Computational Reproducibility in Finance: Evidence from 1,000 Tests

Christophe Pérignon [1,2,*] Olivier Akmansoy [2,3] Christophe Hurlin [2,4]
Anna Dreber [5,6] Felix Holzmeister [6] Jürgen Huber [6] Magnus Johannesson [5]
Michael Kirchler [6] Albert J. Menkveld [7,8] Michael Razen [6] Utz Weitzel [7,8,9]

[1] HEC Paris, France [2] cascad, France [3] CNRS, France [4] University of Orléans, France [5] Stockholm School of Economics, Sweden [6] University of Innsbruck, Austria [7] Vrije Universiteit Amsterdam, Netherlands [8] Tinbergen Institute, Netherlands [9] Radboud University, Netherlands

[*] *Corresponding author: perignon@hec.fr*

This draft: April 14, 2023

Abstract

We analyze the computational reproducibility of more than 1,000 empirical answers to six research questions in finance provided by 168 international research teams. Running the original researchers' code on the same raw data regenerates exactly the same results only 52% of the time. Reproducibility is higher for researchers with better coding skills and for those exerting more effort. It is lower for more technical research questions, more complex code, and for outlier results. Neither researcher seniority, nor peer-review ratings appear to be related to the level of reproducibility. Moreover, researchers exhibit strong overconfidence when assessing the reproducibility of their own research. We provide guidelines for finance researchers and discuss several implementable reproducibility policies for academic journals.

JEL: C80, C87
Keywords: Market microstructure; open science; credibility of research; multi-analyst study; data-availability policy

We are grateful for their valuable comments to Bruno Biais, Abel Brodeur, Jean-Edouard Colliard, Daniel Kahneman, Joan Llull, Ted Miguel, Olivier Sibony, Lars Vilhuber, to participants to the 2022 ECODEC workshop, 2022 French Finance Association meeting, 2022 EH Lausanne Finance & Economics Conference, 2023 FIBSF conference, and 2023 BITSS annual meeting (UC Berkeley), and to seminar participants at HEC Paris. The authors are grateful for financial support from (Dreber) the Knut and Alice Wallenberg Foundation, the Marianne, Marcus Wallenberg Foundation, the Jan Wallander, Tom Hedelius Foundation, (Huber) FWF grant P29362, (Huber and Kirchler) FWF SFB F63, (Hurlin and Pérignon) ACPR Chair in Regulation and Systemic Risk and ANR grants Ecodec ANR-11-LABX-00474, MLEforRisk ANR-21-CE26-0007, CaliBank ANR-19-CE26-0002, (Johannesson) Riksbankens Jubileumsfond grant P21-0168, and (Menkveld) NWO-Vici.

"Non-reproducible single occurrences
are of no significance to science."

—Popper (1959)

1. Introduction

Finance benefits from reanalysis studies, as they reinforce its scientific nature and build trust (Harvey, 2017, 2019; Welch, 2019). Overlooked for decades, reanalyses have recently been given more attention in leading finance journals. Indeed, most top finance journals have introduced and enforced stricter data availability policies (see the policies of the *Journal of Finance*, the *Review of Financial Studies*, the *Journal of Financial Economics*, and the *Review of Finance*). Furthermore, the *Journal of Finance* promoted a section dedicated to reanalyses (Nagel, 2019) and retracted a non-reproducible article (Nagel, 2021). Several authors have recently challenged the validity and robustness of some classic empirical results in corporate finance (Mitton, 2021; Cohn et al., 2023) and in asset pricing (Harvey et al., 2016; McLean and Pontiff, 2016; Linnainmaa and Roberts, 2018; Hou et al., 2020). In contrast, Chen and Zimmermann (2022) and Jensen et al. (2022) depict a much more positive view of the validity of most of the anomalies reported in empirical asset pricing.

There are two main types of reanalyses: reproductions and replications (National Academies of Sciences, Engineering, and Medicine, 2019; Welch, 2019). A reproduction consists of an attempt to regenerate the same result in the same sample with the same method, whereas a replication does so by changing either the sample, the method, or both. As shown in Table 1, several types of reanalysis can be considered in practice. One can first verify whether the regenerated result $\beta_{\boldsymbol{R}}$ obtained by running the original code on the original sample is equal to the original result β. This strict definition, called computational reproducibility (Buckheit and Donoho, 1995; Peng, 2011), is related to the concepts of code/data sharing and code/data quality (Trisovic et al., 2022).

Another definition consists of verifying whether the original and regenerated results are similar: i.e., same sign, same magnitude, and same statistical significance. Furthermore, when the original code or dataset are not available, one needs to write a new code ($c_{\boldsymbol{R}}$) or reconstruct the exact

same sample from the original database (CRSP, Compustat, etc). The latter definition is the most common definition of reproducibility in finance (Welch, 2019; Chen and Zimmermann, 2022). Alternatively, when testing the replicability of a result, one can test whether the original and regenerated results are similar when the data ($d_{\boldsymbol{R}}$), the method ($m_{\boldsymbol{R}}$), or both are modified.

Table 1: Typology of reanalyses. This table displays several types of reanalyses used in economics and finance. In each case, the test is whether the original result $\beta(d, m, c)$ in a given article can be regenerated by another researcher using the same or different data (d vs. d_R), methods (m vs. m_R), and code (c vs. c_R). Depending on the definition, the test is whether the original result (β) and the regenerated result (β_R) are either equal ($=$) or similar ($\approx$).

Concepts	*Definitions*	*Examples*
Reproduction	$\beta_{\boldsymbol{R}}(d, m, c) = \beta(d, m, c)$	Vilhuber (2022), *this study*
	$\beta_{\boldsymbol{R}}(d, m, c) \approx \beta(d, m, c)$	Chang and Li (2017)
	$\beta_{\boldsymbol{R}}(d, m, c_{\boldsymbol{R}}) \approx \beta(d, m, c)$	Chen and Zimmermann (2022)
Replication	$\beta_{\boldsymbol{R}}(d_{\boldsymbol{R}}, m, c_{\boldsymbol{R}}) \approx \beta(d, m, c)$	McLean and Pontiff (2016)
	$\beta_{\boldsymbol{R}}(d, m_{\boldsymbol{R}}, c_{\boldsymbol{R}}) \approx \beta(d, m, c)$	Cohn et al. (2023)
	$\beta_{\boldsymbol{R}}(d_{\boldsymbol{R}}, m_{\boldsymbol{R}}, c_{\boldsymbol{R}}) \approx \beta(d, m, c)$	Jensen et al. (2022), Mitton (2021), and Hou et al. (2020)

In this paper, we conduct a large-scale empirical analysis of the level of computational reproducibility in finance. As in most computational sciences, this property acts as a minimum standard that is expected for any scientific result (Duflo and Hoynes, 2018; Christensen and Miguel, 2018; Welch, 2019). When a result is reproducible, it is easier for other scientists to consider richer concepts such as replication, robustness, or sensitivity-analysis, which examine the general validity of a scientific result.[1] In practice, testing for computational reproducibility is challenging because the code and data associated with publications are not always available in finance. Reasons include the lack of obligations, incentives, or tradition to share and the frequent use of restricted or copyrighted data (Gertler et al., 2018; Colliard et al., 2022). While not yet the norm in finance, the systemic verification of the computational reproducibility of the results of conditionally-accepted papers is now required in many leading academic journals including the

[1] This is emphasized by statements like *"reproducibility of research findings is a basic principle of science and a prerequisite for replicability"* (Nagel, 2018) or "*the best starting point for a replication is always a reproduction. They are the foundation of science*" (Welch, 2019).

American Economic Review, *American Economic Journal*, *Econometrica* (starting Summer 2023), *Review of Economic Studies*, *Economic Journal*, *Journal of the American Statistical Association*, or the *American Journal of Political Science*.

Unlike existing studies that aim to reproduce published papers, we rely on a crowdsourcing study referred to as the *Finance Crowd Analysis Project*, or in short #*fincap* (Menkveld et al., 2023). In the context of this project, 168 research teams from 37 countries were instructed to analyze the same dataset containing 720 million futures transactions to answer the same six research questions. Using the teams' replication kits (common dataset, teams' computer code and readme file), a single verificator attempted to reproduce the results provided by all teams for all the research questions, for a total of $168 \times 6 = 1,008$ empirical findings. For each finding, the verificator produces a *reproducibility score* by contrasting original and regenerated results. The score can take five values: 100 (perfect accuracy), 75 (only small differences), 50 (one large difference), 25 (several large differences), and 0 (no result generated).

When attempting to run a piece of code in our experiment, two outcomes could arise. For 53.9% of the results in our study, the code ran smoothly and for 46.1% of the results, it did not. In the latter cases, the verificator made some changes to the code and/or to the computing environment. When he was still unsuccessful, the situation was systematically discussed with two tenured professors specialized in econometrics, finance, and computational reproducibility. In some cases, the verificator also sought some technical help from the IT team of his university. Besides all the effort exerted and help received, the verificator was unable to make the code run and to produce any result in 29.1% of the cases. At the end of the process, the distribution of the reproducibility scores is: 52% of 100, 11.3% of 75, 2.5% of 50, 5.2% of 25, and 29.1% of 0, and the average reproducibility score is 63. When only focusing on the code that we have been able to run, the average score is 88.8.

Is the glass half empty or half full? A positive interpretation of our results is that our success rate is larger than those reported in the economics literature (Chang and Li, 2017; Gertler et al., 2018) and relatively close to the high success rates on stock market anomalies reported by Chen and Zimmermann (2022) and Jensen et al. (2022). This is particularly remarkable because the studies

we aim to reanalyze are in the market microstructure field, which is known for its huge datasets and complex data processing. Furthermore, one could say that our conclusion can be viewed as a lower bound as someone with more skills, time, and resources than us could reach an even higher success rate. However, a less positive interpretation is that reproduction kits do not always run, and when they do, they do not always produce the results reported in the corresponding paper. This is an important problem as hard-to-run code imposes a cost on other researchers who try to reproduce the results. Similarly, nonworking code and result discrepancies create negative externalities and can lower the trust put in these results, and in finance in general.

To investigate why some empirical results can be reproduced while others cannot, we proceed in three main steps. First, we classify all the problems we faced during our 1,008 reproducibility attempts, both the problems we have been able to solve, and those we have been unable to solve. To do so, we create a typology of 20 types of problems related to the readme file, software, hardware, code, and data. We find that, collectively, these issues prevent us from reproducing 46.1% of the results, of which 17 percentage points could be solved while 29.1 percentage points could not. Approximately half of the problems impacting the code were solved by the verificator, which allowed us to successfully regenerate another 13.5% of the results. In contrast, none of the problems affecting the readme files (4.8%) and those affecting computational capacity (3.9%) could be resolved. To test the external validity of our results, we analyzed 48 replication kits associated with papers published in leading economics journals. We report a comparable reproducibility rate and similar bugs and problems as in our original sample.

Second, we study the cross-section of computational reproducibility among the various teams by considering the following six dimensions of scholarship: (i) the researchers' characteristics and skills, (ii) the type of research question addressed, (iii) the software used, (iv) the complexity of the computer code, (v) the quality of the research, and (vi) the effort exhibited by researchers to make their results reproducible. There are some important differences between these families of variables. On the one hand, variables (i)–(iii) are exogenous and permit to test some intuitive economic channels. For instance, engaging in reproducible research can be less desirable for researchers with high opportunity costs (e.g., for tenure-track researchers), but it can also be more desirable (e.g., for top-tier scholars) because it lowers reputation risk (Colliard et al., 2022).

On the other hand, variables (iv)–(vi) are expected to covary with the level of reproducibility but do not necessarily have a causal effect on it.

In the cross-section, we find that the level of reproducibility is much lower for answers that lie in the tails of the result distribution. This finding points toward a positive relationship between reproducibility and research quality. Indeed, by design, outlier results contribute more to the dispersion of the results across teams (level noise in Kahneman et al., 2021) and can be viewed as a proxy for lower research quality. Furthermore, we show that our measure of computational reproducibility tends to be higher when the original researchers have better coding skills and lower when the research questions are more technically challenging. We also find that code complexity (respectively, the quality of the documentation prepared by the researchers) covaries negatively (positively) with the level of reproducibility. We do not find evidence that peer-review ratings or the academic quality of the researchers, as proxied by seniority, top publications, and citations, has an impact on reproducibility. Other variables that show no effect include software, location, and the presence of a coauthor.

Third, the researchers in our study seem to exhibit strong overconfidence when assessing the level of reproducibility of their own research ex ante. More than 70% of the teams indicate that one could find the exact same results by running their code on the initial dataset, whereas only approximately 30% of the papers are actually fully reproducible. Fourth, we show that participants also severely underestimate the difficulty faced by their peers when attempting to reproduce their findings. Close to 95% of the teams claimed that regenerating their results would be either "*straightforward*" or "*quite easy*". However, the reproducibility verificator disagreed with this self-assessment and assigned these favorable ratings to only 62% of the teams.

In the final section, we draw some implications for researchers and academic journals. Specifically, we provide guidelines for researchers in finance to increase the reproducibility of their own research. Our guidelines are informed by the empirical findings of our study, our own experience as reproducibility verificators, and discussions with data editors at leading economic journals. We also compare alternative policies that finance journals could implement to ensure the research they publish is computationally reproducible.

This paper adds to the growing literature on the reproducibility and credibility of academic research in economics and finance (see the surveys of Christensen and Miguel (2018) and Colliard et al. (2022)). Early evidence was provided by Dewald et al. (1986) and McCullough et al. (2006) for the *Journal of Money, Credit and Banking* and by McCullough and Vinod (2003) and Glandon (2011) for the *American Economic Review*. In a study of 67 articles published in a dozen well-regarded economics journals, Chang and Li (2017) were able to reproduce the results from one-third of these papers using the code and data available in the journals' repositories. In the same vein, Gertler et al. (2018) considered a sample of 203 empirical papers published in nine leading economics journals that did not use any proprietary or otherwise restricted data. They were able to produce final tables and figures from the raw data for only 14% of these 203 studies. Recently, Herbert et al. (2021) attempted to regenerate the findings of all 303 articles published in the *American Economic Journal: Applied Economics* between 2009 and 2018 using dozens of Cornell undergraduate students as verificators. After excluding papers that used confidential data or papers with either no or incomplete data, the students were able to fully reproduce approximately 42% of the papers.

In finance, we are not aware of similar large-scale studies based on original scripts and data. However, most replication studies in finance start with a reproduction. Recent examples include the successful reproductions of Acharya and Pedersen (2005) by Holden and Nam (2019) and Kazumori et al. (2019), of Amihud (2002) by Drienko et al. (2019) and Harris and Amato (2019), or of Pástor and Stambaugh (2003) by Li et al. (2019) and Pontiff and Singla (2019). More recently, Chen and Zimmermann (2022) successfully reproduce the findings of a large number of market anomalies papers. For the 161 characteristics that were clearly significant in the original papers, 98% of their regenerated long-short portfolios lead to t-stats above 1.96. Similarly, Jensen et al. (2022) report a high degree of internal validity of prior research on asset pricing factors. These findings contrast with those of Hou et al. (2020) who find that approximately half of the literature cannot be reproduced when applying the same methods to the same data and same sample periods. They highlight the key role played in the findings by small caps and the weighting schemes used to compute returns.

To conduct such reanalyses, researchers have to reconstruct the original dataset and to rewrite the

script. In practice, these steps can be challenging because some ambiguity may exist regarding the data preparation and analysis, or because data providers sometimes alter the raw data. Unlike the aforementioned reanalyses in finance, our paper focuses on computational reproducibility and only relies on original scripts and raw data. This approach allows us to speak about code/data availability, code/data quality, and the (lack of) efficiency of the data and code availability policies in place in academic journals. Moreover, our findings on the strong positive relationship between research quality and computational reproducibility suggest that more attention should be paid to the latter.

2. Experiment and Data

Several components are required to investigate the level and drivers of computational reproducibility. We need a sample of research papers that report some empirical tests, as well as the code and data used to generate the reported results. In addition, we require detailed information about the authors, the associated computer code, and the research piece itself. As shown below, we collected all these elements in the context of the #*fincap* study.

2.1. The Finance Crowd Analyses Study (#*fincap*)

Most of the data used in our paper come from the #*fincap* study, which is the first multi-analyst study in empirical finance (Menkveld et al., 2023). It aims to analyze the level and dynamics of heterogeneity in empirical results when *different* researchers test the *same* research questions using the *same* dataset. The findings by Menkveld et al. indicate that the variability in results that is due to researchers choosing different analysis pipelines adds uncertainty—which they refer to as *non-standard errors*—to the evidence-generating process. In particular, they show that non-standard errors are comparable in size with standard errors, significantly decline after peer feedback, and are substantially underestimated by participants.

The teams in #*fincap* consisted of one or two researchers, with at least one member holding a Ph.D. degree in finance or economics. In addition, the included teams were required to be

sufficiently skilled in the field of empirical finance, to have an understanding of market liquidity, and to be familiar with the analysis of large datasets. Based on a survey completed by all team members upon their application to join the project, the project coordinators decided whether each team was sufficiently qualified to participate.[2]

The dataset used in *#fincap* was provided by Deutsche Börse and contains 720 million trade records of the EuroStoxx 50 index futures between 2002 and 2018. Based on these data, teams were asked to test six research questions on trends in (i) market efficiency, (ii) the realized bid-ask spread, (iii) the share of client volume in total volume, (iv) the realized spread on client orders, (v) the share of market orders in all client orders, and (vi) the gross trading revenue of clients.[3] For each research question, the teams were required to report both an effect size estimate (measured in terms of the average annualized percentage change in the dependent variable) and an estimate of the corresponding standard error. Together with a short paper summarizing the results of their analyses, the teams were required to submit the analysis scripts that generated their results and a readme file that outlined how to reproduce their estimates.[4] Each paper was evaluated by two anonymous peer evaluators who rated the quality of the papers and provided feedback for making improvements.[5]

The controlled nature of the *#fincap* project provides several advantages. First, as by design, *#fincap* removes the major impediments for reproducing results (i.e., lack of data and code), this project is well suited to identify some of the drivers and barriers of computational reproducibility. Second, given its multi-research question structure, it leads to many more data points (1,000+) than in all the aforementioned studies on the reproducibility of economics or finance research.

[2] A total of 259 teams registered to participate in *#fincap*. Out of them, 231 teams fulfilled the participation requirements and were accepted into the project. A total of 223 teams complied with the requirement to sign the project's non-disclosure agreement and were provided access to the dataset. Eventually, 168 teams completed the required analysis by the due date.

[3] The instructions provided to the research teams, including the verbatim phrasing of the six research questions, are provided in Appendix A.

[4] Although teams were not explicitly informed that their results would be verified, they could infer from the requirement to submit their computer code that their results might be verified. Therefore, the experiment simulated a regime similar to the one currently used by all top-3 finance journals in which authors must submit their code, but they do not know for sure whether someone will ever run it to verify the results.

[5] The assessment was single-blind as the evaluators could see the names of the authors but their names were unknown to the authors. Reputation-wise, this creates strong incentives for authors to exert effort as both their names and their paper are seen by two senior members of the academic finance community (see Section 2.3 for details).

Another important aspect is that we have multiple teams addressing the same research questions using the same database, which is a truly unique feature since in the literature different observations are drawn from papers on different topics. Third, the research teams in #*fincap* are of high quality, in terms of, for instance, tenure and number of publications in the top finance or economics journals. They are also quite representative of the finance research community, in terms of software, gender, and coauthorship.

2.2. Reproducibility Scores

For each team and each research question, the team's analysis script and readme file were used to generate a specific reproducibility score. The latter is obtained at the end of a rigorous reproducibility assessment conducted by *cascad*, which stands for Certification Agency for Scientific Code And Data (www.cascad.tech). It is a nonprofit academic organization that helps individual researchers signal the reproducibility of their research (pre-submission check) and helps academic journals verify the reproducibility of the research they publish (pre-publication check).[6] In practice, this verification was conducted by a single reproducibility verificator working for *cascad*, under the supervision of two reproducibility editors. Having a single verificator for all reproductions brings the benefit that heterogeneity in reproduction outcomes attributable to verificator fixed effects is ruled out by design.

For each research question addressed by a given team, the reproducibility score was determined at the end of the following process. After downloading the common dataset, the verificator attempted to run the computer code by following the various steps indicated in the readme file.[7] The computation was performed either on a workstation or on a virtual machine, depending on

[6] *cascad* regularly acts as a third party to verify the results of papers conditionally accepted by the *American Economic Review*, *American Economic Journal (AEJ): Applied Economics*, *AEJ: Economic Policy*, *AEJ: Macroeconomics*, *AEJ: Microeconomics* (Vilhuber, 2021, 2022), and by the *Economic Journal*.

[7] As in most previous studies (see, e.g., Gertler et al., 2018; Herbert et al., 2021), the reproducibility verificator was prevented from seeking advice or help from the original authors. Another reason for not contacting the researchers was to not interfere with the course of the #*fincap* study.

availability and the required environment (PC, Mac, Linux).[8] To deliver the reproducibility scores in a timely manner and not delay the #*fincap* study, the verificator stayed within a one-week time budget for each team. As shown in Table 4, this time limit corresponds to 18 times the average CPU time and 50 times the median CPU time across all teams. Because of the time limit, no results were generated for four teams.

Then, both the regenerated effect size estimates and standard errors were displayed in a team-specific execution report. This report also included comments about potential problems encountered during the computation phase. For each parameter (effect size and standard-error), the difference, if any, between the original result and the regenerated result was either *small* ($\leq 10\%$ of the original result) or *large* ($> 10\%$ of the original result). Finally, for each research question, the reproducibility score could take one of five values ranging between 0 and 100, which were set as shown in Table 2.

Table 2: Reproducibility ratings and scores. Scores are assigned independently for each of the six research questions addressed by the research teams. *Difference* refers to the potential discrepancies between the original and the regenerated results (size and standard-error): *small* indicates differences smaller than 10% of the original result; *large* indicates differences strictly larger than 10%. For the RR rating, we can detect a difference on one or two parameters (size and/or standard-errors).

Reproducibility Rating	*Reproducibility Score*	*Difference*
RRR	100	None, perfect accuracy
RR	75	Only small differences
R	50	One large difference
D	25	Several large differences
DD	0	No result generated

In our analyses, we use two versions of the reproducibility scores. The first is a binary variable, which contrasts full reproducibility (score of 100) with no/partial reproducibility (score lower than 100). The second one is an ordinal variable with five levels (0, 25, 50, 75, and 100). This five-

[8] The workstation was equipped with 64GB RAM, Intel® Core™ i9-9900K CPU @3.60-5.00GHz, Nvidia Geforce RTX 2060, and Windows 10. It includes R 4.0.5, Matlab R2019b and R2020a, Python 3.9.2 (through Anaconda 2020.11), Stata 16.1, SAS 9.4, Haskell, and PostgreSQL. In addition, we used two Microsoft Azure virtual machines: (1) one machine with Windows 10, 32 cores, 512GB RAM, and with the following software: Microsoft SQL Server, R 4.0.4, Matlab R2020a, Python 3.9.2 (through Anaconda 2020.11), and Julia v1.5.4 and (2) one machine with Linux (Ubuntu distribution), 32 cores, 128GB RAM, and with the following software: R 4.0.5, Python 3.9.2 (through Anaconda 2020.11), and Java SDK 16.0.1.

notch scale is consistent with the one used during the internal reproducibility audit conducted by the *American Economic Review* (Glandon, 2011).

2.3. Covariates

To identify the drivers of the reproducibility scores, we build a rich dataset of variables that characterize several facets of scholarship. To ease exposition, we organize these variables into six categories, namely, variables related to (i) the researchers' characteristics and skills, (ii) the type of research questions, (iii) the software used, (iv) the complexity of the computer code, (v) the quality of the research, and (vi) the effort exhibited by researchers to make their results reproducible. A comprehensive description of each variable is provided in Table B1 in Appendix B.

Researchers' characteristics and skills. We characterize the *academic quality* of a team using four variables. The first two are binary variables that indicate whether the team includes at least one tenured faculty member (associate or full professor) and whether the team includes at least one researcher who has already published in a top-5 economics or top-3 finance journal. In our sample, 52.4% of the teams include at least one tenured faculty member, and 42.9% include a researcher with at least one top publication. Another proxy that we use for academic excellence is the number of Google Scholar citations (maximum number across the team members). The average number of citations—as self-reported by all participants upon registration for #*fincap* in December 2020—is 1,595, with a maximum of 29,000. To assess expertise in the field, we ask each participant to self-assess his or her experience with market liquidity and empirical finance using a scale ranging from 0 to 10. The average score is 8.23 ($sd = 1.39$). Overall, these indicators show that the academic quality of the participants is quite good.

To measure the *coding and data handling skills* of the research teams, we take two actions. First, we ask the team members about the size of the largest database they have worked with and about their self-assessment of their coding skills (low, average, high, excellent). A total of 72.6% of the teams state that they have direct experience with datasets that are comparable in size to the dataset analyzed in #*fincap*, and 32.1% report that their coding skills are excellent. Second, we

Table 3: Summary statistics on researchers and software. m and sd denote the mean and the standard deviation, respectively; p_{25}, p_{50}, and p_{75} indicate the 25th, the 50th, and the 75th percentile; min and max denote the minimum and maximum values. Variables marked with * are dichotomous, i.e., means (m; reported as percentages) correspond to the fraction of "successes;" variables not marked with * are metrics. Superscript [a] indicates that the data were elicited on the individual level in #*fincap*'s entry survey and is, thus, based on individual researchers' self-reports; team-level aggregates correspond to the maximum value per team. Superscript [b] indicates that the data were recorded on the team level in the course of *cascad*'s evaluation process. $n = 168$ for all variables.

Variable	m	sd	min	p_{25}	p_{50}	p_{75}	max
Academic quality:							
» Seniority (assoc. / full professor) [a,*]	52.4%						
» Top publication (top-3 / top-5) [a,*]	42.9%						
» Citations (in thousands) [a]	1.59	3.62	0.00	0.05	0.30	1.42	29.00
» Expertise (emp. fin. & liquidity) [a]	8.23	1.39	3.50	7.50	8.50	9.00	10.00

Variable	m
Coding and data handling skills:	
» Experience with large data [a,*]	72.6%
» Excellent coding skills [a,*]	32.1%
» Parallel computing [b,*]	10.1%
» Loops/matrix operations [b,*]	88.7%
Team composition:	
» Coauthor (team of two) *	78.6%
» Gender (female) [a,*]	29.2%
Location:	
» North America [a,*]	25.0%
» Europe [a,*]	63.1%
» Asia-Pacific [a,*]	15.5%

Variable	m
Software used:	
» Eviews [b,*]	1.8%
» Excel [b,*]	1.8%
» GAUSS [b,*]	0.6%
» Haskell [b,*]	0.6%
» Java [b,*]	0.6%
» Julia [b,*]	0.6%
» Matlab [b,*]	11.9%
» OneTick [b,*]	0.6%
» Python [b,*]	29.2%
» R [b,*]	28.6%
» SAS [b,*]	22.0%
» SQL/PostgreSQL [b,*]	3.6%
» Stata [b,*]	29.2%
» Visual Basic [b,*]	0.6%

measure their skills ourselves by scrutinizing their code. In particular, we systematically check whether the researchers use parallel computing techniques (10.1%) and whether they use loops or matrix operations (88.7%).

We also consider the size of the teams (one vs. two members). In our sample, 78.6% of the teams consist of two researchers, which is quite close to what is observed in practice in the finance field. For instance, Grossmann and Lee (2022) report that the percentage of single-authored papers published in the *Journal of Finance*, the *Journal of Financial Economics*, and the *Review of Financial Studies* between 2015 and 2019 is 16.2%, 13.0%, and 11.3%, respectively.

Other features we control for are gender and location. In the sample, 29.2% of the teams consist of at least one woman. This proportion is close to the actual proportion observed in the economics and finance fields. Indeed, Hengel (2022) finds that the percentage of papers published in top-4 economics journals with at least one female author was approximately 25% in 2015, with an overall positive trend over the past 25 years. Schwert (2021) indicates that, over the 2010–2020 period, the percentage of female coauthors in the *Journal of Financial Economics* was around 15%. Regarding location, our sample appears to be tilted toward Europe. Indeed, 63.1% of the teams involve at least one researcher who is affiliated with a European institution. The corresponding frequencies are 25.0% for North America and 15.5% for the Asia-Pacific region. For the top-3 finance journals, Grossmann and Lee (2022) indicate that the percentage of coauthors located in Europe is between 29.4 and 35.0%, whereas the corresponding percentage range is between 52.8 and 56.5% in North America and between 5.6% and 10.6% in the Asia-Pacific region.

Type of research questions. The six research questions vary in terms of complexity and the econometric techniques required. For instance, *RQ1* focuses on a relatively hard-to-measure and quite abstract concept (i.e., market efficiency) whereas *RQ5* deals with a straightforward measure (i.e., percentage of market orders in all client orders). In our tests, we use research-question specific binary variables in order to estimate fixed effects.

Software. We collect information about the software used by the researchers. Collectively, the researchers used 14 different software languages (see Table 4). The five most common languages are Stata (29.2%), Python (29.2%), R (28.6%), SAS (22.0%), and Matlab (11.9%). The popularity of Stata is consistent with the results of an economics research survey conducted by Vilhuber (2021). However, our sample contains more Python, R, and SAS users than what is typically observed in top finance and economics journals (according to Vilhuber, the market share of Python and R is between 5 and 10% each, and less than 5% for SAS). We believe this is because #*fincap* participants must analyze a particularly large database.

Table 4: Summary statistics on code, research quality, and effort. m and sd denote the mean and the standard deviation, respectively; p_{25}, p_{50}, and p_{75} indicate the 25th, the 50th, and the 75th percentile; min and max denote the minimum and maximum values. Variables marked with * are dichotomous, i.e., means (m; reported as percentages) correspond to the fraction of "successes;" variables not marked with * are metrics. Superscript [a] indicates that the data were elicited on the individual level in #*fincap*'s entry survey and is, thus, based on individual researchers' self-reports; team-level aggregates correspond to the maximum value per team. Superscript [b] indicates that the data were recorded on the team level in the course of *cascad*'s evaluation process. $n = 168$ for all variables except for "Actual CPU time" and "Size of software" for which the numbers of observations amount to $n = 151$ and $n = 167$, respectively.

	m	sd	min	p_{25}	p_{50}	p_{75}	max
Code complexity:							
» Lack of master file [a,*]	33.3%						
» Help of *cascad* verificator [a,*]	17.9%						
» Number of software [a]	1.32	0.62	0.00	1.00	1.00	2.00	5.00
» Size of software (in 100kb) [a]	1.05	4.98	0.03	0.16	0.26	0.42	58.40
» Number of script files [a]	6.25	9.38	1.00	1.00	3.00	7.00	85.00
» CPU time (in hours) [a]	9.66	17.73	0.17	1.67	3.50	8.50	145.00
Peer evaluations:							
» Research Question *RQ1* [b]	−0.25	1.89	−5.33	−1.63	−0.15	1.12	4.14
» Research Question *RQ2* [b]	−0.35	1.73	−5.41	−1.61	−0.27	0.98	3.25
» Research Question *RQ3* [b]	0.81	1.59	−4.91	−0.09	1.03	1.89	4.12
» Research Question *RQ4* [b]	−0.33	1.73	−5.41	−1.69	−0.33	0.99	3.24
» Research Question *RQ5* [b]	0.47	1.53	−4.91	−0.40	0.55	1.53	3.81
» Research Question *RQ6* [b]	−0.23	1.86	−5.41	−1.46	−0.02	1.11	3.76
Documentation quality:							
» Software requirements [a,*]	93.5%						
» Computer specification [a,*]	3.6%						
» Instructions to verificators [a,*]	92.3%						
» Mapping output/results [a,*]	82.1%						
» Runtime [a,*]	7.1%						
» Readme file [a,*]	94.6%						
» Size of readme file (in kb) [a]	1.81	2.05	0.00	0.62	1.17	2.23	13.00

Code complexity. We characterize the complexity of the computer code submitted by the various teams using the following six variables. We start by simply counting the number of software languages used by the researchers. Intuitively, we expect that having to switch from one software to another is more cumbersome for someone who is trying to regenerate the results, as well as being a source of error. On average, teams use 1.32 software applications and/or programming languages; the maximum number of different software applications used is 5. When they utilize more than one software, researchers typically use one software for data cleaning and the preparation phase and another one for the econometric analysis. The other measures we analyze are the number of script files (average number is 6.25), the size of the software (average size is 105 KB),

the computing time (average CPU time is 9.66 hours but the median is only 3.50 hours), the lack of a master file (33.3% do not provide a master file), and the fact that the *cascad* reproducibility verificator had to modify the code to make it run (which applied to 17.0% of the results). We notice that, by construction, the six considered variables are positively correlated with complexity.

Research quality. We rely on a team of 34 experienced peer evaluators to assess the academic quality of each sample paper. Collectively, the evaluators are significantly more senior than the participants: 88.2% are tenured faculty members (associate or full professor), 85.3% have published in a top-5 economics or top-3 finance journal, and their average number of citations is 6,663. Each paper was reviewed twice and the assessment was conducted at the research question level. As a result, each paper received six ratings two times, each of which was provided on a 0–10 scale. To account for peer evaluator fixed-effects, we subtract the mean of all the ratings assigned by an evaluator from his or her scores. Thus, for each research team, the peer evaluation rating per hypothesis is based on the average (demeaned) rating of the two peer evaluators. Table 3 shows that the mean ratings for two of the research questions (*RQ3* and *RQ5*) are positive, whereas the other four are negative; this indicates that the analyses and the results for the two research questions that are the least abstract (*RQ3* and *RQ5*) are on average more positively assessed by peers compared to other research questions.

We also consider a dichotomous variable, called *outlier result*, that indicates whether or not a team's result is an outlier relative to the distribution of all point estimates. We implement a 2.5% threshold on both tails of the distribution as the definition for outlier results, consistent with the threshold used to winsorize the data in #*fincap* (Menkveld et al., 2021). By construction, this variable is directly related to the contribution of a team to the dispersion of the results across teams, also known as non-standard error (Menkveld et al., 2023) or level noise (Kahneman et al., 2021).

Effort. We measure the completeness and quality of the documentation provided by the authors. Specifically, we check whether there is a readme file present (94.6% of the teams provided one). We also measure the size of the readme files (converted to .txt-format) and report large

cross-sectional variation across teams, ranging from minimal to very detailed, 13KB, read-me files ($m = 1.81$KB). In addition, we search for five types of information in the readme files that the current Data Editors of several economics journals have identified as being useful for ensuring that an economic paper is reproducible.[9] The useful types of information consist of information about software requirements (mentioned by 93.5% of the teams), runtime (7.1%), computer specifications (3.6%), instructions to verificators (92.3%), and information that allows one to link the output from the code with the figures and tables in the paper (82.1%).

3. Analysis and Results

3.1. Reproducibility, Bugs, and Problems

Panel (a) of Figure 1 shows the distribution of the 1,008 reproducibility scores, computed at the research question level. A full-reproducibility score of 100 is observed for 52.0% of the results. The frequencies of scores of 75, 50, and 25 are 11.3%, 2.5%, and 5.2%, respectively. Finally, a reproducibility score of zero was assigned to 29.1% of the sample. The average reproducibility score is 63.0 ($sd = 44.3$; $n = 1,008$) and the median score is 100.0. When excluding scores of 0, the average reproducibility score is 88.8.

In the next step, we study the reproducibility scores at the research team and research question levels. Panel (b) of Figure 1 illustrates the average reproducibility score per research team (i.e., the average of the reproducibility scores for the six research questions per team); the mean score across teams is 63.0 ($sd = 40.6$; $n = 168$), and the median score is 83.3. Panel (c) of Figure 1 displays the distribution of the reproducibility scores for the six research questions. For each question, a large proportion—47.6 to 58.3%—of the teams' results is perfectly reproducible. Partial-reproducibility scores (25, 50, or 75) account for 14.3 to 21.4% of the cases, while scores of 0 account for 26.8 to 31.0% of the cases. The overall U-shaped pattern for the reproducibility

[9] The five data editors are Lars Vilhuber (*American Economic Association*), Miklós Koren (*Review of Economic Studies*), Joan Llull (*Royal Economic Society*), Peter Morrow (*Canadian Journal of Economics*), and Marie Connolly (*Canadian Journal of Economics*). Their common continuously-updated template readme is available at https://social-science-data-editors.github.io. In this paper, we use a version retrieved on October 1, 2021.

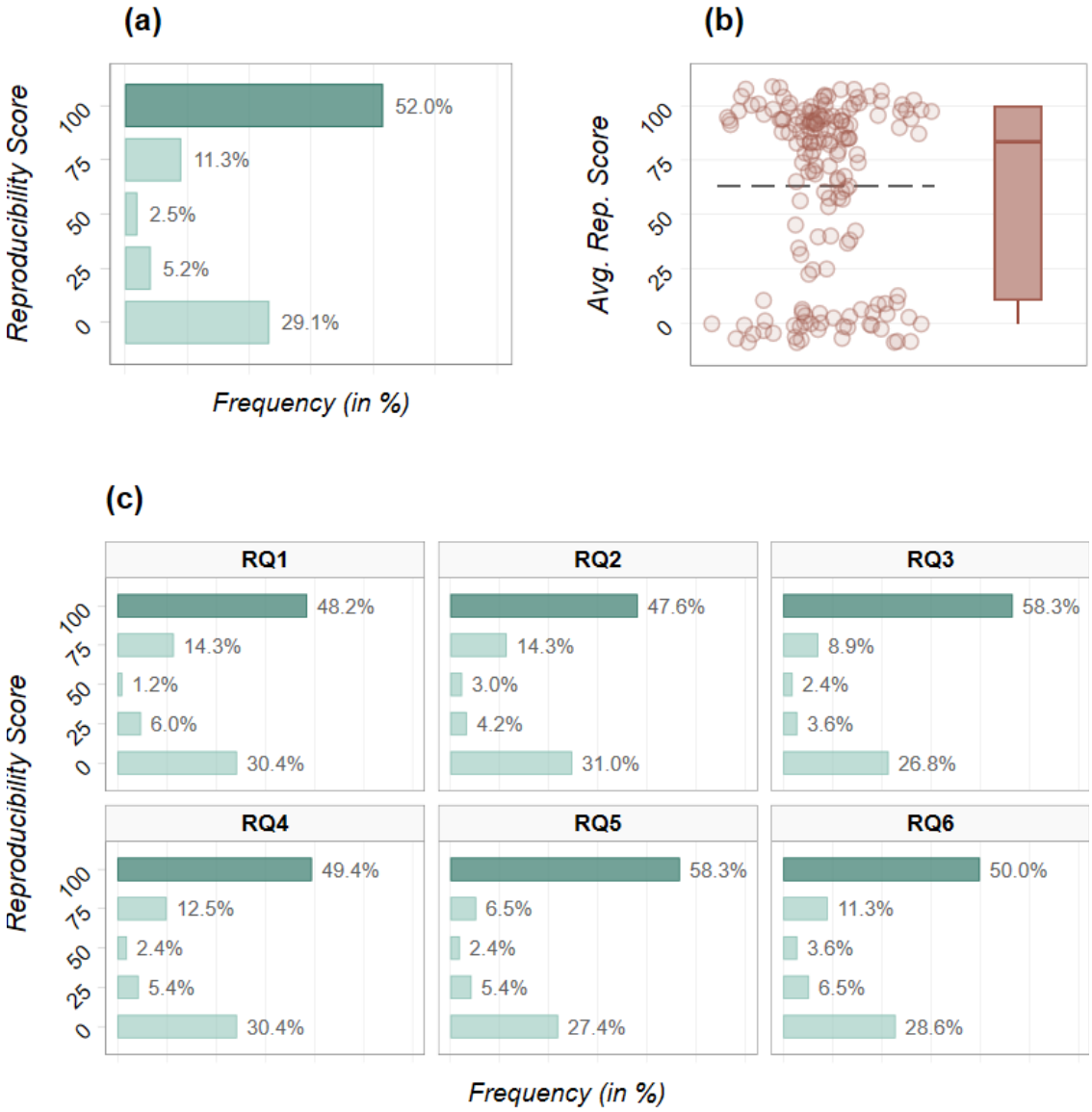

Figure 1: Reproducibility scores. **(a)** Distribution of reproducibility scores at the research question level ($n = 1,008$). **(b)** Strip plot (jittered) of average reproducibility scores at the research team level (i.e., the average of the reproducibility scores for the six research questions per team). The dashed line corresponds to the mean ($m = 63.0$, $sd = 40.6$). The box plot indicates the median ($p_{50} = 83.3$), the interquartile range, and the 5th and 95th percentiles; $n = 168$. **(c)** Distribution of reproducibility scores separated by the six research questions; $n = 168$ in all subpanels.

scores is stable across the research questions. Furthermore, within a given team, we observe some variation across the research questions. The pairwise associations of reproducibility indicators between research questions (mean square contingency coefficients ϕ) vary between 0.545 [H3 vs. H4] and 0.780 [H3 vs. H5], with $p < 0.001$ for all comparisons. Similarly, the pairwise associations of (ordinal) reproducibility scores between research questions (Spearman correlation coefficients ρ_S) vary between 0.712 [H1 vs. H5] and 0.881 [H3 vs. H5], with $p < 0.001$ for all

comparisons.

While the success rate in our sample tends to be higher than that reported by previous empirical studies (Chang and Li, 2017; Gertler et al., 2018; Herbert et al., 2021), one may still wonder why the average reproducibility success rate is far from 100%. This is particularly surprising given that we have access to all data and all computer code, rely on the expertise of an experienced reproducibility verificator, use massive computing resources, allocate an extended amount of time, and have access to any commercial software needed. While this legitimate concern is at the heart of the empirical study presented below, we provide some first elements by studying the problems faced when attempting to regenerate the results. In Panel (a) of Figure 2, we show that 46.1% of the reproduction attempts were initially unsuccessful (i.e., code does not run and reproducibility score (RS) = 0) and that this figure dropped to 29.1% after the intervention of the *cascad* verificator. We find that the most common causes of nonreproducibility are, in decreasing order, bugs and problems affecting the code/scripts, software, readme, CPU/memory, and data. Interestingly, the *cascad* verificator was able to solve approximately 50% of the problems affecting scripts and software. As a complement, we provide in Table C1 in Appendix C, a more granular typology of the problems faced during the verification process, both those that were fixed and those that were not. We also provide actual examples for each of the 20 types of problems.

Since our results are based on an experiment, we test their external validity using 48 real replication kits associated with 32 papers published in leading economics journals (*American Economic Review*, *American Economic Journal*, *Economic Journal*). As these journals are managed by either the *American Economic Association* or by the *Royal Economic Society*, we refer to this extra dataset as the AEA/RES sample. These replication kits were sent to *cascad* by the respective data editors of these journals to request third-party verifications of the computational reproducibility of the results (Vilhuber, 2021). As the verifications for #*fincap* and for AEA/RES were conducted by the same verificator, with the same computing infrastructure, during approximately the same period, and without contacting the authors, we believe this comparison is meaningful. In the AEA/RES sample, 16 papers required only one verification, whereas 16 of the papers had to be verified a second time. The observation that half of the papers have to be verified more than once is consistent with the figures disclosed by the AEA data editor for 2020 and 2021 in its annual report

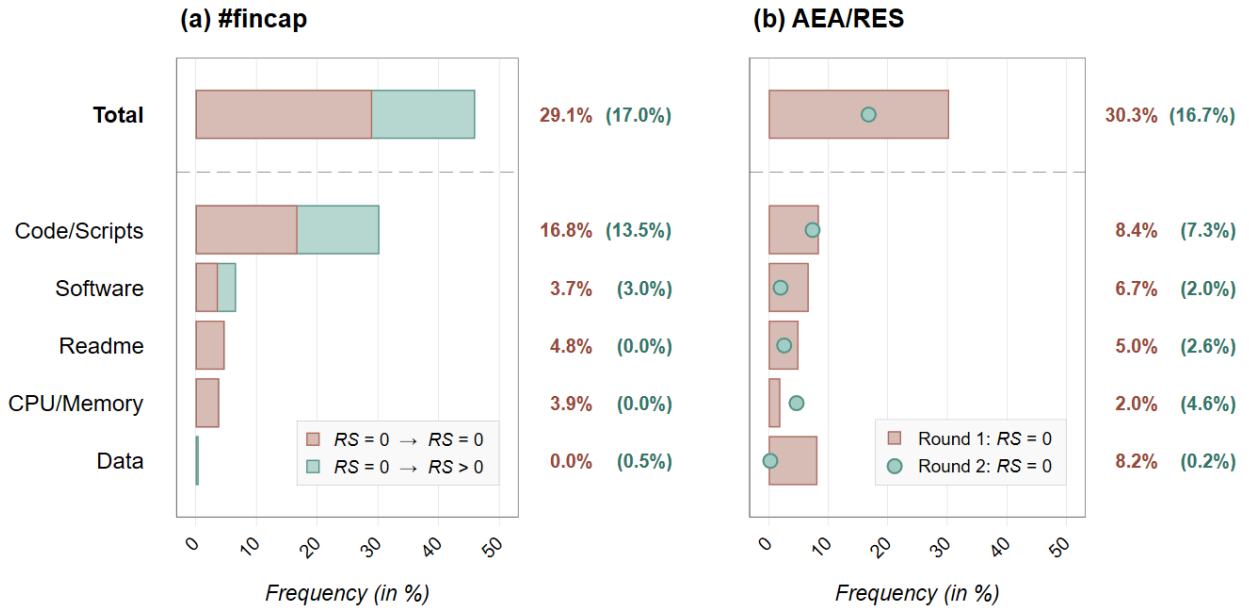

Figure 2: Causes of non-reproducible results. (a) The figure shows the percentage of reproduction attempts for the 1,008 estimates in #*fincap* that did not generate any results before and after possible interventions by the *cascad* verificator. The greenish portion of the bar ($RS = 0 \rightarrow RS > 0$) indicates the fraction of results with a reproducibility score of zero that could be fixed by the verificator to result in a score larger than 0. **(b)** The figure shows the percentage of reproduction attempts for 818 items (tables and figures) in a sample of 32 papers reviewed by *cascad* for the AEA/RES. The bars indicate the fraction of non-generatable results for the initial submissions of replication kits (after potential intervention of the *cascad* verificator; round 1); the green dots indicate the fraction of non-generatable results after the replication kits have been revised by the original authors (round 2).

(Vilhuber, 2022).

Panel (b) in Figure 2 displays the percentage of the 818 results (tables and figures) displayed in these papers that were not reproduced in the first round (R1) or in the second round (R2). We find that 30.3% of the results were not reproduced, which is very similar to the #*fincap* results. As for the breakdown of problems, we observe some reassuring similarities between the #*fincap* project and the real world. However, 8.4% of the results cannot be regenerated in the AEA/RES sample because of data-related issues, which is of course much higher than in #*fincap* as the data were provided to the participants.

3.2. Cross-Sectional Determinants of Reproducibility

The strong cross-sectional variability illustrated in the previous section, both across the research questions and across the teams, should be useful in regard to identifying some of the forces that

drive the level of reproducibility. To identify these factors, we define and test formal hypotheses using several regression models. Then, we investigate whether the researchers in our sample have a good sense of the level of reproducibility of their own research, as well as the difficulty of regenerating their results.

To discipline ourselves, we prespecified the hypotheses to be examined and constructed a detailed analysis plan before examining the data.[10] When we decided to examine the computational reproducibility in #*fincap* in detail, the data collection had already been completed. A formal preregistration was therefore no longer possible, but the prespecified plan is publicly available at https://osf.io/bn7wx/. Analyses not included in our analysis plan are transparently labeled as exploratory analyses below.

Hypotheses. We formally test several hypotheses with the aim of identifying the drivers of computational reproducibility of results in finance. We divide the variables previously presented in Section 2 into two groups. First, the variables that characterize the researchers' characteristics and skills, the type of research questions, and the software used are *pre-determined* variables with respect to the level of computational reproducibility. Indeed, they are provided at the beginning of the experiment, i.e., prior to writing the computer code, the readme file, and the paper. Second, variables describing the complexity of the computer code, the quality of the research as assessed by peer evaluators, and the effort exhibited by researchers to ensure their results are reproducible are *co-determined* variables with respect to the level of computational reproducibility. This partition of the variables is the basis of our two main null hypotheses: *H1. None of the pre-determined variables drive computational reproducibility*; *H2. None of the co-determined variables covary with computational reproducibility, neither with nor without using the pre-determined variables as controls.* While both hypotheses are interesting *per se*, the former allows us to directly and cleanly test for the drivers of reproducibility.

We also analyze whether research teams are able to assess the reproducibility rate of their own

[10] The three team members responsible for the reproducibility assessment had access to all data generated by *cascad* during the reproduction process but they were blinded to the data elicited in #*fincap*. Only three members of the #*fincap*'s coordinator team had access to the data generated in #*fincap*. However, they were blinded to all data generated by *cascad*, except for the reproducibility scores, which enter the analyses as a covariate in Menkveld et al. (2023). The two datasets were merged after the analysis plan was agreed on.

research and whether they are aware of the level of difficulty faced by their peers when attempting to reproduce their findings. To do so, we define the following hypotheses: *H3a. Researchers do not exhibit over- or underconfidence when estimating the computational reproducibility of their results*; *H3b. Researchers do not exhibit over- or underconfidence when estimating the difficulty of reproducing their results.*

Empirical Strategy. Our primary hypothesis tests for *H1* and *H2* are based on Wald tests for joint statistical significance of all explanatory variables included in the respective regression models. If the overall joint test turns out to be significant, we proceed with conducting Wald tests for the joint significance of the coefficient estimates associated with one of the pre-determined or co-determined variables in the particular regression models. In addition, we report the coefficient estimates and the associated p-values of the individual covariates in the regression models. However, we interpret the estimates of the individual predictors and covariates in terms of statistical significance only if the corresponding joint test yields a statistically significant result ($p < 0.05$). Irrespective of the p-value of the joint test, we report the p-values of the individual coefficients.

To test hypotheses *H1* and *H2*, we estimate logistic regressions of the reproducibility indicator on the set of pre-determined variables and co-determined variables, respectively. Standard-errors are clustered at the team level.[11]

As described in detail in Section 2.3, two of the pre-determined variables (i.e., academic quality and coding/data handling skills) and two of the co-determined variables (i.e., code complexity and documentation quality) involve multiple dimensions that are captured by varying sets of factors. In the first step, for each of the four variables, we conduct principal component analyses of the associated factors and use the first principal components as proxies for the relevant characteristics. In a second step, we replace the proxies with the individual input factors. Similar to the constraint that individual variables are only interpreted in terms of statistical significance if the joint Wald

[11] We cluster standard errors on the research team level as there are six observations per team (i.e., one per hypothesis) and it appears sensible to expect that the reproducibility outcomes are correlated within clusters. Note that we do not account for fixed intercepts per research team as characteristics that are constant across hypotheses would get absorbed by the fixed effects.

test results in $p < 0.05$, the coefficient estimates of the factors are only interpreted in terms of statistical significance if the proxy (first principal component) is found to be significant.

For hypothesis *H3a*, we need to test the equality between the probability of underestimating the reproducibility level and the probability of overestimating it. Similarly, for hypothesis *H3b*, we contrast the distribution of the expected level of difficulty (by the researchers) and the distribution of the actual level of difficulty (as measured by *cascad*). All statistical tests are two-tailed. We refer to results with p-values smaller than 0.05 as "statistically significant."[12]

Below, we empirically investigate the drivers of computational reproducibility using the indicator variable for full reproducibility as the outcome variable. In robustness analyses (presented in Appendix D), we replace the dichotomous dependent variable with *cascad*'s ordinal reproducibility scores.[13] In exploratory analyses, we replace the team-level data with individual-level data and restrict the sample to single-authored reports; the results are presented in Appendix E.

Effects of Pre-Determined Variables. Hypothesis *H1* addresses whether the (i) researchers' academic quality, (ii) researchers' coding and data-handling skills, (iii) number of coauthors, (iv) teams' gender composition, (v) teams' location, (vi) software used, and (vii) research questions examined by the teams in #*fincap* systematically affect the computational reproducibility. The model estimates in terms of marginal effects are reported in Table 5; Wald tests for joint significance of groups of independent variables are tabulated in the bottom panel. Robustness analyses using the ordinal reproducibility scores instead of the binary indicator for full reproducibility

[12] We agreed on adopting the conventional cutoff value of 5%, instead of the stricter 0.5% significance threshold proposed by Benjamin et al. (2018), with the constraint that the p-values of individual regressors only be interpreted in terms of statistical significance if the corresponding joint Wald test results in $p < 0.05$. Given the restrictive nature of the latter requirement, we abstain from correcting our results for family-wise error rates. Note that our final decision regarding the significance threshold and using Wald tests as primary hypothesis tests was made *before* we conducted any data analyses (see our analysis plan at https://osf.io/bn7wx/).

[13] We planned to use ordered logistic regressions of the (ordinal) reproducibility scores on the set of pre- and co-determined variables. However, when estimating the models, we realized that the assumptions of the ordinal logit model are violated for our data. In particular, Brant tests indicated that the parallel line assumption must be rejected for all six regression models. Multinomial logit models would qualify as an alternative to ordinal logit models. However, given that the number of observations for partial reproducibility scores (25, 50, and 75) are small (see Figure 1), multinomial models were not a sensible choice either. We therefore opted to estimate linear models instead, knowing well that the ordinary least squares assumptions are not fulfilled either, resulting in inefficient (but unbiased) estimates.

as the dependent variable show qualitatively robust results; see Table D2 in Appendix D for details.

The joint Wald test for all variables in model (1) is statistically significant ($\chi^2(17) = 42.383$; $p = 0.001$), which indicates that the set of pre-determined variables explains a significant part of the variation in teams' reproducibility scores. Regarding the individual predictors included in the model, we do not find evidence for a systematic effect of teams' academic quality on the likelihood of full reproducibility.[14] In model (2), the joint Wald test of the four factors associated with academic quality is insignificant ($\chi^2(4) = 4.782, p = 0.310$), as are the individual coefficient estimates for the four factors. We find this result surprising, as one may expect more experienced and more successful researchers to produce results that are more reproducible than those of less experienced members of the profession. Another theoretical reason to expect more senior researchers to produce reproducible research is to limit reputation risk. However, we do not find evidence in support of this view; rather, our results appear to be in line with an opportunity cost story (Colliard et al., 2022; Miguel, 2021).[15]

The first principal component of variables associated with coding and data handling skills, however, significantly increases the likelihood of computational reproducibility. For a one-standard deviation increase in our proxy for coding skills, the probability that teams' results are perfectly reproducible increases, on average, by 9.2 percentage points, hereafter *pp* ($p = 0.001$).[16] The joint Wald test of the coefficient estimates of the four factors associated with coding skills is statistically significant ($\chi^2(4) = 14.598, p = 0.006$). The individual coefficient estimates (see model (2) for details) suggest that the positive impact on reproducibility is governed by whether teams use loops and/or matrix operations in their analyses ($AME = 33.0$ *pp*, $p < 0.001$), which is the factor

[14] The first principal component captures 61.4% of the variability of the four factors associated with academic quality, indicating that a common dimension explains a substantial part of the overall variation. Components 2, 3, and 4 explain 20.5%, 11.6%, and 6.5% of the variance, respectively.

[15] In exploratory analysis, we re-estimate model (2) on the dataset expanded to the individual level as well as on the sample of single-authored papers. Both analyses suggest that seniority negatively affects computational reproducibility, with marginal effects of -16.1 ($p = 0.023$) and -34.0 ($p = 0.046$) percentage points, respectively. See Tables E1 and E2 in Appendix E for details.

[16] The principal component analysis reveals that the first component captures only 28.5% of the overall variation in the four factors. Components 2, 3, and 4 explain 26.8%, 25.2%, and 19.5% of the variance, respectively, which suggests that the four factors describe distinct characteristics that do not stem from a common dimension. As such, we deem the individual estimates reported in model (2) more informative than the estimate of the first principal component.

Table 5: Logit regressions of the reproducibility indicator on pre-determined covariates. Estimates are reported in terms of marginal effects. *PC-1* indicates the first principal component from a principal component analysis of the covariates associated with academic quality and coding skills, respectively. Variables marked with † are dichotomous; see Table B1 in Appendix B for details. The bottom panel reports $\chi^2(df)$-statistics, with df being defined as the number of coefficients (k), for Wald tests for joint statistical significance of groups of covariates. $n = 1,008$, clustered for 168 research teams, in both models. McFadden's Pseudo R^2 is 0.058 and 0.078 for models (1) and (2), respectively. p-values are reported in parentheses; * $p < 0.05$.

	Model (1)		Model (2)	
Academic Quality:				
» PC-1	−0.021	(0.319)		
» Seniority†			−0.152	(0.070)
» Top Publication†			0.090	(0.287)
» Citations (in logs)			−0.004	(0.853)
» Expertise (0–10)			0.009	(0.691)
Coding Skills:				
» PC-1	0.092*	(0.001)		
» Parallel Comp.†			0.107	(0.319)
» Loops/Matrix Operations†			0.330*	(0.000)
» Large Data†			0.001	(0.982)
» Coding Skills†			−0.015	(0.832)
Coauthor:				
» Team of Two†	0.052	(0.524)	0.038	(0.634)
Gender:				
» Female†	−0.059	(0.408)	−0.046	(0.523)
Location:				
» North America†	−0.010	(0.901)	−0.024	(0.765)
» Asia-Pacific†	−0.138	(0.168)	−0.161	(0.104)
» Other countries†	0.036	(0.755)	0.085	(0.478)
Software:				
» Matlab†	−0.080	(0.473)	−0.069	(0.507)
» Python†	0.014	(0.884)	0.014	(0.881)
» R†	−0.052	(0.552)	−0.063	(0.464)
» SAS†	0.130	(0.131)	0.147	(0.095)
» Stata†	−0.116	(0.132)	−0.110	(0.146)
Research Questions:				
» RQ2†	−0.006	(0.862)	−0.006	(0.862)
» RQ3†	0.101*	(0.003)	0.101*	(0.003)
» RQ4†	0.012	(0.732)	0.012	(0.732)
» RQ5†	0.101*	(0.006)	0.101*	(0.005)
» RQ6†	0.018	(0.613)	0.018	(0.613)
Wald Tests:				
» Overall model	42.383*	(0.001)	56.243*	(0.000)
» Academic quality ($k = 4$)			4.782	(0.310)
» Coding skills ($k = 4$)			14.598*	(0.006)
» Location ($k = 3$)	2.205	(0.531)	3.479	(0.324)
» Software ($k = 5$)	7.828	(0.166)	8.003	(0.156)
» Research questions ($k = 5$)	15.037*	(0.010)	15.021*	(0.010)

that loads the strongest on the first principal component.

We conjecture that the presence of a coauthor can increase the incentives to clean-up and better document the code and could thus act as a monitoring device. However, we report no significant difference in the likelihood of reproducibility for teams comprising one researcher and teams comprising two researchers ($AME = 5.2\,pp$, $p = 0.524$).[17] Likewise, we neither find evidence that the likelihood of reproducibility significantly differs between teams involving a female team member and teams only consisting of males, nor find evidence that reproducibility systematically varies across researchers' location/region.[18] With respect to the software and/or programming languages used by teams, the coefficient estimates and the joint tests are insignificant for both models (1) and (2).[19]

Finally, we report a significant Wald test result for the research question fixed effects ($\chi^2(5) = 15.037$, $p = 0.010$). More particularly, we find significant effects for *RQ3* and *RQ5*; on average, teams' estimates for these two research questions are about 10 *pp* more likely to be fully reproducible than estimates for *RQ1*.[20] Note that the six research questions tested by teams exhibit—on purpose—considerable variation in terms of the level of abstraction. While *RQ1* is based on the relatively abstract notion of market efficiency and calls for advanced econometric methods, *RQ3* and *RQ5* only require the computation of simple ratios for the share of client volume in total volume and the share of market orders in all client orders, respectively. Thus, the positive coefficient

[17] In an exploratory analysis, we examine whether the team composition, i.e., whether the team is composed of two professors, one professor and one early-career researcher (ECR), or two ECRs correlates with reproducibility. We do not find evidence that team composition affects reproducibility. Refer to Table E3 in Appendix E for details.

[18] Not only the coefficient estimates of the three location indicators (relative to the baseline category Europe) are statistically insignificant, but also all the comparisons between them. In particular, Wald tests for model (1) yield the following results: (i) North America vs. Asia-Pacific: $\chi^2(1) = 1.296$, $p = 0.255$; (ii) North America vs. Other Countries: $\chi^2(1) = 0.129$, $p = 0.720$; and (iii) Asia-Pacific vs. Other Countries: $\chi^2(1) = 1.528$, $p = 0.216$. The pairwise comparisons for model (2) yield a similar picture: (i) North America vs. Asia-Pacific: $\chi^2(1) = 1.505$, $p = 0.220$; (ii) North America vs. Other Countries: $\chi^2(1) = 0.669$, $p = 0.414$; and (iii) Asia-Pacific vs. Other Countries: $\chi^2(1) = 2.835$, $p = 0.092$.

[19] Exploratory pairwise comparisons between the coefficient estimates for the different software applications and programming languages used suggest that using SAS tends to increase reproducibility rates as compared to using Stata ($\chi^2(1) = 6.545$, $p = 0.011$ in models (1); and $\chi^2(1) = 6.557$, $p = 0.010$ in model (2), respectively) and using R ($\chi^2(1) = 3.639$, $p = 0.056$ in models (1); and $\chi^2(1) = 4.654$, $p = 0.031$ in model (2), respectively). All remaining pairwise comparisons of coefficient estimates for the software dummies are statistically insignificant.

[20] The coefficient estimates of *RQ3* and *RQ5* are not only significantly higher than the base category (*RQ1*); pairwise Wald tests between the six *RQ*-coefficients indicate that, for both models (1) and (2), the likelihood of being fully reproducible is significantly higher for *RQ3* and *RQ5* than for the remaining four research questions ($p < 0.05$ for all comparisons). All pairwise comparisons between coefficient estimates for *RQ1*, *RQ2*, *RQ4*, and *RQ6* are statistically insignificant ($p > 0.05$).

estimates for *RQ3* and *RQ5* are in line with our expectations.

Effects of Co-Determined Variables. Hypothesis *H2* addresses whether research quality, code complexity, and documentation quality systematically co-vary with computational reproducibility.[21] The model estimates in terms of marginal effects are reported in Table 6; the results of the Wald tests for joint significance of groups of covariates are tabulated in the bottom panel.

Table 6: Logit regressions of the reproducibility indicator on co-determined covariates. Estimates are reported in terms of marginal effects. *PC-1* indicates the first principal component from a principal component analysis of the covariates associated with code complexity and documentation quality, respectively. Variables marked with † are dichotomous; see Table B1 in Appendix B for details. The bottom panel reports $\chi^2(df)$-statistics, with df being defined as the number of coefficients (k), for Wald tests for joint statistical significance of groups of covariates. $n = 906$, clustered for 151 research teams, in both models. McFadden's Pseudo R^2 is 0.030 and 0.045 for models (3) and (4), respectively. p-values are reported in parentheses; * $p < 0.05$.

	Model (3)		Model (4)	
Research Quality:				
» Peer Evaluation (0–10)	0.019	(0.197)	0.017	(0.248)
» Outlier Result†	−0.235*	(0.010)	−0.209*	(0.014)
Code Complexity:				
» PC-1	−0.045*	(0.048)		
» Number of Software			−0.083	(0.209)
» Number of Script Files			0.000	(0.975)
» Size of Software (in kb)			0.000	(0.188)
» Actual CPU Time (in minutes)			0.000	(0.520)
» Lack of Master File†			0.029	(0.686)
» Help from Verificator†			0.065	(0.414)
Documentation Quality:				
» PC-1	0.046*	(0.041)		
» Readme File†			0.090	(0.830)
» Size of Readme File (in kb)			−0.005	(0.743)
» Software Requirements†			−0.047	(0.846)
» Runtime†			0.046	(0.675)
» Computer Specification†			0.114	(0.439)
» Instructions to Verificators†			0.211	(0.478)
» Mapping Output/Results†			0.125	(0.151)
Wald Tests:				
» Overall model	15.216*	(0.004)	25.164*	(0.048)
» Code complexity ($k = 6$)			5.840	(0.441)
» Documentation quality ($k = 7$)			7.155	(0.413)

[21] The number of observations drops from $168 \times 6 = 1,008$ to $151 \times 6 = 906$ when addressing the relationship between reproducibility and the set of co-determined variables since the code of 17 teams could not be executed at all, which resulted in the variable "Actual CPU time" being undefined for 17 teams.

The covariates in models (3) and (4) are jointly significant (see the bottom panel in Table 6 for details), which indicates that the set of co-determined variables explains some of the observed heterogeneity in computational reproducibility. Regarding the estimates of the individual covariates, we do not find evidence that the quality of the short papers, as proxied by the peer-reviewed ratings, significantly correlates with the variation in reproducibility. We deem this null result remarkable as it seems reasonable to conjecture that results that are considered by peers to be of high quality should be more likely to be reproducible.

In contrast, the finding on the effect of the outlier results variable is striking.[22] Indeed, we find strong evidence that the results lying in the tails of the distribution are substantially less likely to be fully reproducible, with a marginal effect of approximately 20 percentage points. By construction, the *outlier result* variable is directly related to the contribution of a team to the dispersion of the results across teams, also referred to as non-standard error (Menkveld et al., 2023) or level noise (Kahneman et al., 2021). One potential interpretation is to assume that deviations from the center of the distribution serve as a proxy for the quality of a team's estimate, presuming that the "concensus" estimate is informative as to the ground truth.[23] With this interpretation, the negative coefficient associated with *outlier result* suggests a positive relationship between computational reproducibility and research quality. To the best of our knowledge, this is the first time that such a positive relationship has been empirically established. In most scientific contexts, the distribution of the results across researchers for a given research question is not observable. This distribution is only available in the following two situations: (1) in meta-analyses and (2) in multi-analyst studies such as *#fincap*. As a result, we are usually unable to determine whether a reported estimate qualifies as an outlier result relative to the latent distribution of estimates. However, we can always assess the computational reproducibility of any result, as long as a replication kit is available; and this can be used as a proxy for research quality. We see this

[22] Note that the inclusion of the variable *outlier result* was not part of our initial analysis plan, but was added to our analysis as per the suggestion of a reviewer and the associate editor. We use the 2.5^{th} and 97.5^{th} percentile of the distribution of point estimates to characterize outlier results, consistent with the winsorization level used in Menkveld et al. (2021); robustness tests using different thresholds are presented in Appendix F and indicate that the correlation is not governed by the threshold value.

[23] Alternatively, ending up in the center of the distribution *could* just as well capture reliance on commonly accepted methodologies in the field (which may or may not be due to skills). For instance, results in the tails might rely on more "exotic" methodologies, calling for less well-established computational methods, which in turn could imply lower reproducibility rates.

as an additional reason to pay attention to the computational reproducibility of research.

We also report that the first principal component of factors associated with code complexity tends to be significantly negatively related to reproducibility. For a one standard deviation increase in the standardized proxy of code complexity, the likelihood of results being fully reproducible decreases by 4.5 *pp* ($p = 0.048$). However, the joint Wald test for the six factors associated with code complexity in model (4) is statistically insignificant ($\chi^2(6) = 5.840$, $p = 0.441$), as are the individual coefficient estimates. At a comparable order of magnitude, we find that the likelihood of full reproducibility tends to increase with documentation quality; a one standard deviation increase in the proxy of documentation quality, on average, is associated with a 4.6 *pp* increase ($p = 0.041$) in the probability of successful reproduction. However again, the joint Wald test of the seven coefficients associated with documentation quality in model (4) is statistically insignificant ($\chi^2(7) = 7.155$, $p = 0.413$).[24]

3.3. Expected Reproducibility and Difficulty

After the teams completed all four stages of *#fincap*, they were asked to answer an exit survey that comprised two questions regarding their expectations about their own code's reproducibility.[25] The first question elicited teams' expectations about the reproducibility of their analyses, as follows: *Do you think it would be possible to reproduce your results from the raw data and your computer code?* The question was answered on an ordinal scale with the following options: *A. one would find exactly the same results*, *B. only minor differences may arise*, *C. major differences may arise*, and *D. it would be impossible to reproduce the results.* Since the scaling of the question differs from the reproducibility scores generated by *cascad*, we focus our attention on full reproducibility and define a dichotomous variable that takes a value of one if the team selects the highest expected reproducibility score and zero otherwise. Comparing teams' expectations

[24] The discrepancy in conclusions between models (3) and (4) in Table 6 is likely due to the fact that *PC-1* captures a relatively small share of the cross-sectional variability (28.4% for code complexity and 40.4% for documentation quality). Auxiliary analyses (see Tables D2 through D4 in Appendix D) controlling for the heterogeneity in the set of pre-determined variables indicate that the results reported above are qualitatively robust.

[25] Note that two teams completed all stages in *#fincap* but failed to complete the exit survey. Thus, the number of observations in all tests reported in this section is $n = 166$ rather than $n = 168$.

about whether or not their results will be fully reproducible to the assessments of actual reproducibility by the *cascad* verificator allows us to address hypothesis *H3a*, i.e., whether researchers systematically over- or underestimate the reproducibility of their results.

The second question in #*fincap*'s exit survey elicited teams' beliefs about how difficult it would be to reproduce their results. In particular, teams answered the following question: *How easy would it be to reproduce your results?* The question was answered on an ordinal scale with the following four possible responses: *A. straightforward*, *B. quite easy*, *C. challenging*, and *D. very difficult*. Comparing teams' expectations about the difficulty of reproducing their estimates to the assessments made by *cascad*'s verificator—who assigned a rating for the actual difficulty of reproducing the results on the same scale—allows us to address hypothesis *H3b*, i.e., whether researchers systematically over- or underestimate the difficulty of reproducing their results.

At the paper level, only 28.3% of the results are fully reproducible; for the majority of the results (71.7%), at least one minor discrepancy emerged between the results reported in the paper and the results obtained from the reproduction exercise. However, 70.5% of the teams in the sample expect that one would find exactly the same results as those reported in their short paper when attempting to reconduct their analyses, whereas only 29.5% indicate that at least minor differences might arise.[26] Panel (a) in Figure 3 illustrates the proportion of teams who expect their results (not) to be fully reproducible, separated by whether the results are actually reproducible. Teams whose results are reproducible turn out to have relatively well-calibrated expectations: 74.5% (21.1% ÷ 28.3%) correctly anticipate that their results will be reproducible, while 25.5% (7.2% ÷ 28.3%) are even underconfident in their expectations (i.e., they expect their results not to be perfectly reproducible even though the reproduction exercise generated exactly the same results). However, zeroing in on the expectations of teams whose results could not be fully reproduced reveals that a substantial share of the sample is overconfident. Indeed, only 31.1% (22.3% ÷ 71.7%) of the teams correctly anticipate that the results reported in their paper will not be fully reproducible, whereas 68.9% (49.4% ÷ 71.7%) erroneously suppose that their estimates could be perfectly reproduced. The McNemar's test indicates that the teams' expectations significantly

[26] We acknowledge that it may be difficult for some researchers to respond in a survey that they believe their results cannot be regenerated. As a result, they may pick response A even if they know this may not be true. This behavior biases upward our overconfidence estimate.

exceed the actual reproducibility assessments ($\chi^2(1) = 52.128$, $p < 0.001$; $n = 166$) with a large standardized effect size (Cohen's d) of $d = 1.060$ (95% CI $[0.723, 1.446]$).

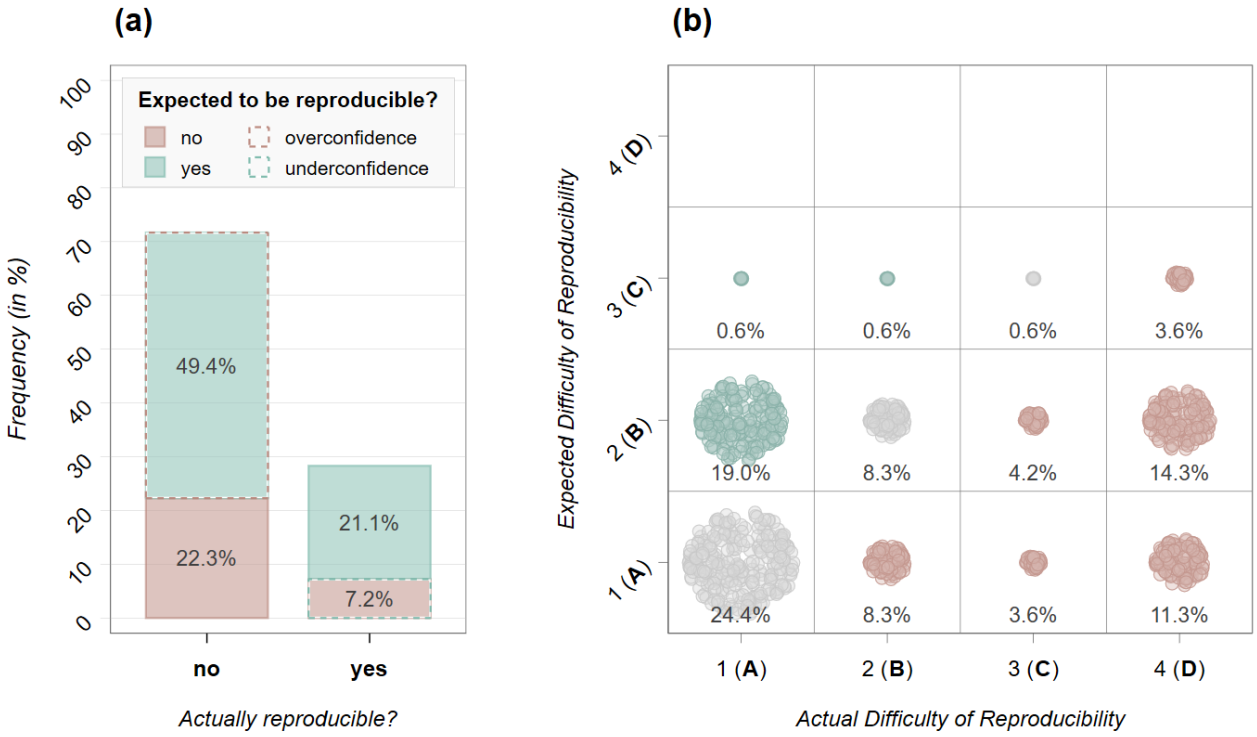

Figure 3: Actual and expected (difficulty of) reproducibility. **(a)** Actual reproducibility separated by research teams' expectation about reproducibility. Both actual and expected reproducibility refers to full reproducibility of all six results reported in the short papers. $n = 166$. **(b)** Scatter plot of expected difficulty of reproducibility over actual difficulty to reproduce the research teams' results. Observations are jittered with weights being determined by the fraction per cell. Observations on the diagonal correspond to accurate expectations; observations below (above) the diagonal indicate overconfidence (underconfidence). $n = 166$.

We report a similar pattern with respect to teams' expectations about the difficulty in reproducing their analyses. Regarding teams' expectations, 94.6% indicate that it would be "straightforward" (*A*) or "quite easy" (*B*) to reproduce their results. Only 5.4% expect the reproduction of their estimates to be "challenging" (*C*); not a single team anticipates that it would be "very difficult" (*D*) to reproduce their findings. The distribution of *cascad*'s assessments of the actual difficulty gives a very different picture; while 62.0% of the analyses are rated *A* or *B*, reproducing the teams' estimates actually turned out to be "challenging" (*C*) or "very difficult" (*D*) in 38% of the cases. Panel (b) in Figure 3 shows a scatter plot of teams' expected difficulty regarding reproduction compared to *cascad*'s assessment of the actual difficulty. While 33.7% of the teams have well-calibrated expectations (on the diagonal) and 20.5% are underconfident (above the diagonal), a

substantial share of the sample, i.e., 45.8%, is overconfident in their expectations (below the diagonal). A Wilcoxon signed-rank test indicates that the teams' anticipations are statistically significantly more positive than *cascad*'s actual difficulty assessments ($z = 5.212$, $p < 0.001$; $n = 166$), with a large standardized effect size of $d = 0.885$ (95% CI $[0.558, 1.049]$).

We believe that the results in this final subsection are particularly important. They provide a potential explanation for why the reproducibility rate is not higher in economics and finance research. Many researchers seem to be unaware of the fact that *(i)* their research may not be fully reproducible and that *(ii)* attempting to regenerate empirical results is much more challenging than they seem to anticipate. As a consequence, these researchers might simply not exert enough effort such that the equilibrium reproducibility level tends to remain low.

4. Implications for Researchers and Journals

Researchers. What are researchers supposed to do to increase the reproducibility of their own research? To answer this question in a very clear way, we provide in Exhibit 1 some guidelines related to the readme file, the code, and the data. Our guidelines are informed by the empirical findings in this paper, in particular the analysis of the problems and bugs in Section 3.1 and of the cross-team variation in reproducibility across teams in Section 3.2. For instance, providing information on the items listed in Exhibit 1 would solve most of the issues listed in Table C1. Items R1-R6 (readme) collectively reflect the positive relationship between reproducibility and documentation quality in Table 6. Furthermore, item D2 addresses the most common bug encountered in *#fincap*, namely missing and misspelled variables. The guidelines are also fueled by our own experience as reproducibility verificators at *cascad*, discussions with data editors of economics journals, and by data and code availability standards endorsed by leading journals in the social sciences (https://datacodestandard.org).

We believe that complying with these guidelines would both boost computational reproducibility and save time for reproducibility verificators and for any researcher willing to verify or build upon existing research in finance. It may also reduce the chance that journals' data editors and

other interested researchers will have to contact the authors with questions about and problems regarding their replication kits, which prolongs the time from an article being accepted to being published. Answering such requests can be particularly challenging if they arise months or years after the paper is published, when all the contributors, including research assistants, may no longer be available, and when some of the details of the analysis may have been forgotten. Moreover, it may significantly reduce the likelihood of someone subsequently raising a lack of reproducibility problem.

As general guidance, we recommend that researchers attempt to reproduce the results reported in their paper before submitting their replication kit to an academic journal. Ideally, reproduction should be undertaken by an independent researcher (e.g., a coauthor or research assistant not involved in the data analysis) and in a "fresh" computing environment. Simple on-site reproduction attempts would likely reveal most of the problems that would be encountered by data editors and their reproducibility teams and qualify as a straightforward means to advance computational reproducibility.

Academic Finance Journals. Our findings can also have valuable implications for academic journals and their editorial teams. Currently, an increasing number of academic finance journals are requesting access to the code and, when possible, to the data associated with published papers (Whited, 2021). However, we are not aware of any finance journals both systematically checking the submitted material for completeness and verifying the computational reproducibility of the reported results. The current journal policies do not prevent publication of papers with incomplete material or information, bugs, and – when the code runs – discrepancies between the regenerated and original results.

One solution, which is being implemented by a growing number of economics journals, is to conduct a systematic reproducibility check for all conditionally-accepted articles. The check consists of verifying the submitted material, requesting any additional information and material, running the code, comparing the results, and potentially requesting revisions of the replication kits by the authors until there is no discrepancy in results remaining. Such verification can be performed either by (i) by the journal (e.g., *Review of Economic Studies*), (ii) a scientific association for all

Exhibit 1: Guidelines to improve computational reproducibility in finance. The guidelines are structured in three parts (readme file, data, and code) and within each part, we distinguish must-have information from good-to-have information; the latter are indicated with an asterisk (*).

PART 1: README FILE

R1. Data availability: (i) For each sharable dataset, mention whether it is directly included in the replication kit or available elsewhere (repository, website). **(ii)** For each nonsharable dataset (copyright, NDA, restricted access), provide the following relevant information on how to obtain it: data provider, database identifier (name, DOI, vintage), application and registration procedures, monetary costs, time requirements, instructions on which range and variables to pick; indicate whether a third party can temporarily access the data (for reproduction purposes). **(iii)*** For nonsharable data (including widely available but proprietary data, such as CRSP or Compustat), provide a synthetic dataset to demonstrate that the code runs and generates outputs for all figures and tables; indicate whether a script needs to be run to generate the synthetic dataset or where to locate it.

R2. Data preparation: Ensure that minimal manual action is needed before running the code, i.e., automate the data preprocessing. If manual action is needed (cleaning, merging, converting), describe the necessary steps in detail.

R3. Computational requirements: Provide information on **(i)** any required software (and its version), **(ii)** any required packages/libraries (and their versions), **(iii)** any required compiler (and its version), **(iv)** the operating system (and its version/distribution), **(v)** the hardware specifications of the computer(s) used (RAM, processor, number of cores, clusters), **(vi)** the runtime, and **(vii)*** the required space on the drive to store intermediary data/results.

R4. List of scripts and their functions:* Do so for all scripts and not only for those which need to be run (the master files). This information can prove useful when the code does not run smoothly.

R5. Intermediary datasets: Some scripts create intermediary datasets from the raw data. As a default, include them in the replication kit; indicate which data files are intermediary and which part of the code generates them. If one is unable to reproduce intermediary data files from the raw data because of bugs, time constraints, or insufficient CPU, the rest of the verification can still be carried out.

R6. Instructions on how to run the code: Describe all steps that need to be followed to generate the results reported in the paper (whether one needs to change the path, which scripts to run and in which order, whether command line/shell prompts need to be executed, whether one needs to change software settings). If the code does not automatically generate the tables and figures, indicate how to recreate them from results in the output/log file.

PART 2. CODE

C1. Structure:* There should be one or a few master files that require minimal modifications, call the required scripts in order, and automate the replication process.

C2. Cleaning: Clean up the code and ensure that no futile parts (functions, commands) are included.

C3. Commenting: Annotate all code such that it can be easily understood by independent researchers, and by yourself in the future. Structure your code and scripts using comments to enhance readability and intelligibility.

PART 3. DATA

D1. Format.* Ensure that the formats and file types of the raw data files match the ones used/required by the code.

D2. Variables. Check for missing and misspelled variables.

D3. Codebook.* Provide a variable dictionary (codebook) and/or assign self-explanatory variable labels (in data formats that allow for labeling variables and values). Describe all variables in the dataset such that they can be easily understood.

its academic journals (e.g, *American Economic Association*, *Royal Economic Studies*), or (iii) by an external third-party verificator (e.g., *cascad*, the *Odum Institute for Research in Social Science* at the University of North Carolina). As shown by Colliard et al. (2022), models (ii) and (iii) permit the exploitation of some important economies of scale and third-parties can have a comparative advantage in accessing restricted data. This systematic computational verification solution solves the problems but is costly for journals and some authors may find it bothersome. An alternative model would be to only verify a subset of conditionally-accepted papers. The composition of the subset could be random or be based on some features of the research (e.g., having important policy implications or challenging the current consensus). This would be less costly for journals and authors, yet maintain some incentives to prepare high-quality replication kits.

A third solution is to create strong incentives for authors and other researchers to improve the computational reproducibility of published results, without conducting any pre-publication reproducibility test. This can be achieved by encouraging and publishing comments on or reanalyses of published papers (Nagel, 2019) and by red-penciling or retracting nonreproducible papers. By design, this would increase both the likelihood of detecting nonreproducible research and the reputational risk for authors. The third solution dominates the first two in terms of monetary and organizational costs (the lower bound of the verification costs estimated by Colliard et al. (2022) is $\dot{3}34 per paper). However, as it relies on ex-post verifications, it requires flagging or retracting scientific results that are already in the public domain. As a result, the scientific damage is more severe, as is the reputational cost for the authors, the journal, and the scientific discipline. Furthermore, it is well known that self-correction within science can be challenging and inefficient (see, e.g., Jamieson, 2018; Serra-Garcia and Gneezy, 2021). In contrast, pre-publication checks allow any discrepancies to be detected before the article becomes an official peer-reviewed contribution to scholarship. In case of minor discrepancies, the paper can still be fixed and eventually published; if some of the main conclusions do not hold true, the article can be rejected at this stage without creating too many negative externalities for the research community.

5. Conclusion

We have presented a large-scale analysis of computational reproducibility in finance using 168 research papers written in the context of a multi-analyst study in market microstructure. Our average success rate compares favorably with existing evidence in economics. However, this paper highlights a problem with the current policies around code and data availability in finance, i.e., when they are available, replication kits sometimes do not run and, when they do, they do not always produce the exact results reported in the corresponding papers. In this paper, we quantify this phenomenon in the field of market microstructure, try to understand some of its causes, and suggest some remedies.

Ensuring that published research in finance can be reproduced helps to boost trust in finance, but this is not a panacea. Computationally reproducible research can still be plagued by various honest mistakes (e.g. typo in the code) or plain fraud (e.g. data alteration). Being computationally reproducible is a minimum requirement that calls for, and facilitates, additional reanalyses such as the ones discussed in Table 1. As called for in his AFA presidential address by Harvey (2017), finance needs to nurture a culture of reanalysis, and we hope this paper will contribute to it.

References

Acharya, V. V., & Pedersen, L. H. (2005). Asset pricing with liquidity risk. *Journal of Financial Economics*, *77*(2), 375–410.

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., & *et al.* (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10.

Buckheit, J. B., & Donoho, D. L. (1995). WaveLab and reproducible research. In A. Antoniadis & G. Oppenheim (Eds.), *Wavelets and statistics. Lecture notes in statistics*. Springer.

Chang, A. C., & Li, P. (2017). A preanalysis plan to replicate sixty economics research papers that worked half of the time. *American Economic Review*, *107*(5), 60–64.

Chen, A. Y., & Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Finance Review*, *11*(2), 207–264.

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, *56*(3), 920–980.

Cohn, J. B., Liu, Z., & Wardlaw, M. I. (2023). Count (and count-like) data in finance. *Journal of Financial Economics*, *146*(2), 529–551.

Colliard, J.-E., Hurlin, C., & Pérignon, C. (2022). The economics of computational reproducibility. *HEC Paris Research Paper*, *FIN-2019-1345*.

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in empirical economics: The Journal of Money, Credit and Banking project. *American Economic Review*, *76*(4), 587–603.

Drienko, J., Smith, T., & von Reibnitz, A. (2019). A review of the return–illiquidity relationship. *Critical Finance Review*, *8*, 127–171.

Duflo, E., & Hoynes, H. (2018). Report of the search committee to appoint a data editor for the AEA. *AEA Papers and Proceedings*, *108*, 745.

Gertler, P., Galiani, S., & Romero, M. (2018). How to make replication the norm. *Nature*, *554*(7693), 417–419.

Glandon, P. J. (2011). Appendix to the report of the editor: Report on the American Economic Review data availability compliance project. *American Economic Review*, *101*(3), 696–699.

Grossmann, A., & Lee, A. (2022). An analysis of finance journal accessibility: Author inclusivity and journal quality. *Journal of Banking and Finance*, *forthcoming*, 106427.

Harris, L., & Amato, A. (2019). Illiquidity and stock returns: Cross-section and time-series effects: A replication. *Critical Finance Review*, *8*, 173–202.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *Journal of Finance*, *72*(4), 1399–1440.

Harvey, C. R. (2019). Editorial: Replication in financial economics. *Critical Finance Review*, *8*(1–2), 1–9.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the cross-section of expected returns. *Review of Financial Studies*, *29*(1), 5–68.

Hengel, E. (2022). Publishing while female: Are women held to higher standards? Evidence from peer review. *The Economic Journal*, *132*(648), 2951–2991.

Herbert, S., Kingi, H., Stanchi, F., & Vilhuber, L. (2021). The reproducibility of economics research: A case study. *Banque de France Working Paper Series*, *WP #853*.

Holden, C. W., & Nam, J. (2019). Do the LCAPM predictions hold? Replication and extension evidence. *Critical Finance Review*, *8*, 29–71.

Hou, K., Xue, C., & Zhang, L. (2020). Replicating anomalies. *Review of Financial Studies*, *33*(5), 2019–2133.

Jamieson, H. (2018). Crisis or self-correction: Rethinking media narratives about the well-being of science. *Proceedings of the National Academy of Sciences*, *115*(11), 2620–2627.

Jensen, T. I., Kelly, B., & Pedersen, L. (2022). Is there a replication crisis in finance? *Journal of Finance*, *forthcoming*.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Harper Collins.

Kazumori, E., Fang, F., Sharman, R., Takeda, F., & Yu, H. (2019). Asset pricing with liquidity risk: A replication and out-of-sample tests with the recent US and the Japanese market data. *Critical Finance Review*, *8*, 73–110.

Li, H., Novy-Marx, R., & Velikov, M. (2019). Liquidity risk and asset Pricing. *Critical Finance Review*, *8*, 223–255.

Linnainmaa, J. T., & Roberts, M. R. (2018). The history of the cross-section of stock returns. *Review of Financial Studies*, *31*(7), 2606–2649.

McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2006). Lessons from the JMCB archive. *Journal of Money, Credit and Banking*, *38*(4), 1093–1107.

McCullough, B. D., & Vinod, H. D. (2003). Verifying the solution from a nonlinear solver: A case study. *American Economic Review*, *93*(3), 873–892.

McLean, R. D., & Pontiff, J. (2016). Does academic publication destroy stock return predictability? *Journal of Finance*, *71*(1), 5–32.

Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Neusüss, S., Razen, M., Weitzel, U., & *et al.* (2021). Non-standard errors. *Tinbergen Institute Discussion Paper 2021-102/IV*. https://bit.ly/3JcSFJ9

Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Neusüss, S., Razen, M., Weitzel, U., & *et al.* (2023). Non-standard errors. *Journal of Finance*, *forthcoming*.

Miguel, E. (2021). Evidence on research transparency in economics. *Journal of Economic Perspectives*, *35*(4), 193–214.

Mitton, T. (2021). Methodological variation in empirical corporate finance. *Review of Financial Studies*, *35*(2), 527–575.

Nagel, S. (2018). Code-sharing policy: Update. *Journal of Finance (Editorial)*.

Nagel, S. (2019). Replication papers in the JF: An update. *Journal of Finance (Editorial)*.

Nagel, S. (2021). Answers to FAQ about the recent retraction of an article in the JF. *Journal of Finance (Editorial)*.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. The National Academies Press.

Pástor, L., & Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy*, *111*(3), 642–685.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227.

Pontiff, J., & Singla, R. (2019). Liquidity risk? *Critical Finance Review*, *8*, 257–276.

Popper, K. R. (1959). *The logic of scientific discovery*. Routledge.

Schwert, G. W. (2021). The remarkable growth in financial economics, 1974–2020. *Journal of Financial Economics*, *140*, 1008–1046.

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, *7*(21), eabd1705.

Trisovic, A., Lau, M., Pasquier, T., & M., C. (2022). A large-scale study on research code quality and execution. *Scientific Data*, (9).

Vilhuber, L. (2021). Report by the AEA data editor. *AEA Papers and Proceedings*, *111*, 808–817.

Vilhuber, L. (2022). Report by the AEA data editor. *AEA Papers and Proceedings*, *112*, 813–823.

Welch, I. (2019). Reproducing, extending, updating, replicating, reexamining, and reconciling. *Critical Finance Review*, *8*, 301–304.

Whited, T. M. (2021). Editorial. *Journal of Financial Economics*, *141*(1), 1–5.

Online Appendices

Computational Reproducibility in Finance: Evidence from 1,000 Tests

Contents

A. Instructions for Research Teams in #*fincap*

The box below presents a reformatted copy of the instructions that were provided to research teams in #*fincap*. The instruction sheet was forwarded to teams when they were provided access to the Deutsche Börse dataset.

Instruction sheet for research teams

This three-page instruction sheet clarifies what is expected of you as a research team in the #*fincap* project. It first provides some context for the hypotheses you are expected to test, then presents the assignment, and finally lists the hypotheses you are asked to test with *only* the Deutsche Börse data that is made available to you by the #*fincap* team. These data contain trade information on the EuroStoxx 50 futures.

A. Context

Electronic order matching systems (automated exchanges) and electronic order generation systems (algorithms) have changed financial markets over time. Investors used to trade through broker-dealers by paying the dealers' quoted ask prices when buying, and accepting their bid prices when selling. The wedge between dealer bid and ask prices, the bid-ask spread, was a useful measure of trading cost, and often still is.

Now, investors more commonly trade in electronic limit-order markets (as is the case for EuroStoxx 50 futures). They still trade at bid and ask prices. They do so by submitting so-called market orders and marketable limit orders. However, investors now also can quote bid and ask prices themselves by submitting (non-marketable) standing limit orders. Increasingly, investors now also use agency algorithms to automate their trades. Concurrently, exchanges have been continuously upgrading their systems to better serve their clients. Has market quality improved, in particular when taking the viewpoint of non-exchange members: (end-user) clients?

B. Assignment

You are expected to write an academic paper that is *maximum five pages long*. To make that feasible you can skip many parts of a typical academic paper. You only need to do the following for all hypotheses listed below:

1. Propose a statistical measure, briefly motivate it, and present the formula to calculate it.
2. For this measure, estimate the average per-year change in percentage terms, based on the full sample (or at least the longest possible period because some series are not available yet at the beginning of the sample). Test it against the null of no change.
3. Report this estimate along with its standard error in four decimals (e.g., "measure Z declined by 1.251% with a standard error 0.241%")
4. Briefly discuss your result.

For example, an appropriate outcome statement for testing hypothesis X which states that Y has not changed is:

> "We propose measure Z to test hypothesis X because [...]. It is calculated as Z = f(DATA). Implementing it leads to the following result: We reject the null of no

change. We find that Y declined as our measure Z declined by 1.251% on average per year where the standard error of this change is 0.421% and the resulting t-statistic is 2.971. This result shows [...]"

We emphasize that you are asked to report your results in a brief manner. *If the paper is longer than five pages we will not consider the paper and we will have to exclude you as co-authors from the project.*

C. Hypotheses

1. Assuming that informationally-efficient prices follow a random walk, did market efficiency change over time?

   Null hypothesis 1: Market efficiency has not changed over time.

2. Did the (realized) bid-ask spread paid on market orders change over time? The realized spread could be thought of as the gross-profit component of the spread as earned by the limit-order submitter.

   Null hypothesis 2: The realized spread on market orders has not changed over time.

   The remaining hypotheses focus on client trades only (i.e., trades implemented by exchange members on behalf of their clients).

3. Did the share of client volume in total volume change over time?

   Null hypothesis 3: Client share volume as a fraction of total volume has not changed over time.

4. On their market orders and marketable limit orders, did the realized bid-ask spread that clients paid, change over time?

   Null hypothesis 4: Client realized spreads have not changed over time.

5. Realized spread is a standard cost measure for market orders, but to what extent do investors continue to use market and marketable limit orders (as opposed to non-marketable limit orders)?

   Null hypothesis 5: The fraction of client trades executed via market orders and marketable limit orders has not changed over time.

6. A measure that does not rely on the classic limit- or market-order distinction is *gross trading revenue* (GTR). Investor GTR for a particular trading day can be computed by assuming a zero position at the start of the day and evaluating an end-of-day position at an appropriate reference price. Relative investor GTR can then be defined as this GTR divided by the investor's total (euro) volume for that trading day. This relative GTR is, in a sense, a realized spread. It reveals what various groups of market participants pay in aggregate for (or earn on) their trading. It transcends market structure as it can be meaningfully computed for any type of trading in any type of market (be it trading through limit-orders only, through market-orders only, through a mix of both, or in a completely different market structure).

   Null hypothesis 6: Relative gross trading revenue (GTR) for clients has not changed over time.

B. Definition of Variables

Table B1: Variable descriptions. This table provides detailed descriptions of how the variables that enter the regression analyses are defined, elicited, and constructed. Columns "*#fincap*" and "*cascad*" indicate whether the variable was elicited in the entry- or exit survey of *#fincap* or whether the variable was recorded during *cascad*'s evaluation process.

Variable	*#fincap*	*cascad*	Description
Academic quality: » *Seniority*	☑	☐	Indicator variable; takes value one if at least one team member is a tenured faculty (i.e., associate or full professor), zero otherwise.
Academic quality: » *Citations*	☑	☐	Upon registration for *#fincap*, participants self-reported their number of Google Scholar citations. Participants were asked to provide an estimate instead in case they did not have a Google Scholar profile. On the team level, the variable is defined as the maximum value per team. In the analyses, the number of citations enters in logs, i.e., as $\log(c+1)$.
Academic quality: » *Top publications*	☑	☐	Indicator variable; takes value one if at least one team member has published in a top-3 finance journal (*JoF*, *JFE*, *RFS*) and/or a top-5 economics journal (*AER*, *ECMA*, *JPE*, *REStud*, *QJE*), zero otherwise.
Academic quality: » *Expertise*	☑	☐	Expertise is defined as the mean of participants' self-assessed expertise in empirical finance and their self-assessed expertise in market liquidity. Both items were indicated on a Likert scale from 0 to 10, i.e., expertise takes values between 0 and 10 in steps of 0.5. On the team level, the variable is defined as the maximum value per team.
Coding skills: » *Parallel computing*	☐	☑	Indicator variable; takes value one if parallel computing techniques were used, zero otherwise.
Coding skills: » *Loops / matrix operations*	☐	☑	Indicator variable; takes value one if loops and/or matrix operations were used instead of copy-pasting code ("Don't Repeat Yourself") were used, zero otherwise.

(continued on next page)

Table B1—continued

Variable	*#fincap*	*cascad*	Description
Coding skills: » *Large data*	☑	☐	Indicator variable; takes value one at least one team member worked with datasets comparable in size to the *#fincap* dataset (720M observations) before, zero otherwise; based on participants' self-reports with respect to the question: "What is the largest dataset you have worked with so far (in terms of observations)?" (answers were provided in terms of log-scaled buckets: <10k, 10k–100k, ..., 1b–10b, >10b).
Coding skills: » *Coding skills*	☑	☐	Indicator variable; takes value one if a research team considers their own coding skills to be "excellent", zero otherwise; based on participants' self-reports with respect to the question: "How would you rate the coding skills of your team?" (answers were provided on an ordinal scale from *A* "excellent" to *D* "low".).
Coauthor: » *Team of two*	☑	☐	Indicator variable; takes value one if the research team consists of two researchers, zero otherwise (i.e., if the team consists of only one researcher).
Gender: » *Female*	☑	☐	Indicator variable; takes value one if at least one of the research team members is female, zero otherwise (i.e., if the team only consists of males).
Location: » *Asia-Pacific*, » *Europe*, & » *North America*	☑	☐	Indicator variables for the Asia-Pacific region, for Europe, and for Northern America; the three variables take value one if at least one of the team members is from the respective region, zero otherwise.
Software: » *Matlab*, » *Python*, » *R*, » *SAS*, & » *Stata*	☐	☑	Indicator variables for Matlab, Python, R, SAS, and Stata, i.e., the five most frequently used software applications/programming languages used in *#fincap*; the five variables take value one if the respective software/programming language is used by the research team, zero otherwise.

(continued on next page)

Table B1—continued

Variable	*#fincap*	*cascad*	Description
Research questions: » *RQ2–RQ6*	☑	☐	Indicator variables for the research questions examined by the research teams in *#fincap*; we take into account five dichotomous variables for *RQ2* through *RQ6* to account for research question fixed effects.
Research Quality: » *Peer evaluation*	☑	☐	Each paper in *#fincap* was assessed by two independent peers. Evaluators assessed the quality of the analysis for each of the six hypotheses and the overall paper on a scale from 0 ("very weak") to 10 ("excellent"). Scores are demeand per evaluator (to account for evaluator fixed-effects) and averaged across the two independent evaluators per research team.
Research Quality: » *Outlier result*	☑	☐	Indicator variable for whether a research team's result is in the 2.5 or 97.5 percentile of the distribution of all teams' reported effect size estimates per hypothesis. (Robustness tests for alternative definitions of "outlier result" are presented in Appendix F.)
Code complexity: » *Number of software*	☐	☑	Number of different software applications and/or programming languages used by a research team.
Code complexity: » *Number of script files*	☐	☑	Number of different files written in the various software applications and/or programming languages.
Code complexity: » *Size of software*	☐	☑	Aggregate size of all files written in the various software applications and/or programming languages, measured in kilobytes (kb).
Code complexity: » *Lack of master file*	☐	☑	Indicator variable; takes value one if no master-file was provided by the research team, zero otherwise.
Code complexity: » *Actual CPU time*	☐	☑	Actual CPU time (in minutes) required in *cascad*'s computational reproduction attempt (see Footnote 8 for details about the computer specification used by *cascad*).

(continued on next page)

Table B1—continued

Variable	*#fincap*	*cascad*	Description
Code complexity: » *Help from verificator*	☐	☑	Indicator variable; takes value one if help from the reproducibility verificator was needed to attempt the reproduction (e.g., whether the verificator had to modify the code, change paths to make the code run, debug, etc.), zero otherwise.
Documentation Quality: » *Readme file*	☐	☑	Indicator variable; takes value one if a readme file was provided by the research team, zero otherwise.
Documentation Quality: » *Size of readme file*	☐	☑	Size of the readme file (in .txt format) in kilobytes (kb). Files provided in other formats have been converted into .txt before measuring the file size. The variable takes value zero if no readme file was provided.
Documentation Quality: » *Software requirements*	☐	☑	Indicator variable; takes value one if software requirements were specified, zero otherwise.
Documentation Quality: » *Runtime*	☐	☑	Indicator variable; takes value one if runtime was specified, zero otherwise.
Documentation Quality: » *Computer specification*	☐	☑	Indicator variable; takes value one if computer specifications were provided, zero otherwise.
Documentation Quality: » *Instructions to verificator*	☐	☑	Indicator variable; takes value one if instructions how to reproduce the results were specified, zero otherwise.
Documentation Quality: » *Mapping output/results*	☐	☑	Indicator variable; takes value one if a mapping of analysis outputs and results in the paper was provided, zero otherwise. The variable also takes value one if the analysis outputs were automatically displayed in the order in which they were presented in the paper, which makes any mapping unnecessary.

C. Bugs and Problems

Table C1: Bugs and problems preventing the generation of results. The table lists bugs and problems identified in attempting to reproduce results in #*fincap* and in the AEA/RES sample that resulted in a failure to generate a result. All numbers are in percentage terms, with the overall number of results as the base rate: for #*fincap*, percentages refer to 1,008 tests; for AEA/RES, percentages refer to 818 items (i.e., results reported in tables and/or figures). For #*fincap*, the columns *R0* and *R1* refer indicate the percentages of non-generatable results before and after the interventions of the *cascad* verificator, respectively; for the AEA/RES sample, *R1* and *R2* indicate initial submissions (Round 1) and replication kits revised by the original authors (Round 2), respectively.

			#fincap		**AEA/RES**	
Type	*Problems*	*Examples*	*R0*	*R1*	*R1*	*R2*
CPU/Memory	Time constraint exceeded	Code fails to complete within 168 hours (1 week)	2.5	2.5	2.0	2.0
	Insufficient CPU/memory	Code requires more than 512 GB of RAM	1.4	1.4	—	2.7
Code/Scripts	Improperly named variables	Code uses one variable that do not exist or one which is misspelled	10.8	5.7	0.9	—
	Code fails to import the raw data	Python code attempts to load the data but cannot parse it	5.8	4.2	5.3	5.7
	Data conversion failure	Code fails to convert the .csv into .Rdata files	3.0	3.0	—	—
	Missing/unreadable scripts	Scripts become unreadable due to conversion errors from MAC	2.1	2.1	—	—
	Code runs but does not produce results	SAS code runs but does not generate results for one #*fincap* hypothesis	1.6	1.6	—	—
	Problems with intermediary results	Intermediary results generated with SAS but R fails to use them	3.3	0.3	0.9	0.2
	Lines of code must be added/removed	Part of the code was bugged but not used to generate results	3.8	—	—	—
	Code fails to compile	Fortran code does not generate any executable file	—	—	1.3	1.3
Data	Missing data	Missing required dataset from the American Time Use Survey (ATUS)	—	—	4.9	—
	Altered data	Data provider (French Customs) updated the raw data	—	—	3.1	—
	Restricted data access	No access to National Center for Education Stat. (NCES) outside the US	—	—	0.2	0.2
	Inappropriate data format	Code expected .dta files instead of .csv	0.5	—	—	—
Readme	Information to map output/results	Missing explanation on where to find results in a 100-page log file	3.0	3.0	0.1	0.4
	Information about code/software	Missing command to run a given Fortran code	1.8	1.8	2.4	2.2
	Information about data access	Missing instructions to download specific data from IPUMS platform	—	—	2.4	—
Software	Incompatible environment	Inability to run SAS code in Linux	1.8	1.8	6.7	2.0
	Unavailable libraries/software	Specific R library removed from the CRAN repository	1.9	1.9	—	—
	Versioning	Code runs in Matlab R2019b but not in more recent versions	3.0	—	—	—
Total			**46.1**	**29.1**	**30.3**	**19.7**

D. Robustness Tests: Alternative Model Specifications

Below, we present robustness tests for the analyses on hypotheses *H1* (impact of pre-determined variables on reproducibility) and *H2* (association of co-determined variables and reproducibility) in the main text.

In a first set of robustness tests, we replace the binary dependent variable indicating full reproducibility by the ordinal reproducibility score: Models (1a) and (2a) reported in Table D1 replicate models (1) and (2) presented in Table 5 in the main text; models (3a) and (4a) reported Table D2 replicates models (3) and (4) presented in Table 6.

In a second set of robustness tests, we extend models (3) and (4) reported in Table 6 by controlling for the set of pre-determined variables used in model (1). Table D3 replicates models (3) and (4) presented in Table 6 in the main text but controls for the variation in pre-determined variables (based on a logistic model). For the sake of completeness, Table D4 replicates models (3) and (4) presented in Table 6 in the main text but *(i)* replaces the binary dependent variable by the ordinal reproducibility score, and *(ii)* controls for the set of pre-determined variables.

Table D1: Ordinary least squares regressions of the reproducibility score (0, 25, ..., 100) on pre-determined covariates. *PC-1* indicates the first principal component from a principal component analysis of the covariates associated with academic quality and coding skills, respectively. Variables marked with † are dichotomous; see Table B1 in Appendix B for details. The bottom panel reports $F(df_1, df_2)$-statistics, with df_1 being defined as the number of coefficients k and $df_2 = 167$, for Wald tests for joint statistical significance of groups of covariates. $n = 1,008$, clustered for 168 research teams, in both models. Adj. R^2 is 0.080 and 0.104 for models (1a) and (2a), respectively. p-values are reported in parentheses; * $p < 0.05$.

	Model (1a)		Model (2a)	
Academic Quality:				
» PC-1	−3.012	(0.149)		
» Seniority†			−14.246	(0.083)
» Top Publication†			4.816	(0.582)
» Citations (in logs)			−0.294	(0.881)
» Expertise (0–10)			0.512	(0.827)
Coding Skills:				
» PC-1	6.741*	(0.026)		
» Parallel Comp.†			5.519	(0.608)
» Loops/Matrix Operations†			28.291*	(0.004)
» Large Data†			0.164	(0.981)
» Coding Skills†			−5.147	(0.464)
Coauthor:				
» Team of Two†	0.681	(0.934)	−0.250	(0.975)
Gender:				
» Female†	−9.658	(0.190)	−8.468	(0.260)
Location:				
» North America†	−2.842	(0.726)	−3.314	(0.692)
» Asia-Pacific†	−13.266	(0.195)	−14.561	(0.166)
» Other Continent†	0.814	(0.950)	5.770	(0.652)
Software:				
» Matlab†	−9.005	(0.420)	−8.559	(0.418)
» Python†	−0.773	(0.935)	−0.904	(0.922)
» R†	−6.350	(0.472)	−8.249	(0.349)
» SAS†	14.480	(0.115)	16.021	(0.080)
» Stata†	−11.509	(0.157)	−11.455	(0.150)
Research Questions:				
» RQ2†	−0.149	(0.950)	−0.149	(0.950)
» RQ3†	6.101*	(0.007)	6.101*	(0.008)
» RQ4†	0.298	(0.900)	0.298	(0.901)
» RQ5†	4.762*	(0.046)	4.762*	(0.047)
» RQ6†	0.893	(0.711)	0.893	(0.712)
Constant	69.308*	(0.000)	48.687*	(0.023)
Wald Tests:				
» Overall model	2.073*	(0.010)	2.054*	(0.005)
» Academic quality ($k = 4$)			1.115	(0.351)
» Coding skills ($k = 4$)			2.238	(0.067)
» Location ($k = 3$)	0.605	(0.613)	0.828	(0.480)
» Software ($k = 5$)	1.743	(0.127)	1.955	(0.088)
» Research questions ($k = 5$)	2.524*	(0.031)	2.509*	(0.032)

Table D2: Ordinary least squares regressions of the reproducibility score (0, 25, ..., 100) on co-determined covariates. *PC-1* indicates the first principal component from a principal component analysis of the covariates associated with code complexity and documentation quality, respectively. Variables marked with † are dichotomous; see Table B1 in Appendix B for details. The bottom panel reports $F(df_1, df_2)$-statistics, with df_1 being defined as the number of coefficients k and $df_2 = 150$, for Wald tests for joint statistical significance of groups of covariates. $n = 906$, clustered for 151 research teams, in both models. Adj. R^2 is 0.052 and 0.092 for models (3a) and (4a), respectively. p-values are reported in parentheses; * $p < 0.05$.

	Model (3a)		Model (4a)	
Research Quality:				
» Peer Evaluation (0–10)	0.779	(0.508)	0.510	(0.665)
» Outlier Result†	−23.848*	(0.008)	−20.272*	(0.015)
Code Complexity:				
» PC-1	−5.697*	(0.003)		
» Number of Software			−10.132	(0.138)
» Number of Script Files			−0.240	(0.406)
» Size of Software (in kb)			−0.001	(0.698)
» Actual CPU Time (in minutes)			0.002	(0.419)
» Lack of Master File†			3.748	(0.561)
» Help from Verificator†			10.170	(0.080)
Documentation Quality:				
» PC-1	5.407*	(0.018)		
» Readme File†			2.301	(0.946)
» Size of Readme File (in kb)			0.226	(0.867)
» Software Requirements†			−9.294	(0.452)
» Runtime†			1.812	(0.853)
» Computer Specification†			11.850	(0.189)
» Instructions to Verificators†			34.279	(0.213)
» Mapping Output/Results†			17.150	(0.067)
Constant	70.622*	(0.000)	39.848*	(0.048)
Wald Tests:				
» Overall model	5.200*	(0.001)	2.684*	(0.001)
» Code complexity ($k = 6$)			1.933	(0.079)
» Documentation quality ($k = 7$)			2.204*	(0.037)

Table D3: Logit regressions of the reproducibility indicator on co-determined covariates, controlling for the variation in pre-determined variables. *PC-1* indicates the first principal component from a principal component analysis of the covariates associated with code complexity and documentation quality, respectively. Both models control for the first principal components of academic quality and coding skills and indicator variables for teams of two, teams involving a female team mate, location, software, and research question fixed effects. Variables marked with † are dichotomous; see Table B1 in Appendix B for details. The bottom panel reports $\chi^2(df)$-statistics, with df being defined as the number of coefficients (k), for Wald tests for joint statistical significance of groups of covariates. $n = 906$, clustered for 151 research teams, in both models. McFadden's Pseudo R^2 is 0.104 and 0.142 for models (3b) and (4b), respectively. p-values are reported in parentheses; * $p < 0.05$.

	Model (3b)		Model (4b)	
Research Quality:				
» Peer Evaluation (0–10)	0.006	(0.710)	0.004	(0.807)
» Outlier Result†	−0.245*	(0.008)	−0.210*	(0.013)
Code Complexity:				
» PC-1	−0.043	(0.126)		
» Number of Software			−0.220*	(0.040)
» Number of Script Files			0.005	(0.199)
» Size of Software (in kb)			0.000	(0.182)
» Actual CPU Time (in minutes)			0.000	(0.814)
» Lack of Master File†			0.076	(0.317)
» Help from Verificator†			0.094	(0.244)
Documentation Quality:				
» PC-1	0.050	(0.061)		
» Readme File†			0.113	(0.758)
» Size of Readme File (in kb)			0.000	(0.985)
» Software Requirements†			−0.077	(0.707)
» Runtime†			−0.003	(0.975)
» Computer Specification†			0.154	(0.287)
» Instructions to Verificators†			0.269	(0.235)
» Mapping Output/Results†			0.136	(0.083)
Wald Tests:				
» Overall model	45.887*	(0.001)	78.766*	(0.000)
» Code complexity ($k = 6$)			8.228	(0.222)
» Documentation quality ($k = 7$)			10.263	(0.174)

Table D4: Ordinary least squares regressions of the reproducibility score (0, 25, ..., 100) on co-determined covariates, controlling for the variation in pre-determined variables. *PC-1* indicates the first principal component from a principal component analysis of the covariates associated with code complexity and documentation quality, respectively. Both models control for the first principal components of academic quality and coding skills and indicator variables for teams of two, teams involving a female team mate, location, software, and research question fixed effects. Variables marked with † are dichotomous; see Table B1 in Appendix B for details. The bottom panel reports $F(df_1, df_2)$-statistics, with df_1 being defined as the number of coefficients k and $df_2 = 150$, for Wald tests for joint statistical significance of groups of covariates. $n = 906$, clustered for 151 research teams, in both models. Adj. R^2 is 0.144 and 0.194 for models (3c) and (4c), respectively. p-values are reported in parentheses; * $p < 0.05$.

	Model (3c)		Model (4c)	
Research Quality:				
» Peer Evaluation (0–10)	0.752	(0.579)	0.466	(0.733)
» Outlier Result†	−23.300*	(0.005)	−19.418*	(0.009)
Code Complexity:				
» PC-1	−4.827*	(0.023)		
» Number of Software			−23.350*	(0.027)
» Number of Script Files			0.189	(0.565)
» Size of Software (in kb)			−0.002	(0.641)
» Actual CPU Time (in minutes)			0.002	(0.454)
» Lack of Master File†			5.030	(0.469)
» Help from Verificator†			15.149*	(0.015)
Documentation Quality:				
» PC-1	5.770*	(0.025)		
» Readme File†			8.419	(0.786)
» Size of Readme File (in kb)			0.723	(0.597)
» Software Requirements†			−12.640	(0.292)
» Runtime†			−3.431	(0.739)
» Computer Specification†			15.799	(0.133)
» Instructions to Verificators†			36.737	(0.088)
» Mapping Output/Results†			18.399*	(0.027)
Constant	68.535*	(0.000)	23.068	(0.619)
Wald Tests:				
» Overall model	4.557*	(0.002)	2.881*	(0.001)
» Code complexity ($k = 6$)			2.349*	(0.034)
» Documentation quality ($k = 7$)			2.697*	(0.012)

E. Robustness Tests: Individual Level

Below, we present auxiliary analyses for the results of the pre-determined cross-sectional determinants of reproducibility with a focus on whether full reproducibility rates are sensitive to the variation in team composition. In particular, we present two sets of additional analyses: (i) Table E1 reports regression estimates of models (1) and (2) on a dataset expanded to the individual-level (300 individuals $\times$ 6 hypotheses = 1,800 observations), with standard errors clustered on the research team level (see Table 5 in the main text for the team-level estimates); (ii) Table E2 tabulates the estimates for the same regression models on the sample of single-authored papers (36 authors $\times$ 6 hypothesis = 216 observations). Note that we do not estimate models (3) and (4) on the expanded dataset since all variables entering these models are measured on the team level. Notably, the results turn out to be qualitatively robust for all but one independent variable: seniority. In both analyses, the indicator for holding an associate or full professorship turns out to be significantly *negative*, with a marginal effect of -16.1 and -34.0 percentage points in (i) and (ii), respectively.

In addition, we examine (on the team-level) whether the team composition (i.e., "junior-junior," "junior-senior," or "senior-senior") systematically correlates with reproducibility. We re-estimate models (1) and (2) but replace the indicator for "team-of-two" by three dummies for the team composition (with single-authored papers constituting the base category). The results are reported in Table E3. Notably, none of the three dichotomous team composition variables turns out to be statistically significant, and the three pairwise comparisons (Wald tests; not reported) between the coefficient estimates are insignificant.

Table E1: Logit regressions of the reproducibility indicator on pre-determined covariates on the individual level. Estimates are reported in terms of marginal effects. *PC-1* indicates the first principal component from a principal component analysis of the covariates associated with academic quality and coding skills, respectively. Variables marked with † are dichotomous; see Table B1 in Appendix B for details. The bottom panel reports $\chi^2(df)$-statistics, with df being defined as the number of coefficients (k), for Wald tests for joint statistical significance of groups of covariates. $n = 1,800$, clustered for 168 research teams, in both models. McFadden's Pseudo R^2 is 0.047 and 0.066 for models (1I) and (2I), respectively. p-values are reported in parentheses; * $p < 0.05$.

	Model (1i)		Model (2i)	
Academic Quality:				
» PC-1	−0.015	(0.308)		
» Seniority†			−0.161*	(0.023)
» Top Publication†			0.073	(0.343)
» Citations (in logs)			0.003	(0.807)
» Expertise (0–10)			0.014	(0.387)
Coding Skills:				
» PC-1	0.075*	(0.015)		
» Parallel Comp.†			0.094	(0.383)
» Loops/Matrix Operations†			0.303*	(0.001)
» Large Data†			−0.030	(0.594)
» Coding Skills†			−0.011	(0.880)
Coauthor:				
» Team of Two†	0.049	(0.530)	0.021	(0.789)
Gender:				
» Female†	−0.059	(0.304)	−0.037	(0.525)
Location:				
» North America†	0.013	(0.872)	0.012	(0.882)
» Asia-Pacific†	−0.135	(0.195)	−0.143	(0.154)
» Other Continent†	0.116	(0.297)	0.159	(0.146)
Software:				
» Matlab†	−0.106	(0.355)	−0.106	(0.324)
» Python†	0.012	(0.904)	0.006	(0.951)
» R†	−0.065	(0.466)	−0.068	(0.435)
» SAS†	0.085	(0.357)	0.116	(0.204)
» Stata†	−0.117	(0.135)	−0.116	(0.137)
Research Questions:				
» RQ2†	−0.007	(0.852)	−0.007	(0.852)
» RQ3†	0.100*	(0.003)	0.100*	(0.003)
» RQ4†	0.017	(0.650)	0.017	(0.650)
» RQ5†	0.097*	(0.011)	0.097*	(0.011)
» RQ6†	0.020	(0.578)	0.020	(0.578)
Wald Tests:				
» Overall model	35.795*	(0.005)	54.020*	(0.000)
» Academic quality ($k = 4$)			6.329	(0.176)
» Coding skills ($k = 4$)			13.786*	(0.008)
» Location ($k = 3$)	3.328	(0.344)	4.791	(0.188)
» Software ($k = 5$)	6.055	(0.301)	7.004	(0.220)
» Research questions ($k = 5$)	14.409*	(0.013)	14.411*	(0.013)

Table E2: Logit regressions of the reproducibility indicator on pre-determined covariates on the subsample of teams involving only one researcher. Estimates are reported in terms of marginal effects. *PC-1* indicates the first principal component from a principal component analysis of the covariates associated with academic quality and coding skills, respectively. Variables marked with † are dichotomous; see Table B1 in Appendix B for details. The bottom panel reports $\chi^2(df)$-statistics, with df being defined as the number of coefficients (k), for Wald tests for joint statistical significance of groups of covariates. $n = 216$, clustered for 36 researchers, in both models. McFadden's Pseudo R^2 is 0.140 and 0.231 for models (1S) and (2S), respectively. p-values are reported in parentheses; * $p < 0.05$.

	Model (1s)		Model (2s)	
Academic Quality:				
» PC-1	−0.057	(0.433)		
» Seniority†			−0.340*	(0.046)
» Top Publication†			0.245	(0.324)
» Citations (in logs)			−0.006	(0.915)
» Expertise (0-10)			0.014	(0.786)
Coding Skills:				
» PC-1	0.133*	(0.014)		
» Parallel Comp.†			0.232	(0.238)
» Loops†			0.456*	(0.006)
» Large Data†			−0.064	(0.701)
» Coding Skills†			−0.203	(0.329)
Gender:				
» Female†	0.035	(0.852)	0.013	(0.947)
Location:				
» North America†	0.077	(0.692)	−0.025	(0.866)
» Asia-Pacific†	0.067	(0.692)	0.036	(0.836)
» Other Continent†	0.022	(0.913)	0.114	(0.602)
Software:				
» Matlab†	0.192	(0.485)	−0.015	(0.958)
» Python†	−0.081	(0.720)	−0.209	(0.446)
» R†	−0.170	(0.543)	−0.277	(0.249)
» SAS†	0.148	(0.471)	0.132	(0.488)
» Stata†	−0.115	(0.573)	−0.289	(0.188)
Research Questions:				
» RQ2†	0.000	(1.000)	0.000	(1.000)
» RQ3†	0.110	(0.243)	0.111	(0.241)
» RQ4†	−0.028	(0.659)	−0.027	(0.660)
» RQ5†	0.138*	(0.047)	0.139*	(0.048)
» RQ6†	0.000	(1.000)	0.000	(1.000)
Wald Tests:				
» Overall model	39.247*	(0.001)	205.661*	(0.000)
» Academic quality ($k = 4$)			8.917	(0.063)
» Coding skills ($k = 4$)			23.733*	(0.000)
» Location ($k = 3$)	0.237	(0.971)	0.526	(0.913)
» Software ($k = 5$)	6.513	(0.259)	12.776*	(0.026)
» Research questions ($k = 5$)	7.862	(0.164)	7.325	(0.198)

Table E3: Logit regressions of the reproducibility indicator on pre-determined covariates, replacing the "team-of-two" indicator by indicators "junior/junior," "junior/senior," and "senior/senior." Estimates are reported in terms of marginal effects. *PC-1* indicates the first principal component from a principal component analysis of the covariates associated with academic quality and coding skills, respectively. Variables marked with † are dichotomous; see Table B1 in Appendix B for details. The bottom panel reports $\chi^2(df)$-statistics, with df being defined as the number of coefficients (k), for Wald tests for joint statistical significance of groups of covariates. $n = 216$, clustered for 36 researchers, in both models. McFadden's Pseudo R^2 is 0.140 and 0.231 for models (1S) and (2S), respectively. p-values are reported in parentheses; * $p < 0.05$.

	(1t)		(2t)	
Academic Quality:				
» PC-1	0.000	(0.994)		
» Seniority†			−0.211	(0.152)
» Top Publication†			0.064	(0.482)
» Citations (in logs)			−0.001	(0.959)
» Expertise (0–10)			0.008	(0.718)
Coding Skills:				
» PC-1	0.095*	(0.001)		
» Parallel Comp.†			0.108	(0.311)
» Loops/Matrix Operations†			0.334*	(0.000)
» Large Data†			0.000	(1.000)
» Coding Skills†			−0.011	(0.876)
Team Composition:				
» Junior-Junior†	0.085	(0.319)	0.018	(0.837)
» Junior-Senior†	0.016	(0.907)	0.113	(0.437)
» Senior-Senior†	−0.100	(0.458)	0.040	(0.797)
Gender:				
» Female†	−0.058	(0.410)	−0.051	(0.474)
Location:				
» North America†	−0.011	(0.894)	−0.018	(0.822)
» Asia-Pacific†	−0.151	(0.127)	−0.159	(0.106)
» Other Continent†	0.044	(0.721)	0.105	(0.403)
Software:				
» Matlab†	−0.060	(0.580)	−0.062	(0.553)
» Python†	−0.009	(0.921)	−0.007	(0.944)
» R†	−0.054	(0.541)	−0.076	(0.375)
» SAS†	0.132	(0.127)	0.148	(0.090)
» Stata†	−0.114	(0.137)	−0.119	(0.114)
Research Questions:				
» RQ2†	−0.006	(0.862)	−0.006	(0.862)
» RQ3†	0.101*	(0.003)	0.101*	(0.003)
» RQ4†	0.012	(0.732)	0.012	(0.732)
» RQ5†	0.101*	(0.006)	0.101*	(0.005)
» RQ6†	0.018	(0.613)	0.018	(0.613)
Wald χ^2	33.938		38.599	
$p > \chi^2$	0.019		0.040	
Pseudo R^2	0.064		0.080	
No. of Clusters	168		168	
No. of Observations	1008		1008	

F. Robustness Tests: Outlier Results

Below we present a supplementary analysis with respect to the relationship between full reproducibility and outlier results in #*fincap* (panel (a) in Figure F1) as well as a robustness perspective on the estimated coefficient of "outlier result" in Table 6 in the main text (panel (b) in Figure F1). Panel (a) in Figure F1 displays the frequency of full reproducibility scores for each decile of the results provided by all teams for a given hypothesis in #*fincap*. The inverted u-shaped relationship indicates that results lying in the left and right tails of the distribution are, on average, associated with lower reproducibility rates. Panel (b) in Figure F1 highlights that the estimated effect of "outlier result" in our multivariate regression setting (Table 6) is robust to various thresholds to define the "outlier results" indicator variable.

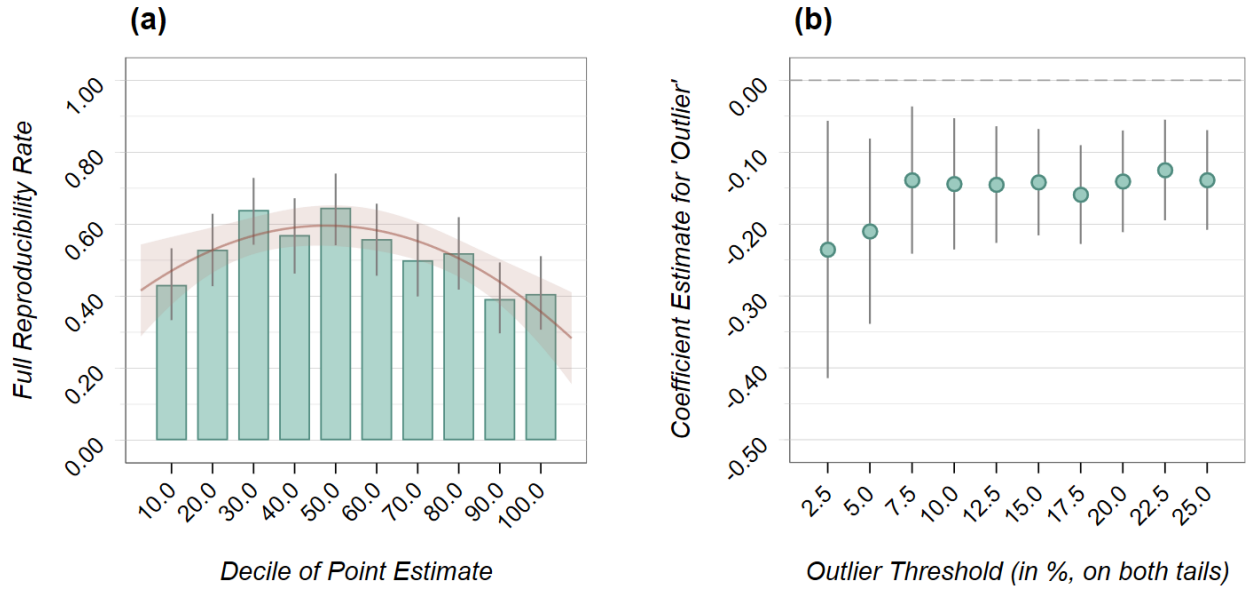

Figure F1: Reproducibility of outlier results. **(a)** Reproducibility rates and 95% confidence intervals (CI) as a function of the deciles of the distribution of all teams' effect size estimates and the two-way quadratic prediction (and corresponding 95% CI). x-axis labels indicate the upper bound of the interval of percentiles that are aggregated (e.g., $x = 20$ corresponds to the interval $(10, 20]$. $n \in [93, 111]$ for each bar. **(b)** Coefficient estimates (in terms of marginal effects at means) for varying thresholds of the "outlier result" indicator variable in regression models (3). x-axis labels correspond to %-thresholds on both tails (e.g., $x = 10$ implies that the outlier dummy in the regression model takes value one for the 10% smallest *and* the 10% largest effect size estimates per hypothesis, zero otherwise). $n = 1,008$ in each regression model.