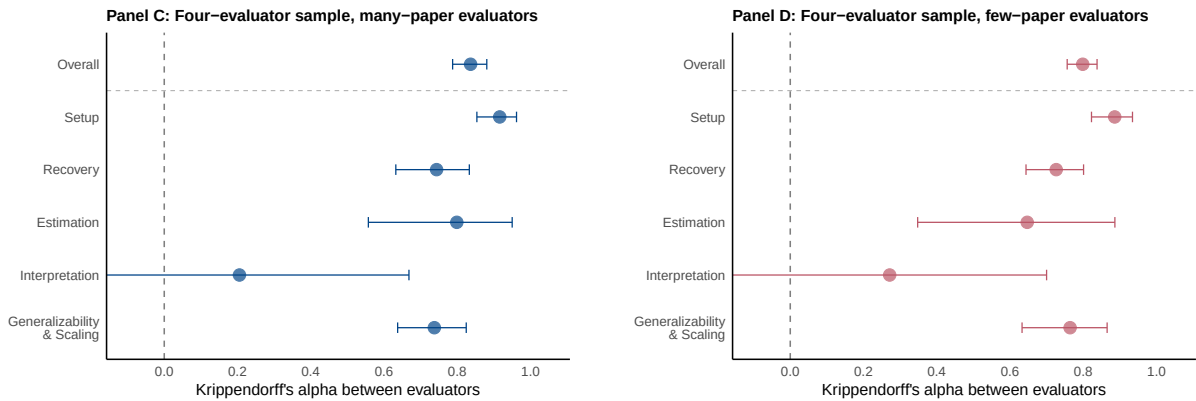
B.1 Reliability robustness figures and heterogeneity

Figure B.1: Inter-evaluator reliability (Krippendorff's α)

All eligible evaluators



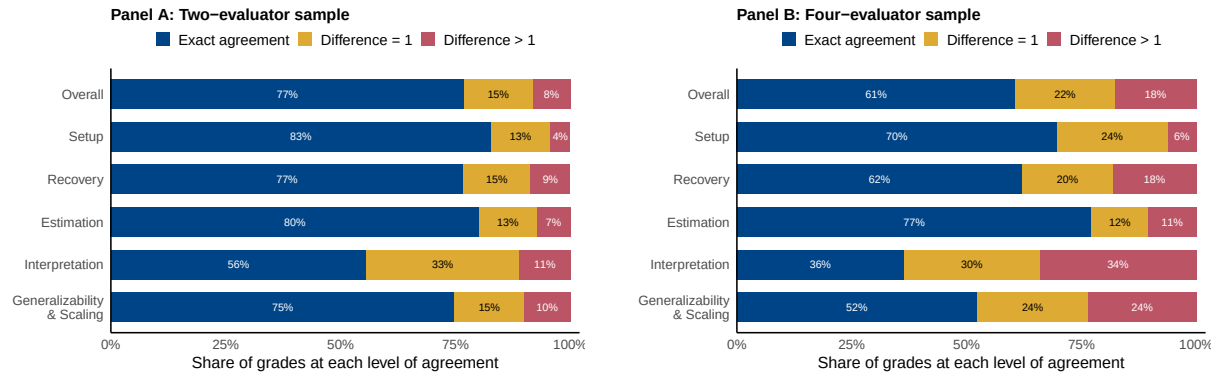
Subset of eligible evaluators



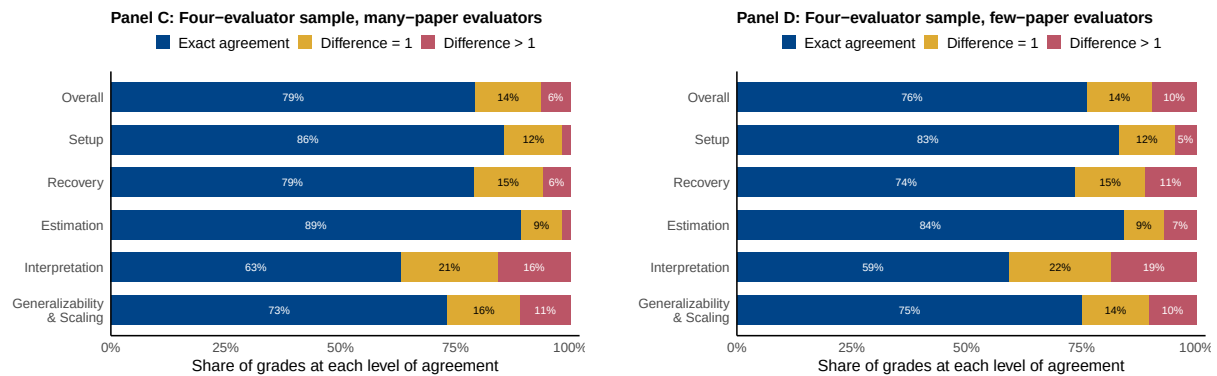
Notes: This figure reports the inter-evaluator reliability of EDAT scores, i.e., the extent to which independent evaluators assign the same grade when scoring the same paper, across various subsamples. Each evaluator grades the applicable components of a paper on a four-point scale from 1 (Insufficient) to 4 (Gold). Reliability is assessed using Krippendorff's α , computed over the component grades of all evaluators of a paper. Panel A shows reliability in the *Two-evaluator sample*: the 75 papers each graded by two *Many-paper evaluators*. Panel B shows reliability in the *Four-evaluator sample*: the 28 papers each graded by four evaluators (two *Many-paper evaluators* and two *Few-paper evaluators*). Panels C and D restrict the *Four-evaluator sample* to evaluations done only by the two *Many-paper evaluators* and by the two *Few-paper evaluators*, respectively. The bars are 95% confidence intervals obtained via bootstrapping clustered at the paper level. Panels A and B repeat Panels A and C of Figure 3.

Figure B.2: Inter-evaluator reliability (level of agreement)

All eligible evaluators



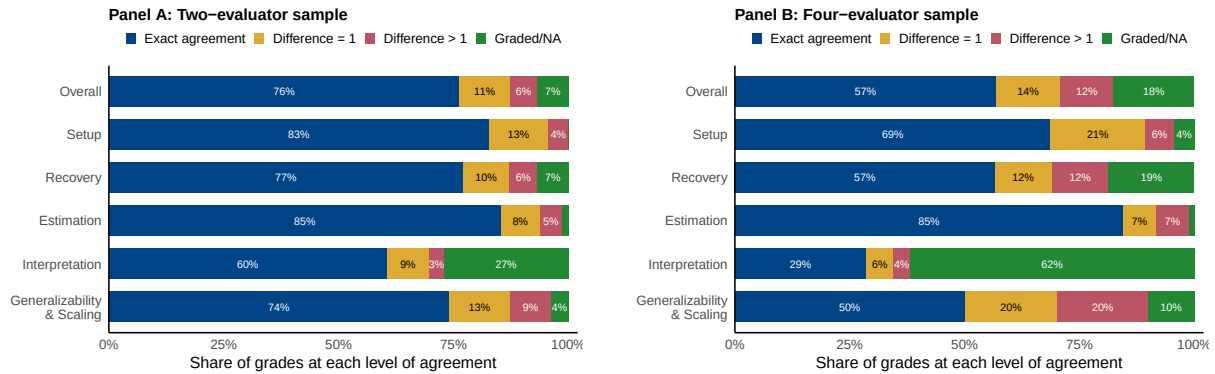
Subset of eligible evaluators



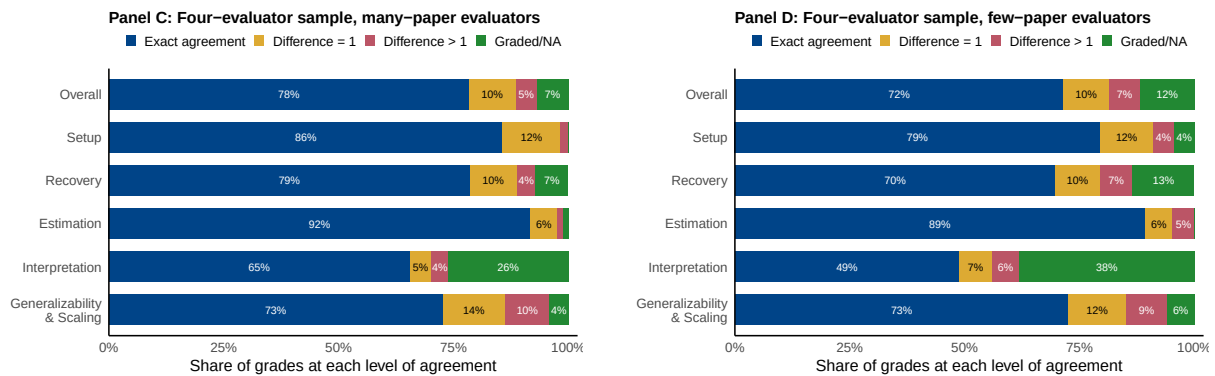
Notes: This figure reports the inter-evaluator reliability of EDAT scores, i.e., the extent to which independent evaluators assign the same grade when scoring the same paper, across various subsamples. Each evaluator grades the applicable components of a paper on a four-point scale from 1 (Insufficient) to 4 (Gold). Reliability is assessed using agreement shares, which report the fraction of graded components on which the evaluators' grades coincide exactly, span one grade level, or span more than one grade level. Panel A shows reliability in the *Two-evaluator sample*: the 75 papers each graded by two *Many-paper evaluators*. Panel B shows reliability in the *Four-evaluator sample*: the 28 papers each graded by four evaluators (two *Many-paper evaluators* and two *Few-paper evaluators*). Panels C and D restrict the *Four-evaluator sample* to evaluations done only by the two *Many-paper evaluators* and by the two *Few-paper evaluators*, respectively. Panels A and B repeat Panels B and D of Figure 3.

Figure B.3: Inter-evaluator reliability (level of agreement, including missing grades)

All eligible evaluators



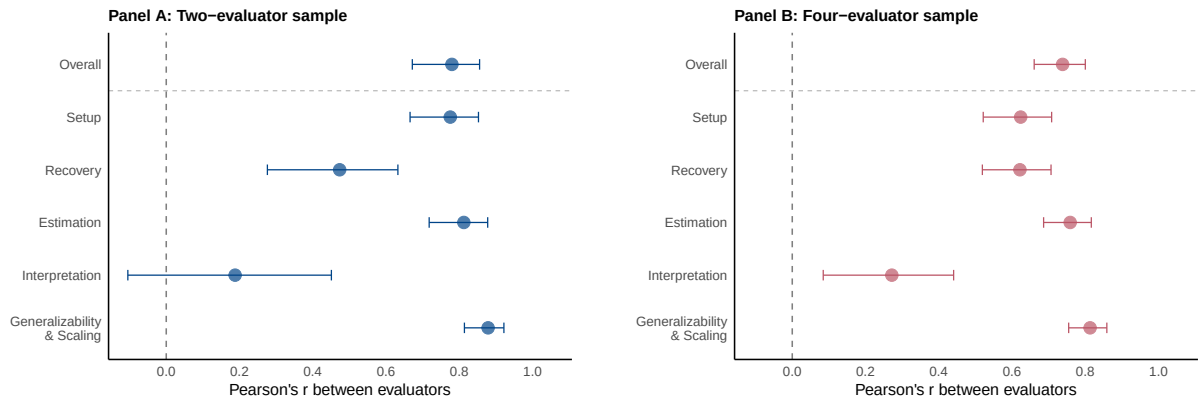
Subset of eligible evaluators



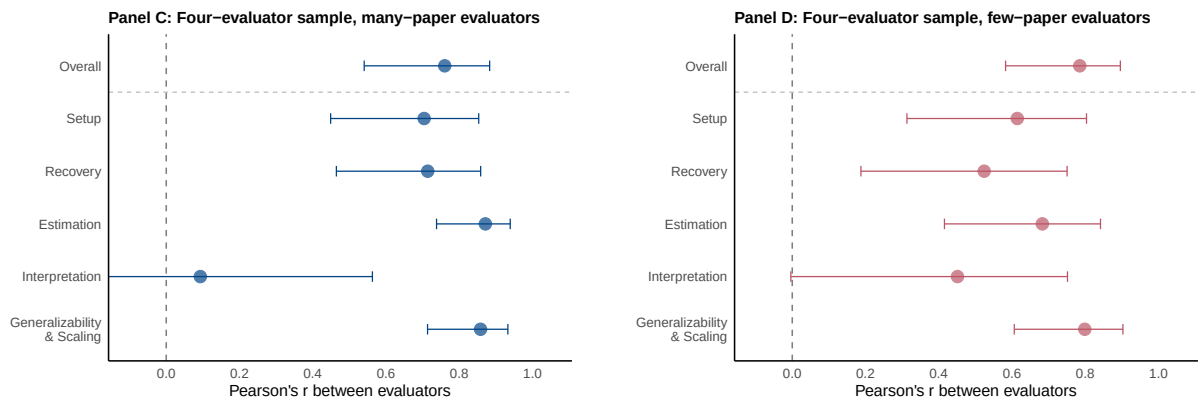
Notes: This figure reports the inter-evaluator reliability of EDAT scores, i.e., the extent to which independent evaluators assign the same grade when scoring the same paper, across various subsamples. Each evaluator grades the applicable components of a paper on a four-point scale from 1 (Insufficient) to 4 (Gold). Reliability is assessed using agreement shares, which report the fraction of graded components on which the evaluators' grades coincide exactly, span one grade level, or span more than one grade level. *Graded/NA* denotes components that one evaluator graded but another left ungraded. Panel A shows reliability in the *Two-evaluator sample*: the 75 papers each graded by two *Many-paper evaluators*. Panel B shows reliability in the *Four-evaluator sample*: the 28 papers each graded by four evaluators (two *Many-paper evaluators* and two *Few-paper evaluators*). Panels C and D restrict the *Four-evaluator sample* to evaluations done only by the two *Many-paper evaluators* and by the two *Few-paper evaluators*, respectively.

Figure B.4: Inter-evaluator reliability (Pearson correlations)

All eligible evaluators



Subset of eligible evaluators



Notes: This figure reports the inter-evaluator reliability of EDAT scores, i.e., the extent to which independent evaluators assign the same grade when scoring the same paper, across various subsamples. Each evaluator grades the applicable components of a paper on a four-point scale from 1 (Insufficient) to 4 (Gold). Reliability is assessed using Pearson correlations between the average dimension grades of all pairs of evaluators who graded the same paper. Panel A shows reliability in the *Two-evaluator sample*: the 75 papers each graded by two *Many-paper evaluators*. Panel B shows reliability in the *Four-evaluator sample*: the 28 papers each graded by four evaluators (two *Many-paper evaluators* and two *Few-paper evaluators*). Panels C and D restrict the *Four-evaluator sample* to evaluations done only by the two *Many-paper evaluators* and by the two *Few-paper evaluators*, respectively. The bars are 95% confidence intervals based on the Fisher z -transformation.

Table B.1: Predicting inter-evaluator reliability and agreement

	Krippendorff's α					Share exact agreement				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
PDF pages (per 100)	-0.12** (0.05)	0.05 (0.07)			0.05 (0.07)	-0.04 (0.03)	0.05 (0.04)			0.05 (0.04)
Flesch–Kincaid grade level			-0.03* (0.02)	-0.01 (0.02)	-0.01 (0.02)			-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)
Year FE		✓		✓	✓		✓		✓	✓
JEL code FE		✓		✓	✓		✓		✓	✓
log(Number of authors)		✓		✓	✓		✓		✓	✓
N	75	74	75	74	74	75	74	75	74	74
R^2	0.06	0.59	0.03	0.58	0.59	0.02	0.48	0.01	0.46	0.48

Notes: This table reports OLS estimates of the relationship between paper characteristics and inter-evaluator reliability. The unit of observation is the paper, using the 75 papers of the *Two-evaluator sample*: the papers each graded by two *Many-paper evaluators*. Each column is a separate OLS regression of overall inter-evaluator reliability (Krippendorff's α , columns 1–5) or the share of exact agreement (columns 6–10) on the listed regressors. Within each outcome, the specifications are: (1)/(6) PDF page length (in hundreds of pages) alone; (2)/(7) PDF page length with controls; (3)/(8) reading difficulty (Flesch–Kincaid grade level; higher values indicate harder text) alone; (4)/(9) reading difficulty with controls; and (5)/(10) both characteristics together with controls. Controls are log(Number of authors), publication-year fixed effects, and JEL-code fixed effects (coefficients omitted). Robust standard errors in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

B.2 Predicting quality

Table B.2: Predicting quality proxies using EDAT (averaged grades)

	Top five (0/1)		log(Citations)		Syllabus (0/1)		log(1 + Altmetric score)		log(1 + Policy mentions)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Average grade	0.56*** (0.11)	0.39*** (0.14)	0.72* (0.37)	1.35*** (0.41)	0.39*** (0.12)	0.41*** (0.14)	2.50*** (0.38)	1.55*** (0.42)	1.38*** (0.38)	0.68* (0.39)
Mean dep. var.	0.34	0.34	4.37	4.37	0.41	0.41	2.39	2.39	1.23	1.23
Mean indep. var.	2.61	2.61	2.61	2.61	2.61	2.61	2.64	2.64	2.64	2.64
R^2	0.10	0.42	0.02	0.51	0.05	0.39	0.20	0.46	0.08	0.56
N	205	205	204	204	205	205	181	181	181	181
JEL code FE		✓		✓		✓		✓		✓
Year FE		✓		✓		✓		✓		✓
log(No. authors)		✓		✓		✓		✓		✓

Notes: This table redoes the OLS-regression analysis of Table 1 with the following modification: in cases of disagreement among evaluators, it uses the average grade among the evaluators rather than the grade resolved by the third evaluator. The dependent variables are the same as those in Table 1. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Controls are log(Number of authors), publication-year fixed effects, and JEL-code fixed effects. Robust standard errors in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

Table B.3: Predicting quality proxies using EDAT with equally weighted dimensions

	Top five (0/1)		log(Citations)		Syllabus (0/1)		log(1 + Altmetric score)		log(1 + Policy mentions)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Average grade	0.50*** (0.09)	0.37*** (0.12)	0.57* (0.34)	1.06*** (0.36)	0.29*** (0.10)	0.25** (0.13)	2.02*** (0.33)	0.89** (0.38)	1.29*** (0.32)	0.70** (0.33)
Mean dep. var.	0.34	0.34	4.37	4.37	0.41	0.41	2.39	2.39	1.23	1.23
Mean indep. var.	2.38	2.38	2.38	2.38	2.38	2.38	2.41	2.41	2.41	2.41
R^2	0.11	0.42	0.01	0.51	0.03	0.37	0.18	0.43	0.09	0.56
N	205	205	204	204	205	205	181	181	181	181
JEL code FE		✓		✓		✓		✓		✓
Year FE		✓		✓		✓		✓		✓
log(No. authors)		✓		✓		✓		✓		✓

Notes: This table redoes the OLS-regression analysis of Table 1 with the following modification: rather than calculating a paper's average grade as an average across all components, it first calculates the average grade within each dimension and then averages the available dimension-level grades, giving each graded dimension equal weight. The dependent variables are the same as those in Table 1. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Controls are log(Number of authors), publication-year fixed effects, and JEL-code fixed effects. Robust standard errors in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

Table B.4: Predicting quality proxies using EDAT (logit and negative binomial models)

	Top five (0/1)		Citations		Syllabus (0/1)		Altmetric score		Policy mentions	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Average grade	0.54*** (0.09)	0.46*** (0.13)	-0.45 (0.47)	1.67*** (0.43)	0.36*** (0.11)	0.41*** (0.14)	2.30*** (0.58)	2.03*** (0.47)	0.88* (0.52)	2.18*** (0.63)
Mean dep. var.	0.34	0.34	262.41	262.41	0.41	0.41	37.01	37.01	8.36	8.36
Mean indep. var.	2.62	2.62	2.62	2.62	2.62	2.62	2.65	2.65	2.65	2.65
Pseudo- R^2	0.09	0.40	0.00	0.08	0.03	0.34	0.03	0.08	0.00	0.12
N	205	205	205	205	205	205	181	181	181	181
JEL code FE		✓		✓		✓		✓		✓
Year FE		✓		✓		✓		✓		✓
log(No. authors)		✓		✓		✓		✓		✓

Notes: This table redoes the regression analysis of Table 1 with the following modification: it uses logit regressions (reported as average marginal effects) for the binary outcomes and negative binomial regressions for the count outcomes. The dependent variables are the same as those in Table 1, except that the count variables enter in levels rather than logarithms. The independent variable is a paper's average EDAT grade across all components. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Controls are log(Number of authors), publication-year fixed effects, and JEL-code fixed effects. Pseudo- R^2 is McFadden's pseudo- R^2 . Robust standard errors in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

Table B.5: Predicting Altmetric mentions over fixed time windows using EDAT

	log(1 + Altmetric mentions in past week)		log(1 + Altmetric mentions in past month)		log(1 + Altmetric mentions in past 3 months)		log(1 + Altmetric mentions in past 6 months)		log(1 + Altmetric mentions in past year)		log(1 + All-time mentions)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Average grade	-0.00 (0.02)	-0.00 (0.04)	0.10 (0.07)	0.03 (0.05)	0.33** (0.13)	0.16* (0.09)	0.50*** (0.17)	0.25* (0.12)	1.01*** (0.22)	0.55*** (0.18)	2.82*** (0.38)	1.63*** (0.44)
Mean dep. var.	0.01	0.01	0.02	0.02	0.09	0.09	0.14	0.14	0.29	0.29	2.23	2.23
Mean indep. var.	2.65	2.65	2.65	2.65	2.65	2.65	2.65	2.65	2.65	2.65	2.65	2.65
R^2	0.00	0.09	0.03	0.23	0.09	0.29	0.11	0.31	0.21	0.46	0.25	0.50
N	181	181	181	181	181	181	181	181	181	181	181	181
JEL code FE		✓		✓		✓		✓		✓		✓
Year FE		✓		✓		✓		✓		✓		✓
log(No. authors)		✓		✓		✓		✓		✓		✓

Notes: This table reports OLS estimates of the relationship between mention counts recorded by Altmetric over fixed time windows and EDAT scores. The dependent variable is the logarithm of one plus the number of mentions recorded for each paper during the week (columns 1–2), month (3–4), three months (5–6), six months (7–8), and one year (9–10) preceding Altmetric data collection on May 10, 2026, as well as all-time mentions (11–12) for comparison. The independent variable is a paper's average EDAT grade across all components. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Controls are log(Number of authors), publication-year fixed effects, and JEL-code fixed effects. Robust standard errors in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

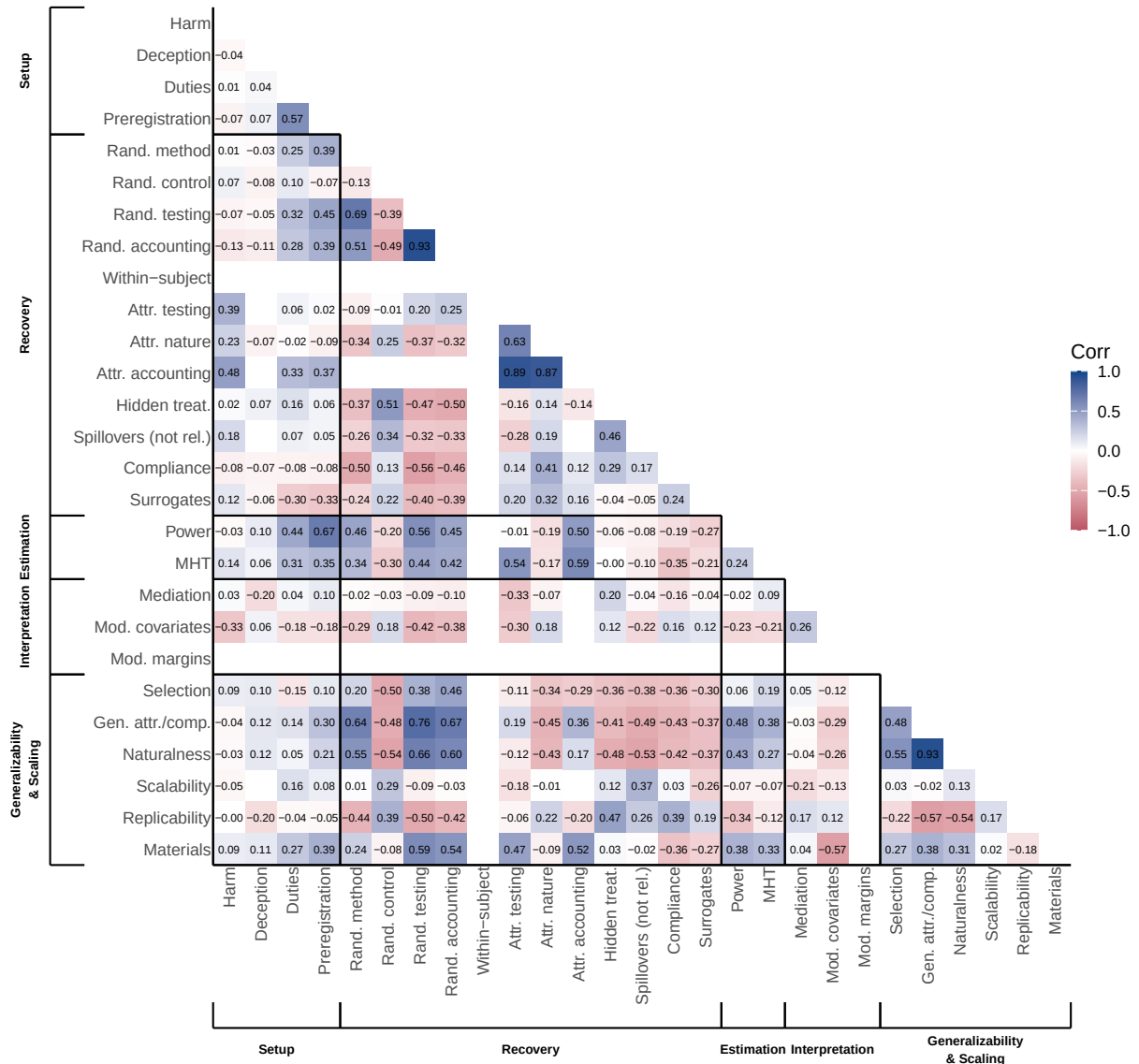
Table B.6: Predicting quality proxies using EDAT, controlling for experiment type

	Top five (0/1)	log(Citations)	Syllabus (0/1)	log(1 + Altmetric score)	log(1 + Policy mentions)
	(1)	(2)	(3)	(4)	(5)
Average grade	0.31** (0.15)	1.30*** (0.46)	0.34** (0.15)	1.32*** (0.48)	0.45 (0.44)
Mean dep. var.	0.34	4.37	0.41	2.39	1.23
Mean indep. var.	2.62	2.62	2.62	2.65	2.65
R^2	0.45	0.53	0.39	0.47	0.60
N	205	204	205	181	181
Experiment type FE	✓	✓	✓	✓	✓
JEL code FE	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
log(No. authors)	✓	✓	✓	✓	✓

Notes: This table reports OLS estimates of the relationship between proxies for scientific quality and EDAT scores. The dependent variable in column (1) is an indicator for whether the paper was published in a top-five instead of a non-top-five journal; in column (2), it is the logarithm of the paper's total citations; in column (3), it is an indicator for whether the paper was included in a course syllabus; in column (4), it is the logarithm of one plus the paper's Altmetric score; and in column (5), it is the logarithm of one plus the number of times the paper was mentioned in a policy document. The independent variable is a paper's average EDAT grade across all components. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Controls are experiment-type fixed effects distinguishing Lab/AFE from FFE/NFE, log(Number of authors), publication-year fixed effects, and JEL-code fixed effects. Robust standard errors in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

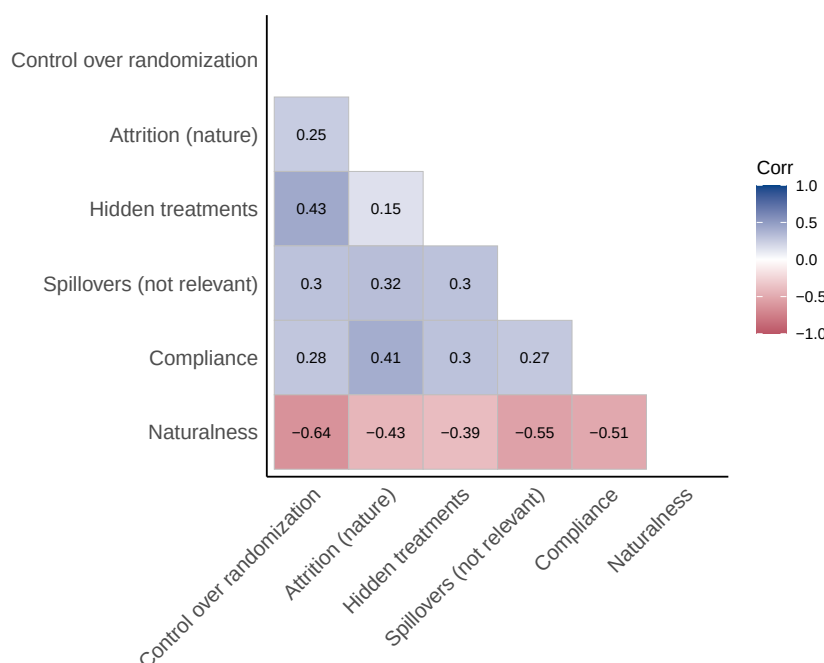
B.3 Correlation matrix robustness and heterogeneity

Figure B.5: Correlations between individual components, top-five journals only



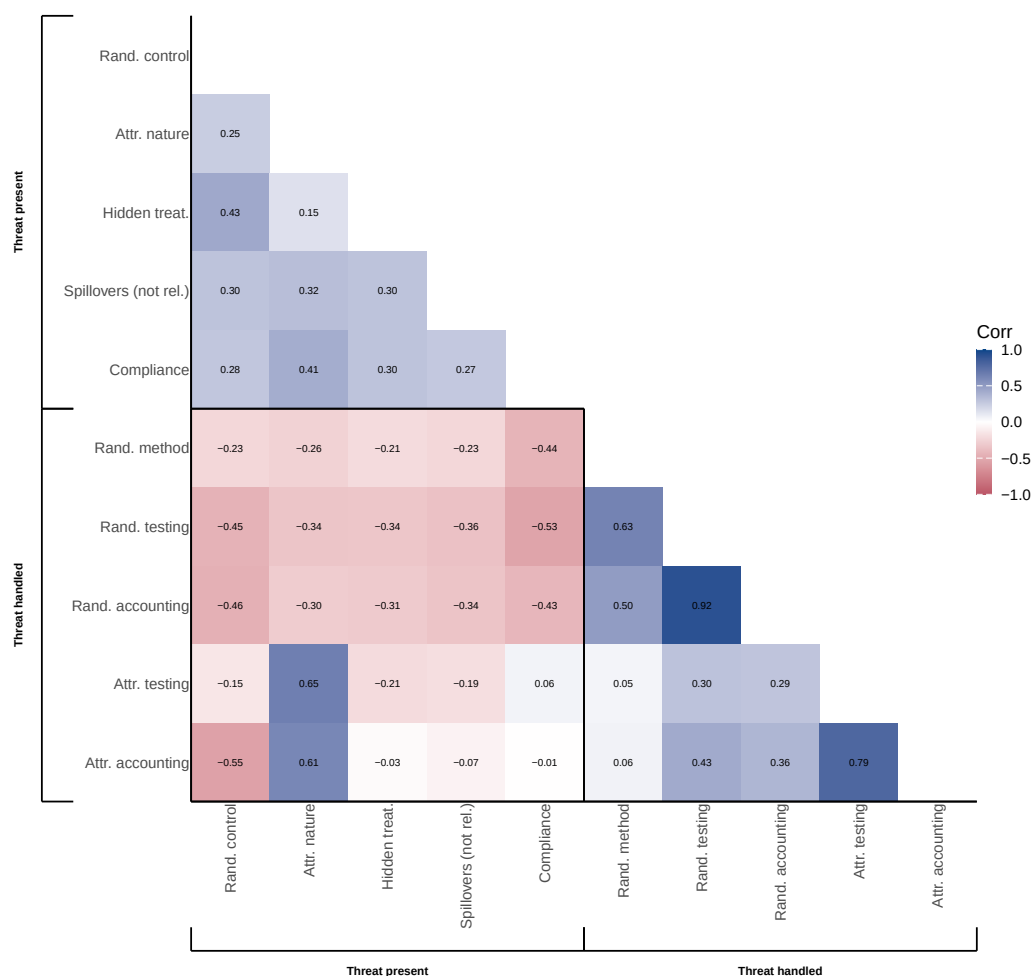
Notes: This figure displays the correlation matrix between component grades for papers published in top-five journals. Each cell represents the pairwise Pearson correlation between the grades of two components, using all such papers for which both components were graded; the number in each cell is the correlation coefficient. Blue (red) cells indicate positive (negative) correlations. Cells with fewer than 15 paired observations are left blank, and two components (*Spillovers relevant to the research question* and *Design of treatment cells*) are omitted entirely, as in Figure 4. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Full names for the abbreviated axis labels are given in Appendix Table A.1. This is the top-five-journal counterpart to Figure 4, which uses all papers.

Figure B.6: Correlations between core internal-validity assumption components and Naturalness



Notes: This figure displays the correlation matrix between the grades of five components that primarily capture core threats to internal validity—*Control over randomization*, *Attrition (nature)*, *Hidden treatments*, *Spillovers (not relevant)*, and *Compliance*—and the *Naturalness* component. Each cell represents the pairwise Pearson correlation between the grades of two components, using all papers for which both components were graded; the number in each cell is the correlation coefficient. Blue (red) cells indicate positive (negative) correlations. Cells with fewer than 15 paired observations are left blank. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Full names for the abbreviated axis labels are given in Appendix Table A.1. These correlations are a subset of those reported in Figure 4.

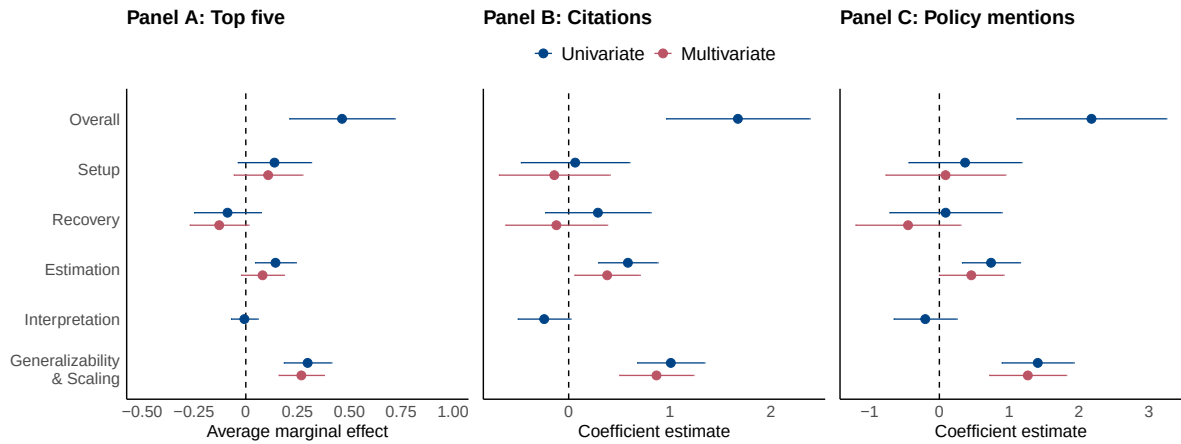
Figure B.7: Correlations between Recovery components capturing the presence and handling of threats



Notes: This figure displays the correlation matrix between the grades of five *Recovery* components that primarily capture core threats to internal validity (*Control over randomization*, *Attrition (nature)*, *Hidden treatments*, *Spillovers (not relevant)*, and *Compliance*) and five that capture how these threats are addressed (*Randomization method*, *Testing for compromised randomization*, *Accounting for compromised randomization*, *Testing for attrition*, and *Accounting for non-random attrition*). Each cell represents the pairwise Pearson correlation between the grades of two components, using all papers for which both components were graded; the number in each cell is the correlation coefficient. Blue (red) cells indicate positive (negative) correlations. Cells with fewer than 15 paired observations are left blank. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Higher grades in the *Threat present* (*Threat handled*) group generally indicate fewer threats to identification (better handling of threats to identification). Full names for the abbreviated axis labels are given in Appendix Table A.1. These correlations are a subset of those reported in Figure 4.

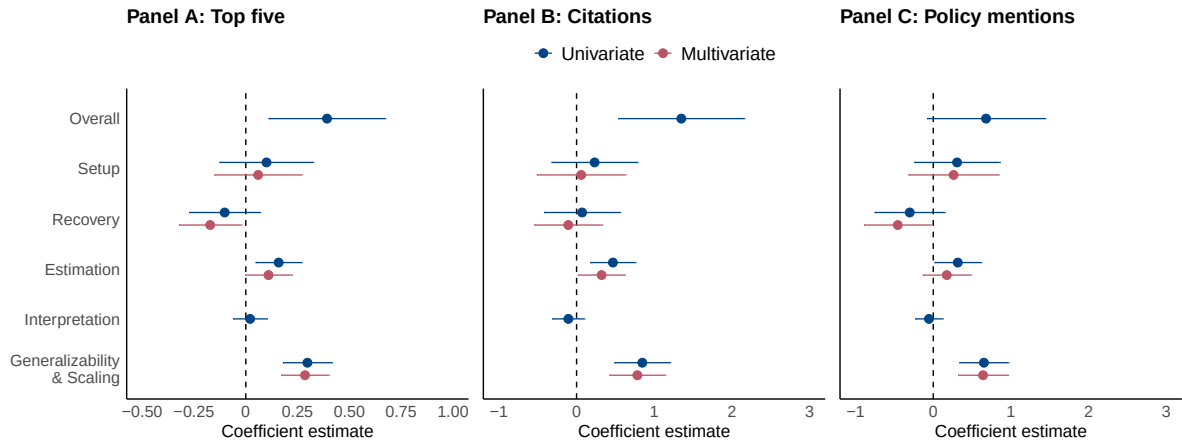
B.4 How dimensions are valued

Figure B.8: Relating dimension grades to scientific outcomes: logit and negative binomial models



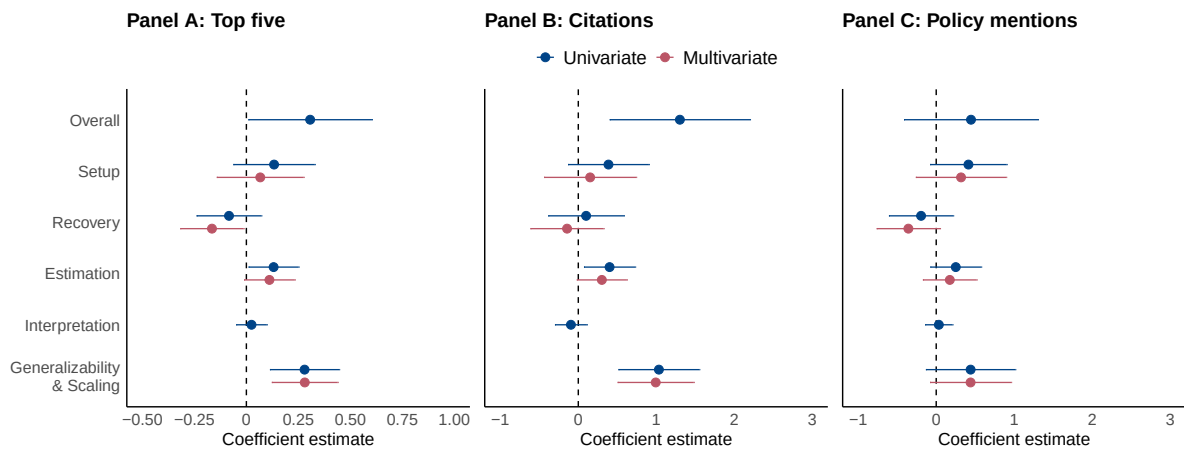
Notes: This figure displays coefficient plots summarizing the relationship between proxies for scientific quality and EDAT scores. Each point is a coefficient estimate from a regression of the outcome on the paper's average grade for a given dimension. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Panel A reports average marginal effects from a logit model for publication in a top-five journal; Panels B and C are negative binomial regressions with dependent variables Citations and Policy mentions, respectively. *Univariate* estimates (blue) are from regressions that feature only one dimension grade at a time; *multivariate* estimates (red) are from a joint regression of the outcome on grades for *Setup*, *Recovery*, *Estimation*, and *Generalizability & Scaling*. *Interpretation* is excluded from the joint regression due to missing data. All specifications include controls for $\log(\text{Number of authors})$, publication-year fixed effects, and JEL-code fixed effects. Bars are 95% confidence intervals based on robust standard errors. This is the non-linear counterpart to Figure 5, which uses a linear probability model and OLS regressions.

Figure B.9: Relating dimension grades to scientific outcomes: averaged grades



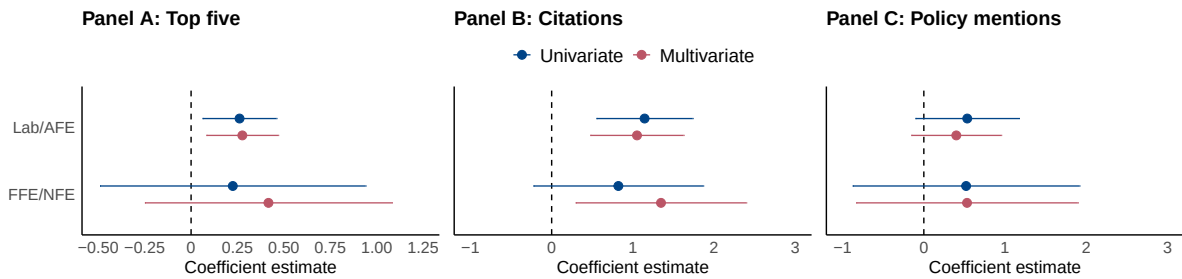
Notes: This figure displays coefficient plots summarizing the relationship between proxies for scientific quality and EDAT scores. Each point is a coefficient estimate from a regression of the outcome on the paper's average grade for a given dimension. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Panel A reports estimates from a linear probability model for publication in a top-five journal; Panels B and C are OLS regressions with dependent variables $\log(\text{Citations})$ and $\log(1 + \text{Policy mentions})$, respectively. *Univariate* estimates (blue) are from regressions that feature only one dimension grade at a time; *multivariate* estimates (red) are from a joint regression of the outcome on grades for *Setup*, *Recovery*, *Estimation*, and *Generalizability & Scaling*. *Interpretation* is excluded from the joint regression due to missing data. All specifications include controls for $\log(\text{Number of authors})$, publication-year fixed effects, and JEL-code fixed effects. Bars are 95% confidence intervals based on robust standard errors. This is the averaged-grade counterpart to Figure 5: in cases of disagreement among evaluators, it uses the average grade among the evaluators rather than the grade resolved by the third evaluator.

Figure B.10: Relating dimension grades to scientific outcomes, controlling for experiment type



Notes: This figure displays coefficient plots summarizing the relationship between proxies for scientific quality and EDAT scores. Each point is a coefficient estimate from a regression of the outcome on the paper's average grade for a given dimension. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Panel A reports estimates from a linear probability model for publication in a top-five journal; Panels B and C are OLS regressions with dependent variables $\log(\text{Citations})$ and $\log(1 + \text{Policy mentions})$, respectively. *Univariate* estimates (blue) are from regressions that feature only one dimension grade at a time; *multivariate* estimates (red) are from a joint regression of the outcome on grades for *Setup*, *Recovery*, *Estimation*, and *Generalizability & Scaling*. *Interpretation* is excluded from the joint regression due to missing data. All specifications include controls for experiment-type fixed effects distinguishing Lab/AFE from FFE/NFE, $\log(\text{Number of authors})$, publication-year fixed effects, and JEL-code fixed effects. Bars are 95% confidence intervals based on robust standard errors.

Figure B.11: Relating Generalizability & Scaling grades to scientific outcomes, separately by experiment-type subsample



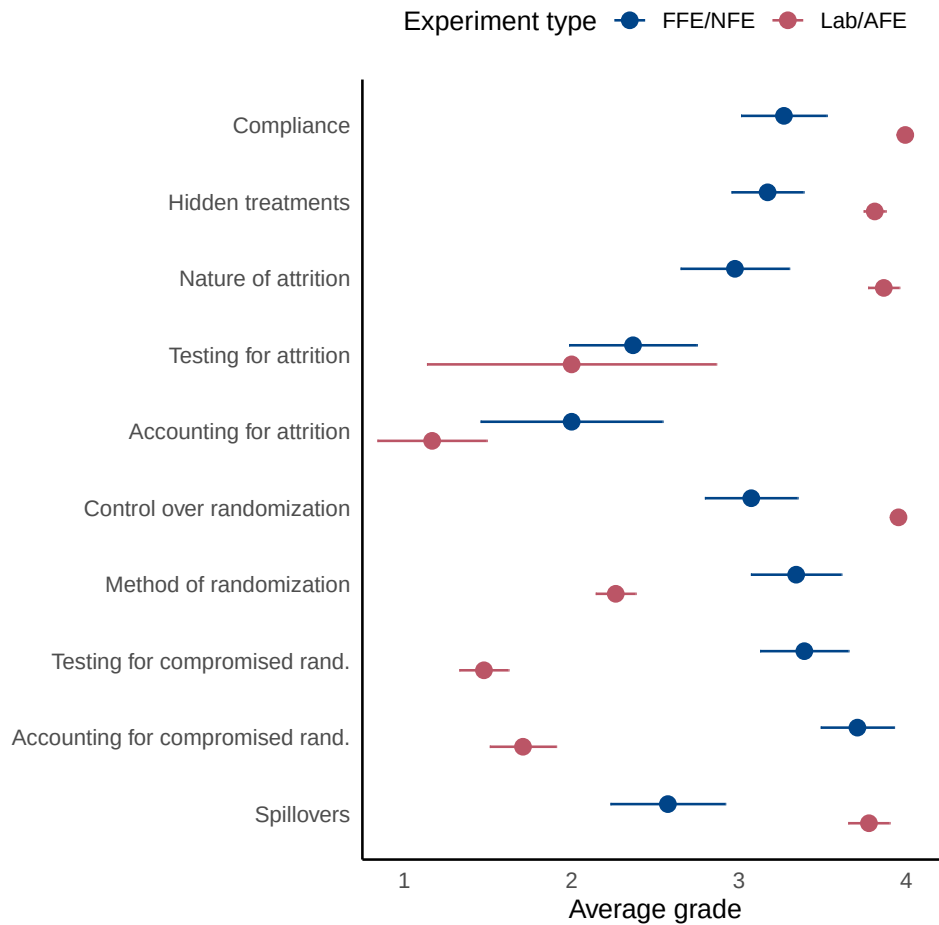
Notes: This figure displays coefficient plots summarizing the relationship between proxies for scientific quality and the *Generalizability & Scaling* dimension. Each point is a coefficient estimate from a regression of the outcome on the paper's average grade on this dimension. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Panel A reports estimates from a linear probability model for publication in a top-five journal; Panels B and C are OLS regressions with dependent variables $\log(\text{Citations})$ and $\log(1 + \text{Policy mentions})$, respectively. Regressions are run separately for laboratory and artefactual field experiments (Lab/AFE) and framed and natural field experiments (FFE/NFE). *Univariate* estimates (blue) are from regressions that feature only one dimension grade at a time; *multivariate* estimates (red) are from a joint regression of the outcome on grades for *Setup*, *Recovery*, *Estimation*, and *Generalizability & Scaling*. *Interpretation* is excluded from the joint regression due to missing data. All specifications include controls for $\log(\text{Number of authors})$, publication-year fixed effects, and JEL-code fixed effects. Bars are 95% confidence intervals based on robust standard errors.

Table B.7: Generalizability & Scaling grade across research fields

	Average Generalizability & Scaling grade	
	(1)	(2)
Microeconomics (JEL D)	0.08 (0.09)	0.13** (0.06)
Labor (JEL J)	0.48*** (0.13)	0.16* (0.09)
Development (JEL O)	0.81*** (0.13)	0.30** (0.14)
Field experiment (FFE/NFE)		1.18*** (0.10)
N	193	193
R^2	0.26	0.66

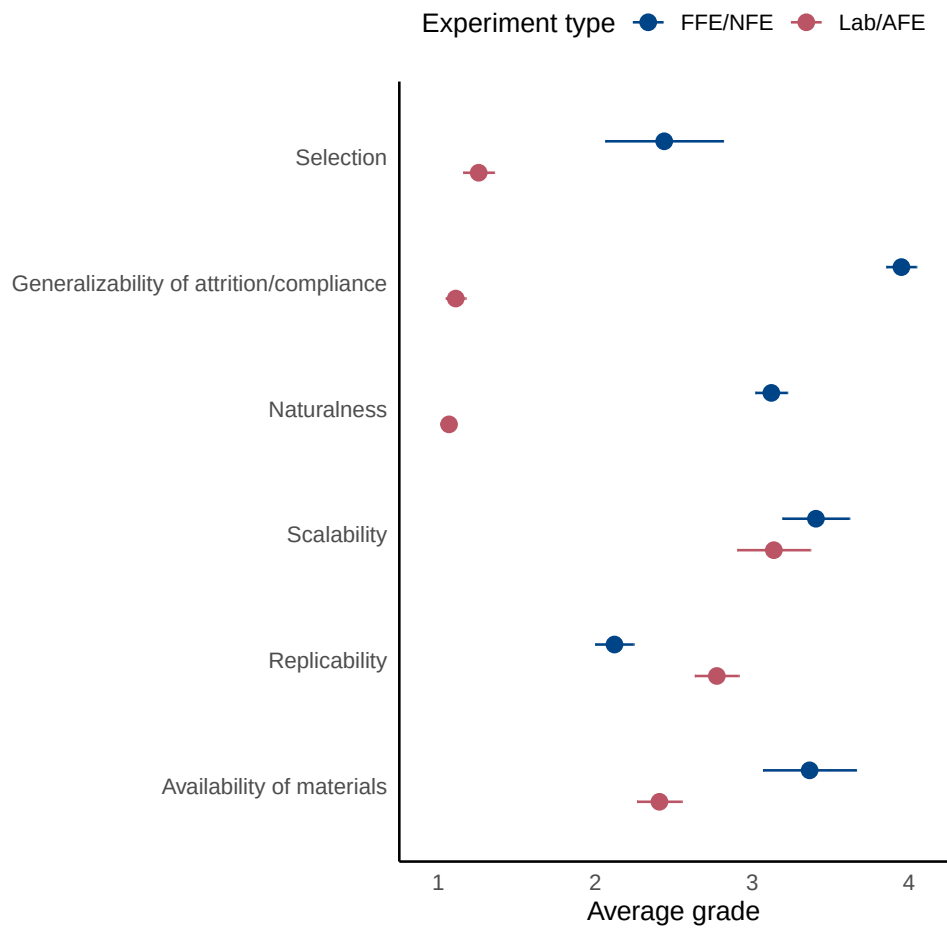
Notes: This table reports OLS estimates of the relationship between a paper's research field and its *Generalizability & Scaling* grade. The dependent variable is a paper's average grade in the *Generalizability & Scaling* dimension. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). The independent variables are indicators for the paper's research field, as measured by its JEL codes; the omitted research-field category is Quantitative Methods (JEL C). Column (2) additionally includes an indicator for framed and natural field experiments (FFE/NFE); its omitted category is laboratory and artefactual field experiments (Lab/AFE). Robust standard errors in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

Figure B.12: Internal validity by experiment type



Notes: This figure displays heterogeneity in component grades by experiment type. Each point shows the mean grade of a given component of the *Recovery* dimension for that particular subsample of papers. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Papers are split by experiment type, pooling laboratory and artefactual field experiments (Lab/AFE) and framed and natural field experiments (FFE/NFE). The figure is based on the 189 papers graded on *Control over randomization*, of which 148 are Lab/AFE and 41 are FFE/NFE. Bars are 95% confidence intervals based on the standard error of the mean. Full names for the abbreviated axis labels are given in Appendix Table A.1. This figure decomposes the *Recovery* dimension reported in Panel A of Figure 6 into its individual components.

Figure B.13: External validity by experiment type



Notes: This figure displays heterogeneity in component grades by experiment type. Each point shows the mean grade of a given component of the *Generalizability & Scaling* dimension for that particular subsample of papers. Grades are on a four-point scale from 1 (Insufficient) to 4 (Gold). Papers are split by experiment type, pooling laboratory and artefactual field experiments (Lab/AFE) and framed and natural field experiments (FFE/NFE). The figure is based on the 205 papers graded on *Selection*, of which 164 are Lab/AFE and 41 are FFE/NFE. Bars are 95% confidence intervals based on the standard error of the mean. Full names for the abbreviated axis labels are given in Appendix Table A.1. This figure decomposes the *Generalizability & Scaling* dimension reported in Panel A of Figure 6 into its individual components.

C Example survey instructions

Observability (general)

Are the main outcomes observable for almost all ($\geq 95\%$) of the participants in every experimental condition? ([Doc.](#))

[Click here to hide information](#)

One of the four key assumptions required for recovering an internally valid average treatment effect (ATE) is that the main outcomes are observed by the researchers for all subjects who participate in the experiment; violations of this assumption are called attrition. For practical purposes, we can define participation in the experiment as having been assigned to a treatment or control condition, regardless of whether they are informed of or have experienced their treatment status. Attrition that occurs before this step may affect the representativeness of the sample but does not threaten internal validity. For an example of why attrition can matter, consider an experiment where students who benefit less from treatment are more likely to drop out of the study and render their outcomes unobservable than those who benefit more. This type of attrition may induce between-condition imbalance in potential outcomes in the remaining sample, thus complicating the calculation of the ATE. The most effective way of dealing with attrition is to minimize it altogether. As such, please report whether the experiment achieves perfect or near-perfect ($\geq 95\%$) observability in every experimental condition. The authors may make a statement to this effect, e.g., "outcomes are only missing for 15 out of 500 participants in the control arm and 30 out of 500 participants in the treatment arm" (observability rates of 97% and 94%). Alternatively, you may have to calculate the observability rate by comparing sample sizes at initial assignment to those when outcomes are measured; any loss of outcome data between these two stages counts as attrition. Please select "no" if this rate is below 95% for any experimental conditions. In the example above, you would select "no" because the treatment arm has an observability rate of 94%.

- ☐ Yes
- ☐ No

Reminder: the question on this page has a significant impact on the survey flow and/or multiple grades. Please take special care with responding.



Testing for attrition (i)

What type of test or check for non-random attrition do the authors conduct, if any? [\(Doc.\)](#)

[Click here to hide information](#)

Even though missing main outcome data counts as a violation of observability, an internally valid ATE can still be recovered under modified assumptions. Most importantly, we require that response status (observable/non-observable) is unrelated to potential outcomes (i.e., randomly-occurring attrition). This assumption would be violated if, e.g., treated participants who do not benefit from an intervention are less likely to respond to a follow-up survey. While this assumption is not directly testable, researchers can instead test whether baseline outcomes differ systematically by treatment and response status. This could work if, for example, participants with low test scores *ex-ante* are the most likely to benefit from an educational intervention. Ghanem, Hirshleifer and Ortiz-Becerra (2020) recommend regressing baseline outcomes on indicators for treatment and response status. In case researchers do not have access to baseline outcome variables, they can instead check whether baseline covariates differ systematically by treatment and response status. Tests that do this include the selective attrition and the determinants of attrition test, with which researchers could, e.g., check for a relationship between gender and treatment/response status. This can make sense if gender is a predictor of potential outcomes. Finally, if the researchers lack access to baseline outcomes or covariates, they can run a rate of attrition test. This has researchers check whether attrition rates differ significantly by treatment condition. Some other alternatives include informal checks for non-random attrition, such as interviewing a sample of study dropouts to inquire about their reasons for attriting.

- ☐ Test for differences in baseline outcomes between attrited and non-attrited participants, e.g., Ghanem, Hirshleifer and Ortiz-Becerra (2020)
- ☐ Test for differences in covariates between attrited and non-attrited participants, e.g., selective attrition or determinants of attrition test
- ☐ Test for differences in attrition rate by experimental condition
- ☐ Informal check of attrition patterns or participant motivations for attrition, e.g., by surveying a sample of attritors
- ☐ None of the above; the authors do not run any tests or checks for non-random attrition



Extent of attrition (i)

Do the results of the statistical tests indicate random or non-random attrition? (Judg.)

[Click here to hide information](#)

In the previous block of questions, you indicated that the authors test or check for non-random attrition. Please answer here whether the results of their statistical tests are consistent with random or non-random attrition. For most tests, insignificant results indicate random attrition, e.g., when the authors find insignificant differences in baseline outcomes or covariates or attrition rates between responders and non-responders and/or between treatment conditions. By contrast, significant differences in these tests are usually indicative of non-random attrition. Please answer based on your best judgment about the statistical results that the authors present. You may use the typical significance level of 5% for this purpose. If many tests of differences are run, consider that some may appear significant by chance due to multiple hypothesis testing. The expected number of significant differences should be $c \times \alpha$, where c is the number of variables tested and α is the significance level. Applying this rule with $\alpha=0.05$ plus some approximate bounds, we might expect 0-2, 0-3, 0-6, and 1-9 significant differences for $c=10$, $c=20$, $c=50$, and $c=100$, respectively.

- ☐ Random attrition
- ☐ Non-random attrition



Nature of attrition (ii)

Based on your previous responses, this experiment receives a grade of **Gold** on the **nature of attrition** dimension.

The grades on this dimension are determined as follows:

- A grade of **Gold** is given to experiments for which there is perfect or near-perfect observability ($\geq 95\%$) of main outcomes among participants who received a randomized assignment.
- A grade of **Silver** is given to experiments for which the evidence from statistical tests indicates that attrition occurred randomly.
- A grade of **Bronze** is given to experiments for which the evidence from statistical tests indicates that attrition occurred non-randomly.
- A grade of **Insufficient** is given to experiments for which the authors do not run any statistical tests to check for non-random attrition, so the nature of potential attrition is unknown.

Based on these definitions, is the grade of **Gold** correct?

- ☐ Yes, it is correct
- ☐ No, it is incorrect



D List of coded papers

Table D.1: List of coded papers

No.	Title	Authors	Journal	Year	JEL code(s)
1	Accuracy and retaliation in repeated games with imperfect private monitoring: experiments	Yutaka Kayaba; Hitoshi Matsushima; Tomohisa Toyama	<i>GEB</i>	2020	C70; C71; C72; C73; D03
2	Adaptive learning versus punishment in ultimatum bargaining	Klaus Abbink; Gary E. Bolton; Abdolkarim Sadrieh; Fang-Fang Tang	<i>GEB</i>	2001	C78; C91; D83
3	Aggregation mechanisms for crowd predictions	Stefan Palan; Jürgen Huber; Larissa Senninger	<i>ExEc</i>	2020	C53; C83; G14
4	A gift with thoughtfulness: a field experiment on work incentives	Cangjian Cao; Sherry Xin Li; Tracy Xiao Liu	<i>GEB</i>	2020	C93; D9; J3
5	Altruism, fast and slow? Evidence from a meta-analysis and a new experiment	Hanna Fromell; Daniele Nosenzo; Trudy Owens	<i>ExEc</i>	2020	C91
6	A market-based mechanism for allocating space shuttle secondary payload priority	John Ledyard; David Porter; Randii Wessen	<i>ExEc</i>	2000	C92
7	An experimental analysis of intertemporal allocation behavior	Vital Anderhub; Werner Güth; Wieland Müller; Martin Strobel	<i>ExEc</i>	2000	C91; D81; D90
8	An experimental study of the effect of private information in the Coase theorem	Richard D. McKelvey; Talbot Page	<i>ExEc</i>	2000	C78; C92; D82
9	An experiment to evaluate Bayesian learning of Nash equilibrium play	James C. Cox; Jason Shachat; Mark Walker	<i>GEB</i>	2001	D83; C72; C92
10	A simple mechanism for the efficient provision of public goods: experimental evidence	Josef Falkinger; Ernst Fehr; Simon Gächter; Rudolf Winter-Ebmer	<i>AER</i>	2000	H41
11	Asset markets: how they are affected by tournament incentives for individuals	Duncan James; R. Mark Isaac	<i>AER</i>	2000	G12; D82
12	A test of game-theoretic and behavioral models of play in exchange and insurance environments	Cary A. Deck	<i>AER</i>	2001	C72; D81; G22
13	Auctions in near-continuous time	Cary Deck; Bart J. Wilson	<i>ExEc</i>	2020	D4; D8; C9
14	Behavior and learning in the “dirty faces” game	Roberto A. Weber	<i>ExEc</i>	2001	C91; C92; C72
15	Behavior in multi-unit demand auctions: experiments with uniform price and dynamic Vickrey auctions	John H. Kagel; Dan Levin	<i>ECMA</i>	2001	D44
16	Bribing the self	Uri Gneezy; Silvia Saccardo; Marta Serra-Garcia; Roel van Veldhuizen	<i>GEB</i>	2020	D03; D83; C91
17	Can intermediaries assure contracts? Experimental evidence	Michael Mitzkewitz; Tibor Neugebauer	<i>GEB</i>	2020	C91; D03
18	Can people form links to efficiently access information?	A Stefano Caria; Marcel Fafchamps	<i>EJ</i>	2020	D85; D90; O12
19	Can there be a market for cheap-talk information? An experimental investigation	Antonio Cabrales; Francesco Feri; Piero Gottardi; Miguel A. Meléndez-Jiménez	<i>GEB</i>	2020	D83; C72; G14
20	Cognition and behavior in normal-form games: an experimental study	Miguel Costa-Gomes; Vincent P. Crawford; Bruno Broseta	<i>ECMA</i>	2001	C72; C92; C51

Continued on next page

Table D.1 – continued from previous page

No.	Title	Authors	Journal	Year	JEL code(s)
21	Collective choice in the commons: experimental results on proposed allocation rules and votes	J. M. Walker; Roy Gardner; Andrew Herr; Elinor Ostrom	<i>EJ</i>	2000	C92; D70
22	Competition and entry in agricultural markets: experimental evidence from Kenya	Lauren Falcao Bergquist; Michael Dinerstein	<i>AER</i>	2020	L13; O13; Q11; Q12; Q13
23	Competitive preferences and ethnicity: experimental evidence from Bangladesh	Abu Siddique; Michael Vlassopoulos	<i>EJ</i>	2020	C9; C91; C93; J15
24	Compromise and coordination: an experimental study	Simin He; Jiabin Wu	<i>GEB</i>	2020	C72; C92
25	Concentration and variability of forecasts in artificial investment games: an online experiment on WeChat	Xiu Chen; Fuhai Hong; Xiaojian Zhao	<i>ExEc</i>	2020	C90; D83; D84; G11
26	Control without deception: individual behaviour in free-riding experiments revisited	Nicholas Bardsley	<i>ExEc</i>	2000	H41; C9; C92
27	Cooperation and punishment in public goods experiments	Ernst Fehr; Simon Gächter	<i>AER</i>	2000	H41
28	Cooperation in an overlapping generations experiment	Theo Offerman; Jan Potters; Harrie A.A. Verbon	<i>GEB</i>	2001	C72; C92; D90
29	Cooperation in stochastic games: a prisoner's dilemma experiment	Andrew Kloosterman	<i>ExEc</i>	2020	C73; C92
30	Correlated beliefs: predicting outcomes in 2 × 2 games	Timothy N. Cason; Tridib Sharma; Radovan Vadovič	<i>GEB</i>	2020	C72; C91; D84; D8
31	Corrupt police	Klaus Abbink; Dmitry Ryvkin; Danila Serra	<i>GEB</i>	2020	D73; K42; C92
32	Cortisol meets GARP: the effect of stress on economic rationality	Elena Cettolin; Patricio S. Dalton; Willem J. Kop; Wanqing Zhang	<i>ExEc</i>	2020	C90; C91; D01; D91
33	Costly voting: a large-scale real effort experiment	Marco Faravelli; Kenan Kalayci; Carlos Pimienta	<i>ExEc</i>	2020	C93; D72; C72
34	Cutthroat capitalism versus cuddly socialism: are Americans more meritocratic and efficiency-seeking than Scandinavians?	Ingvid Almås; Alexander W. Cappelen; Bertil Tungodden	<i>JPE</i>	2020	C93; D31; D63
35	Cycles of aggregate behavior in theory and experiment	Antoni Bosch-Doménech; María Sáez-Martí	<i>GEB</i>	2001	C7; C9; E3
36	Debates: voting and expenditure responses to political communication	Kelly Bidwell; Katherine Casey; Rachel Glennerster	<i>JPE</i>	2020	D72; D83; O17
37	Decentralized matching markets with(out) frictions: a laboratory experiment	Joana Pais; Ágnes Pintér; Robert F. Veszteg	<i>ExEc</i>	2020	C78; C91; D82
38	Decision making costs and problem solving performance	Tanga M. McDaniel; E. Elisabet Rutström	<i>ExEc</i>	2001	C91
39	Delegation and coordination with multiple threshold public goods: experimental evidence	Luca Corazzini; Christopher Cotton; Tommaso Reggiani	<i>ExEc</i>	2020	C91; C92; H40; H41; L31
40	Demand reduction in multiunit auctions: evidence from a sportscard field experiment	John A. List; David Lucking-Reiley	<i>AER</i>	2000	D44; H63
41	Determinants of trust: the role of personal experiences	Frederik Schwerter; Florian Zimmermann	<i>GEB</i>	2020	C91; D03; D81

Continued on next page

Table D.1 – continued from previous page

No.	Title	Authors	Journal	Year	JEL code(s)
42	Diffusion of being pivotal and immoral outcomes	Armin Falk; Thomas Neuber; Nora Szech	REStud	2020	C91; C92; D01; D02; D23; D63; D71
43	Discrimination in a segmented society: an experimental approach	Chaim Fershtman; Uri Gneezy	QJE	2001	D63; D70
44	Do economic inequalities affect long-run cooperation and prosperity?	Gabriele Camera; Cary Deck; David Porter	ExEc	2020	C70; C90; D03; E02
45	Does advertising serve as a signal? Evidence from a field experiment in mobile search	Navdeep S. Sahni; Harikesh S. Nair	REStud	2020	C93; D82; M3
46	Does information break the political resource curse? Experimental evidence from Mozambique	Alex Armand; Alexander Coutts; Pedro C. Vicente; Inês Vilela	AER	2020	C73; D72; D74; O13; O17; Q33; Q34
47	Does money illusion matter?	Ernst Fehr; Jean-Robert Tyran	AER	2001	E31; E52; E32; E51
48	Does poverty negate the impact of social norms on cheating?	Suparee Boonmanunt; Agne Kackaite; Stephan Meier	GEB	2020	C91; C93; D82; D91
49	Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards	John A. List	AER	2001	D44; Q26
50	Dominant strategy adoption and bidders' experience with pricing rules	Ronald M. Harstad	ExEc	2000	D44; C9
51	Durable goods, Coasian dynamics, and uncertainty: theory and experiments	Timothy N. Cason; Tridib Sharma	JPE	2001	L12; D82; C73; C92*
52	Dynamic inconsistency in food choice: experimental evidence from two food deserts	Sally Sadoff; Anya Samek; Charles Sprenger	REStud	2020	C91; D12; D81
53	Dynamic runs and circuit breakers: an experiment	Jacopo Magnani; David Munro	ExEc	2020	G02; G18; G01
54	Economic rationality under cognitive load	Andreas C Drichoutis; Rodolfo M Nayga, Jr.	EJ	2020	C91; D11; D12; D91; G11
55	Education quality and teaching practices	Marina Bassi; Costas Meghir; Ana Reynoso	EJ	2020	I21; I24; I25; I3
56	Efficiency versus gender roles and stereotypes: an experiment in domestic production	Hélène Couprie; Elisabeth Cuddeville; Catherine Sofer	ExEc	2020	D13; J16; J22; C91; C92
57	Elicitation of strategy profiles in large group coordination games	Darryl A. Seale; Amnon Rapoport	ExEc	2000	C72; C92; L13; D83*
58	Eliciting and utilizing willingness to pay: evidence from field trials in Northern Ghana	James Berry; Greg Fischer; Raymond Guiteras	JPE	2020	C93; D12; D82; L11; L31
59	Endogenous entry and exit in common value auctions	James C. Cox; Sam Dinkin; James T. Swarthout	ExEc	2001	D44; C92; C72; D82
60	Endogenous group formation and responsibility diffusion: an experimental study	Katharina Brütt; Arthur Schram; Joep Sonnemans	GEB	2020	C92; D64; D71
61	Escalation in conflict games: on beliefs and selection	Kai A. Konrad; Florian Morath	ExEc	2020	C90; D72; D74; D83

Continued on next page

Table D.1 – continued from previous page

No.	Title	Authors	Journal	Year	JEL code(s)
62	Estimating the production function for human capital: results from a randomized controlled trial in Colombia	Orazio Attanasio; Sarah Cattani; Emla Fitzsimons; Costas Meghir; Marta Rubio-Codina	AER	2020	I24; I28; J13; J24; O15
63	‘Everybody’s doing it’: on the persistence of bad social norms	David Smerdon; Theo Offerman; Uri Gneezy	ExEc	2020	Z10; D90; D70; C92
64	Experience and rationality under risk: re-examining the impact of sampling experience	Ilke Aydogan; Yu Gao	ExEc	2020	D81; D83; C91
65	Experimental evidence on the economics of rural electrification	Kenneth Lee; Edward Miguel; Catherine Wolfram	JPE	2020	L12; L94; O13; Q41
66	Experiments in high-frequency trading: comparing two market institutions	Eric M. Aldrich; Kristian López Vargas	ExEc	2020	D53; D44; D47; G12; C91; G14
67	Fairness considerations in joint venture formation	Tanjim Hossain; Elizabeth Lyons; Aloysius Siow	ExEc	2020	C92; D91; L14; D83
68	Fair shares: accountability and cognitive dissonance in allocation decisions	James Konow	AER	2000	D63
69	Feedback spillovers across tasks, self-confidence and competitiveness	Ritwik Banerjee; Nabanita Datta Gupta; Marie Claire Villeval	GEB	2020	C91; J15; M52
70	Fool me once: an experiment on credibility and leadership	David J Cooper; John R Hamman; Roberto A Weber	EJ	2020	C72; C92; D82
71	Framing and repeated competition	Aidas Masiliūnas; Heinrich H. Nax	GEB	2020	C72; C91; D43; D74; D83
72	Framing effects in public goods experiments	R. Cookson	ExEc	2000	H41; C92
73	Friendship and female education: evidence from a field experiment in Bangladeshi primary schools	Youjin Hahn; Asadul Islam; Eleonora Patacchini; Yves Zenou	EJ	2020	I25; J16; O12
74	From extreme to mainstream: the erosion of social norms	Leonardo Bursztyn; Georgy Egorov; Stefano Fiorin	AER	2020	D72; D85; Z13
75	Frustration and anger in the ultimatum game: an experiment	Chiara Aina; Pierpaolo Battigalli; Astrid Gamba	GEB	2020	C72; C91; D03
76	Funding public goods with lotteries: experimental evidence	John Morgan; Martin Sefton	REStud	2000	D44; H41
77	Gain-loss framing in interdependent choice	Susann Fiedler; Adrian Hillenbrand	GEB	2020	D91; C91; D64; D03*
78	Games played through agents in the laboratory – a test of Prat & Rustichini’s model	Ludwig Ensthaler; Steffen Huck; Johannes Leutgeb	GEB	2020	D44; C91; D72; D83
79	Gender-based focal points	Håkan J Holm	GEB	2000	J7; C7; C9
80	Gender differences in networking	Friederike Mengel	EJ	2020	C92; J16; J71
81	Gender differences in sabotage: the role of uncertainty and beliefs	Simon Dato; Petra Nieken	ExEc	2020	J16; M12; C91
82	God insures those who pay? Formal insurance and religious offerings in Ghana	Emmanuelle Auriol; Julie Lassébie; Amma Panin; Eva Raiber; Paul Seabright	QJE	2020	D14; G22; O12; O17; Z12
83	Good news and bad news: search from unknown wage offer distributions	James C. Cox; Ronald L. Oaxaca	ExEc	2000	J64; C91; D83

Continued on next page

Table D.1 – continued from previous page

No.	Title	Authors	Journal	Year	JEL code(s)
84	Habits as adaptations: an experimental study	Ludmila Matysková; Brian Rogers; Jakub Steiner; Keh-Kuan Sun	GEB	2020	C91; D8; D9
85	Health services as credence goods: a field experiment	Felix Gottschalk; Wanda Mimra; Christian Waibel	EJ	2020	C93; D82; I11; L00
86	Honesty and self-selection into cheap talk	Sebastian Fehrler; Urs Fischbacher; Maik T Schneider	EJ	2020	C92; D71; D83
87	Hot vs. cold: sequential responses and preference stability in experimental games	Jordi Brandts; Gary Charness	ExEc	2000	C92
88	How soon is now? Evidence of present bias from convex time budget experiments	Uttara Balakrishnan; Johannes Haushofer; Pamela Jakiela	ExEc	2020	C91; D90; O12
89	I (don't) like you! But who cares? Gender differences in same-sex and mixed-sex teams	Leonie Gerhards; Michael Kosfeld	EJ	2020	C90; J16
90	Improving college instruction through incentives	Andy Brownback; Sally Sadoff	JPE	2020	I23; M52; M55; C93
91	Inflation expectations and firm decisions: new causal evidence	Olivier Coibion; Yuriy Gorodnichenko; Tiziano Ropele	QJE	2020	E02; E03
92	Informational channels of financial contagion	Isabel Trevino	ECMA	2020	C7; C9; D8; G15
93	Information cascades: replication and an extension to majority rule and conformity-rewarding institutions	Angela A. Hung; Charles R. Plott	AER	2001	D83; D71
94	Information frictions and market power: a laboratory study	Anna Bayona; Jordi Brandts; Xavier Vives	GEB	2020	C92; D43; L13
95	In-group bias in prisons	Shiqi Guo; Pinghan Liang; Erte Xiao	GEB	2020	C93; D64; D91; K14
96	Iterative versus standard deferred acceptance: experimental evidence	Inácio Bó; Rustamdjan Hakimov	EJ	2020	C78; C92; D63; D78; D82
97	It's getting crowded in here: experimental evidence of demand constraints in the gender profit gap	Morgan Hardy; Gisella Kagy	EJ	2020	D22; J16; J31; L25; L26; L67; O12
98	Laboratory behavior in spot and forward auction markets	Owen R. Phillips; Dale J. Menkhaus; Joseph L. Krogmeier	ExEc	2001	D4; C9; C1
99	Learning to accept in ultimatum games: evidence from an experimental design that generates low offers	John A. List; Todd L. Cherry	ExEc	2000	C7; C9
100	Losing prosociality in the quest for talent? Sorting, selection, and productivity in the delivery of public services	Nava Ashraf; Oriana Bandiera; Edward Davenport; Scott S. Lee	AER	2020	H83; I11; I13; J24; M51; O15; Z13
101	Marketmaking in the laboratory: does competition matter?	Jan Pieter Krahnén; Martin Weber	ExEc	2001	C91; G14
102	Maternal depression, women's empowerment, and parental investment: evidence from a randomized controlled trial	Victoria Baranov; Sonia Bhalotra; Pietro Biroli; Joanna Maseko	AER	2020	G51; I12; J16; O15
103	Measuring and controlling for the compromise effect when estimating risk preference parameters	Jonathan P. Beauchamp; Daniel J. Benjamin; David I. Laibson; Christopher F. Chabris	ExEc	2020	B49; D03; D14; D83; G11

Continued on next page

Table D.1 – continued from previous page

No.	Title	Authors	Journal	Year	JEL code(s)
104	Measuring beliefs in an experimental lost wallet game	Martin Dufwenberg; Uri Gneezy	<i>GEB</i>	2000	C72; C92
105	Measuring “schmeduling”	Alex Rees-Jones; Dmitry Taubinsky	<i>REStud</i>	2020	D03; H20; H30
106	Measuring the indirect effects of adverse employer behaviour on worker productivity: a field experiment	Matthias Heinz; Sabrina Jeworrek; Vanessa Mertins; Heiner Schumacher; Matthias Sutter	<i>EJ</i>	2020	C93; J50; J63
107	Measuring trust	Edward L. Glaeser; David I. Laibson; José A. Scheinkman; Christine L. Soutter	<i>QJE</i>	2000	C9; D7; J7
108	Misperceived social norms: women working outside the home in Saudi Arabia	Leonardo Bursztyn; Alessandra L. González; David Yanagizawa-Drott	<i>AER</i>	2020	D83; J16; J22; O15; Z13
109	Moving to opportunity in Boston: early results of a randomized mobility experiment	Lawrence F. Katz; Jeffrey R. Kling; Jeffrey B. Liebman	<i>QJE</i>	2001	H43; I18; J18
110	Nash versus coarse correlation	Konstantinos Georgalos; Indrajit Ray; Sonali SenGupta	<i>ExEc</i>	2020	C72; C91; C92; D63; D83
111	Negotiating a better future: how interpersonal skills facilitate intergenerational investment	Nava Ashraf; Natalie Bau; Corinne Low; Kathleen McGinn	<i>QJE</i>	2020	D13; I24; J16; O15
112	Never underestimate your opponent: hindsight bias causes overplacement and overentry into competition	David Danz	<i>GEB</i>	2020	C91; D80; D82; D83; D84
113	Nonspeculative bubbles in experimental asset markets: lack of common knowledge of rationality vs. actual irrationality	Vivian Lei; Charles N. Noussair; Charles R. Plott	<i>ECMA</i>	2001	C90; E32; G10; G11
114	Norm enforcement in markets: group identity and the volunteering of feedback	Gary E Bolton; Johannes Mans; Axel Ockenfels	<i>EJ</i>	2020	C91; D4
115	(Not) alone in the world: cheating in the presence of a virtual observer	Jantsje M. Mol; Eline C. M. van der Heijden; Jan J. M. Potters	<i>ExEc</i>	2020	B41; C91; C93
116	On self-serving strategic beliefs	Nadja R. Ging-Jehli; Florian H. Schneider; Roberto A. Weber	<i>GEB</i>	2020	C72; D83; C92
117	On the empirical validity of axioms in unstructured bargaining	Noemí Navarro; Róbert F. Veszteg	<i>GEB</i>	2020	C78; C91; D63
118	On the relation between asset ownership and specific investments	Joep Sonnemans; Hessel Oosterbeek; Randolph Sloof	<i>EJ</i>	2001	C91; D23
119	On the roots of the intrinsic value of decision rights: experimental evidence	João V. Ferreira; Nobuyuki Hanaki; Benoît Tarrow	<i>GEB</i>	2020	C91; D03; D23
120	Outsourcing education: experimental evidence from Liberia	Mauricio Romero; Justin Sandefur; Wayne Aaron Sandholtz	<i>AER</i>	2020	H41; I21; I28; O15
121	Partners in crime? Corruption as a criminal network	Romain Ferrali	<i>GEB</i>	2020	C92; C93; D73; D85
122	Pathways of persuasion	Lucas Coffman; Paul Niehaus	<i>GEB</i>	2020	C90; C92; D03
123	Pay enough or don't pay at all	Uri Gneezy; Aldo Rustichini	<i>QJE</i>	2000	D1; D8; D9
124	Payoff-based belief distortion	Peiran Jiao	<i>EJ</i>	2020	C91; D81; D83

Continued on next page

Table D.1 – continued from previous page

No.	Title	Authors	Journal	Year	JEL code(s)
125	Peer effects with random assignment: results for Dartmouth roommates	Bruce Sacerdote	<i>QJE</i>	2001	I2; J0
126	Peers or police?: the effect of choice and type of monitoring in the provision of public goods	Gregory DeAngelo; Laura K. Gee	<i>GEB</i>	2020	C72; C91; C92; D7; H41
127	Perfect versus imperfect observability—an experimental test of Bagwell's result	Steffen Huck; Wieland Müller	<i>GEB</i>	2000	C72; C92
128	Player types and self-impression management in dictatorship games: two experiments	J. Keith Murnighan; John M. Oesch; Madan Pillutla	<i>GEB</i>	2001	C78; C91; D63
129	Preference discovery	Jason Delaney; Sarah Jacobson; Thorsten Moenig	<i>ExEc</i>	2020	D81; D83; D01; D03
130	Preordered service in contract enforcement	Jan U. Auerbach; Miguel A. Fonseca	<i>GEB</i>	2020	K00; K12; K40; O17; C92
131	Present-bias, procrastination and deadlines in a field experiment	Alberto Bisin; Kyle Hyndman	<i>GEB</i>	2020	D03
132	Quantifying the life-cycle benefits of an influential early-childhood program	Jorge Luis García; James J. Heckman; Duncan Ermini Leaf; María José Prados	<i>JPE</i>	2020	C93; I28; J13
133	Rational altruism? On preference estimation and dictator game experiments	Philip D. Grech; Heinrich H. Nax	<i>GEB</i>	2020	A13; C72; D01; D64
134	Rational inattention, competitive supply, and psychometrics	Andrew Caplin; Dániel Csaba; John Leahy; Oded Nov	<i>QJE</i>	2020	D83; D90
135	Reactions to (the absence of) control and workplace arrangements: experimental evidence from the internet and the laboratory	Katrin Schmelz; Anthony Ziegelmeyer	<i>ExEc</i>	2020	C81; C90; C93; M52
136	Rehabilitation and social behavior: experiments in prison	Loukas Balafoutas; Aurora García-Gallego; Nikolaos Georgantzis; Tarek Jaber-Lopez; Evangelos Mitrokostas	<i>GEB</i>	2020	C93; K14; K42
137	Reinforcement-based vs. belief-based learning models in experimental asymmetric-information games	Nick Feltovich	<i>ECMA</i>	2000	C72; C73; C92; D82; D83*
138	Robustness of the incentive compatible combinatorial auction	R. Mark Isaac; Duncan James	<i>ExEc</i>	2000	D44; C90
139	Room composition effects on risk taking by gender	Marco Castillo; Greg Leo; Ragan Petrie	<i>ExEc</i>	2020	C91; D81; J16
140	Seasonal liquidity, rural labor markets, and agricultural production	Günther Fink; B. Kelsey Jack; Felix Masiye	<i>AER</i>	2020	O13; O15; O18; Q11; Q12; R23
141	Self-serving cheap talk: a test of Aumann's conjecture	Gary Charness	<i>GEB</i>	2000	A13; C72; C91; D82; D84
142	Sentiments, strategic uncertainty, and information structures in coordination games	Michal Szkup; Isabel Trevino	<i>GEB</i>	2020	C72; C9; D82; D9
143	Silence is golden: team problem solving and communication costs	Gary Charness; David J. Cooper; Zachary Grossman	<i>ExEc</i>	2020	C91; C92; D03; D20
144	Social and strategic ambiguity versus betrayal aversion	Chen Li; Uyanga Turmunkh; Peter P. Wakker	<i>GEB</i>	2020	C72; D91; C81

Continued on next page

Table D.1 – continued from previous page

No.	Title	Authors	Journal	Year	JEL code(s)
145	Social norms with private values: theory and experiments	Giovanna d'Adda; Martin Dufwenberg; Francesco Passarelli; Guido Tabellini	<i>GEB</i>	2020	C91; D91
146	Stackelberg beats Cournot: on collusion and efficiency in experimental markets	Steffen Huck; Wieland Müller; Hans-Theo Normann	<i>EJ</i>	2001	L13; C92; C72; D63*
147	Status in markets	Sheryl Ball; Catherine Eckel; Philip J. Grossman; William Zame	<i>QJE</i>	2001	C90; A13; Z10
148	Step returns in threshold public goods: a meta and experimental analysis	Rachel T. A. Croson; Melanie Beth Marks	<i>ExEc</i>	2000	C9; C7; H4
149	Strategically delusional	Alice Soldà; Changxia Ke; Lionel Page; William von Hippel	<i>ExEc</i>	2020	C91; D03; D83
150	Strategic reasoning in persuasion games: an experiment	Ying Xue Li; Burkhard C. Schipper	<i>GEB</i>	2020	C72; C92; D82; D83
151	Strength of the social dilemma in a public goods experiment: an exploration of the error hypothesis	Marc Willinger; Anthony Ziegelmeyer	<i>ExEc</i>	2001	H41; C92; C72
152	Subsidy policies and insurance demand	Jing Cai; Alain de Janvry; Elisabeth Sadoulet	<i>AER</i>	2020	G22; G52; G53; Q54
153	Substitution and dropout bias in social experiments: a study of an influential social experiment	James Heckman; Neil Hohmann; Jeffrey Smith; Michael Khoo	<i>QJE</i>	2000	H43; C93; C14
154	Tackling youth unemployment: evidence from a labor market experiment in Uganda	Livia Alfonsi; Oriana Bandiera; Vittorio Bassi; Robin Burgess; Imran Rasul; Munshi Sulaiman; Anna Vitali	<i>ECMA</i>	2020	J2; M5
155	Targeted debt relief and the origins of financial distress: experimental evidence from distressed credit card borrowers	Will Dobbie; Jae Song	<i>AER</i>	2020	G56; K35
156	Teachers' pay for performance in the long-run: the dynamic pattern of treatment effects on students' educational and labour market outcomes in adulthood	Victor Lavy	<i>REStud</i>	2020	J24; J3
157	Ten little treasures of game theory and ten intuitive contradictions	Jacob K. Goeree; Charles A. Holt	<i>AER</i>	2001	C70
158	Testing for the emergence of spontaneous order	Konstantinos Georgalos; John Hey	<i>ExEc</i>	2020	C92; C72; L23; D83*
159	Testing models of social learning on networks: evidence from two experiments	Arun G. Chandrasekhar; Horacio Larreguy; Juan Pablo Xandri	<i>ECMA</i>	2020	C91; C92; C93; D83
160	Tests of fairness models based on equity considerations in a three-person ultimatum game	John H. Kagel; Katherine Willey Wolfe	<i>ExEc</i>	2001	D63; C78; C90
161	The attraction effect and its explanations	Geoffrey Castillo	<i>GEB</i>	2020	C91; D11; D80; D90
162	The benefit of the doubt: willful ignorance and altruistic punishment	Robert Stüber	<i>ExEc</i>	2020	C91; D01; D63; D83
163	The carrot vs. the stick in work team motivation	David L. Dickinson	<i>ExEc</i>	2001	H41; J4; J33; C92
164	The dynamics of motivated beliefs	Florian Zimmermann	<i>AER</i>	2020	C91; D83; D91

Continued on next page

Table D.1 – continued from previous page

No.	Title	Authors	Journal	Year	JEL code(s)
165	The dynamics of reorganization in matching markets: a laboratory experiment motivated by a natural experiment	John H. Kagel; Alvin E. Roth	<i>QJE</i>	2000	C78; C92; D47; J44*
166	The economic effects of Facebook	Roberto Mosquera; Mofioluwasademi Odunowo; Trent McNamara; Xiongfei Guo; Ragan Petrie	<i>ExEc</i>	2020	C93; D91; D83; I31
167	The effect of attending a small class in the early grades on college-test taking and middle school test results: evidence from Project Star	Alan B. Krueger; Diane M. Whitmore	<i>EJ</i>	2001	I2
168	The effect of collusion on efficiency in experimental auctions	Charles N. Noussair; Gyula Seres	<i>GEB</i>	2020	D44; D82; L41; C92
169	The effect of early-childhood education on social preferences	Alexander Cappelen; John List; Anya Samek; Bertil Tungodden	<i>JPE</i>	2020	C9; C93; D01
170	The effect of gender-targeted conditional cash transfers on household expenditures: evidence from a randomized experiment	Alex Armand; Orazio Attanasio; Pedro Carneiro; Valérie Lechene	<i>EJ</i>	2020	D12; D13; E21; O12
171	The effects of conflict budget on the intensity of conflict: an experimental investigation	Kyung Hwan Baik; Subhasish M. Chowdhury; Abhijit Ramalingam	<i>ExEc</i>	2020	C72; C91; D72; D74
172	The endowment effect and repeated market trials: is the Vickrey auction demand revealing?	Jack L. Knetsch; Fang-Fang Tang; Richard H. Thaler	<i>ExEc</i>	2001	D44; C91
173	The exposure problem and market design	Jacob K. Goeree; Luke Lindsay	<i>REStud</i>	2020	C78; C92; D47
174	The false consensus effect disappears if representative information and monetary incentives are given	Dirk Engelmann; Martin Strobel	<i>ExEc</i>	2000	D84; C91; D83
175	The first 2,000 days and child skills	Orla Doyle	<i>JPE</i>	2020	C93; D13; I26; J13
176	The formation of prosociality: causal evidence on the role of social environment	Fabian Kosse; Thomas Deckers; Pia Pinger; Hannah Schildberg-Hörisch; Armin Falk	<i>JPE</i>	2020	C90; D64
177	The impact of exchange context on the activation of equity in ultimatum games	Elizabeth Hoffman; Kevin McCabe; Vernon Smith	<i>ExEc</i>	2000	B40; C78; C92
178	The impact of management practices on employee productivity: a field experiment with airline captains	Greer K. Gosnell; John A. List; Robert D. Metcalfe	<i>JPE</i>	2020	D01; J3; Q5; R4
179	The joy of lottery play: evidence from a field experiment	Martijn J. Burger; Martijn Hendriks; Emma Pleeging; Jan C. van Ours	<i>ExEc</i>	2020	C93; D12; I31
180	The participation dividend of taxation: how citizens in Congo engage more with the state when it tries to tax them	Jonathan L. Weigel	<i>QJE</i>	2020	H20; P48; D73
181	The rationale of in-group favoritism: an experimental test of three explanations	Giuseppe Ciccaraone; Giovanni Di Bartolomeo; Stefano Papa	<i>GEB</i>	2020	A13; C91; D03; D64; D90
182	The relative income effect: an experiment	John Ifcher; Homa Zarghamee; Dan Houser; Lina Diaz	<i>ExEc</i>	2020	C91; D31; D63; I31
183	The sequential prisoner's dilemma: evidence on reciprocity	Kenneth Clark; Martin Sefton	<i>EJ</i>	2001	C72; C91; D91; D83*

Continued on next page

Table D.1 – continued from previous page

No.	Title	Authors	Journal	Year	JEL code(s)
184	The strength of weak leaders: an experiment on social influence and social learning in teams	Berno Buechel; Stefan Klößner; Martin Lochmüller; Heiko Rauhut	ExEc	2020	D83; D85; C91
185	The timing of discretionary bonuses – effort, signals, and reciprocity	Luke Boosey; Sebastian Goerg	GEB	2020	M5
186	The welfare effects of social media	Hunt Allcott; Luca Braghieri; Sarah Eichmeyer; Matthew Gentzkow	AER	2020	D12; D72; D90; I31; L82; L86; Z13
187	Time lotteries and stochastic impatience	Patrick DeJarnette; David Dillenberger; Daniel Gottlieb; Pietro Ortoleva	ECMA	2020	C91; D81; D90
188	Timing of communication	Puja Bhattacharya; Kirby Nielsen; Arjun Sengupta	EJ	2020	C72; C91; D83
189	Toward an understanding of the welfare effects of nudges: evidence from a field experiment in the workplace	Erwin Bulte; John A List; Daan van Soest	EJ	2020	C93; D03
190	Trust in state and nonstate actors: evidence from dispute resolution in Pakistan	Daron Acemoglu; Ali Cheema; Asim I. Khwaja; James A. Robinson	JPE	2020	C91; C93; D02; D73; D74; D83; K40; O17; P16
191	Trust, reciprocity, and social history: a re-examination	Andreas Ortmann; John Fitzgerald; Carl Boeing	ExEc	2000	C72; C91
192	Two experiments to test a model of herd behaviour	Louise Allsopp; John D. Hey	ExEc	2000	C91; C92
193	Underpricing of initial public offerings in experimental asset markets	Sascha Füllbrunn; Tibor Neugebauer; Andreas Nicklisch	ExEc	2020	C9; D02; D40; D83; G02; G32
194	Understanding gender differences in leadership	Sule Alan; Seda Ertac; Elif Kubilay; Gyongyi Loranth	EJ	2020	C91; C93; D03; I28
195	Upping the ante: the equilibrium effects of unconditional grants to private schools	Tahir Andrabi; Jishnu Das; Asim I. Khwaja; Selcuk Ozyurt; Niharika Singh	AER	2020	I21; I22; I25; I28; L22; L26; N75
196	Urban poverty and juvenile crime: evidence from a randomized housing-mobility experiment	Jens Ludwig; Greg J. Duncan; Paul Hirschfield	QJE	2001	K42; I30
197	Value orientations, income and displacement effects, and voluntary contributions	Neil Buckley; James Chowhan; Mohamed Shehata; Kenneth S. Chan; Stuart Mestelman	ExEc	2001	H41; C91; Z13
198	Voting on the threat of exclusion in a public goods experiment	Astrid Dannenberg; Corina Haita-Falah; Sonja Zitzelsberger	ExEc	2020	H41; D02; C91; C92; C72; D71
199	What makes a rule complex?	Ryan Oprea	AER	2020	C73; D11; D12; D83; D91
200	What you see is all there is	Benjamin Enke	QJE	2020	D03; D80; D84
201	When does less information translate into more giving to public goods?	Billur Aksoy; Silvana Krasteva	ExEc	2020	H41; C72; C90
202	When fair isn't fair: understanding choice reversals involving social preferences	James Andreoni; Deniz Aydin; Blake Barton; B. Douglas Bernheim; Jeffrey Naecker	JPE	2020	D03; D63

Continued on next page

Table D.1 – continued from previous page

No.	Title	Authors	Journal	Year	JEL code(s)
203	Where to look for the morals in markets?	Matthias Sutter; Jürgen Huber; Michael Kirchler; Matthias Stefan; Markus Walzl	ExEc	2020	C92; D03; D62
204	Workplace knowledge flows	Jason J. Sandvik; Richard E. Saouma; Nathan T. Seegert; Christopher T. Stanton	QJE	2020	D8; J3; L2; M5
205	Would depositors pay to show that they do not with- draw? Theory and experiment	Markus Kinatader; Hubert János Kiss; Ágnes Pintér	ExEc	2020	C72; C91; D80; G21

Notes: The table lists all 205 papers in the coded sample and their recorded authors, publication outlets, publication years, and JEL codes. Papers are listed in alphabetical order. Journals are abbreviated as AER (*American Economic Review*), ECMA (*Econometrica*), EJ (*Economic Journal*), ExEc (*Experimental Economics*), GEB (*Games and Economic Behavior*), JPE (*Journal of Political Economy*), QJE (*Quarterly Journal of Economics*), and REStud (*Review of Economic Studies*). Multiple JEL codes may be reported for one paper. JEL codes were taken from the published version of the paper if available there, or a previous working paper version if available there. In total, 8 papers reported no JEL codes in either the published version or any working paper version. These are classified by us and marked with an asterisk (*) next to their JEL codes.