(Forthcoming in Journal of Finance)

Non-Standard Errors*

Albert J. Menkveld[200,121], *Anna Dreber*[113], *Felix Holzmeister*[149], *Juergen Huber*[149], *Magnus Johannesson*[113], *Michael Kirchler*[149], *Sebastian Neusüss*[1], *Michael Razen*[149], *Utz Weitzel*[200,100,121], David Abad-Díaz[133], Menachem Abudy[12], Tobias Adrian[58], Yacine Ait-Sahalia[95], Olivier Akmansoy[17,21], Jamie T. Alcock[169], Vitali Alexeev[179], Arash Aloosh[81], Livia Amato[139], Diego Amaya[204], James J. Angel[44], Alejandro T. Avetikian[93], Amadeus Bach[154], Edwin Baidoo[118], Gaetan Bakalli[6], Li Bao[122], Andrea Barbon[174], Oksana Bashchenko[104], Parampreet C. Bindra[149], Geir H. Bjønnes[8], Jeffrey R. Black[158], Bernard S. Black[85], Dimitar Bogoev[32], Santiago Bohorquez Correa[129], Oleg Bondarenko[147], Charles S. Bos[200], Ciril Bosch-Rosa[116], Elie Bouri[69], Christian Brownlees[130], Anna Calamia[115], Viet Nga Cao[78], Gunther Capelle-Blancard[131], Laura M. Capera Romero[200], Massimiliano Caporin[170], Allen Carrion[158], Tolga Caskurlu[134], Bidisha Chakrabarty[107], Jian Chen[97], Mikhail Chernov[124], William Cheung[202], Ludwig B. Chincarini[172], Tarun Chordia[37], Sheung Chi Chow[7], Benjamin Clapham[46], Jean-Edouard Colliard[49], Carole Comerton-Forde[157], Edward Curran[74], Thong Dao[87], Wale Dare[47], Ryan J. Davies[9], Riccardo De Blasis[67], Gianluca F. De Nard[193], Fany Declerck[122], Oleg Deev[75], Hans Degryse[61], Solomon Y. Deku[87], Christophe Desagre[125], Mathijs A. van Dijk[38], Chukwuma Dim[43], Thomas Dimpfl[146], Yun Jiang Dong[97], Philip A. Drummond[78], Tom Dudda[117], Teodor Duevski[49], Ariadna Dumitrescu[36], Teodor Dyakov[33], Anne Haubo Dyhrberg[178], Michał Dzieliński[114], Asli Eksi[108], Izidin El Kalak[20], Saskia ter Ellen[22], Nicolas Eugster[171], Martin D. D. Evans[44], Michael Farrell[189], Ester Felez-Vinas[179], Gerardo Ferrara[11], El Mehdi Ferrouhi[56], Andrea Flori[92], Jonathan T. Fluharty[203], Sean D. V. Foley[74], Kingsley Y. L. Fong[165], Thierry Foucault[49], Tatiana Franus[14], Francesco Franzoni[195], Bart Frijns[89], Michael Frömmel[45], Servanna M. Fu[141], Sascha C. Füllbrunn[100], Baoqing Gan[179], Ge Gao[135], Thomas P. Gehrig[188], Roland Gemayel[65], Dirk Gerritsen[197], Javier Gil-Bazo[130,13], Dudley Gilder[20], Lawrence R. Glosten[25], Thomas

Gomez[197], Arseny Gorbenko[78], Joachim Grammig[184], Vincent Grégoire[48], Ufuk Güçbilmez[143], Björn Hagströmer[114], Julien Hambuckers[47], Erik Hapnes[1], Jeffrey H. Harris[3], Lawrence Harris[173], Simon Hartmann[199], Jean-Baptiste Hasse[2], Nikolaus Hautsch[188], Xue-Zhong (Tony) He[205], Davidson Heath[185], Simon Hediger[193], Terrence Hendershott[123], Ann Marie Hibbert[203], Erik Hjalmarsson[144], Seth Hoelscher[77], Peter Hoffmann[39], Craig W. Holden[57], Alex R. Horenstein[159], Wenqian Huang[10], Da Huang[185], Christophe Hurlin[168,21], Konrad Ilczuk[1], Alexey Ivashchenko[200], Subramanian R. Iyer[164], Hossein Jahanshahloo[20], Naji P. Jalkh[106], Charles M. Jones[24], Simon Jurkatis[11], Petri Jylhä[1], Andreas T. Kaeck[177], Gabriel Kaiser[152], Arzé Karam[30], Egle Karmaziene[200], Bernhard Kassner[162], Markku Kaustia[1], Ekaterina Kazak[153], Fearghal Kearney[98], Vincent van Kervel[94], Saad A. Khan[48], Marta K. Khomyn[179], Tony Klein[98], Olga Klein[190], Alexander Klos[63], Michael Koetter[50], Aleksey Kolokolov[153], Robert A. Korajczyk[85], Roman Kozhan[190], Jan P. Krahnen[46], Paul Kuhle[127], Amy Kwan[178], Quentin Lajaunie[91], F. Y. Eric C. Lam[51], Marie Lambert[47], Hugues Langlois[49], Jens Lausen[46], Tobias Lauter[71], Markus Leippold[193], Vladimir Levin[152], Yijie Li[109], Hui Li[68], Chee Yoong Liew[126], Thomas Lindner[201], Oliver Linton[138], Jiacheng Liu[96], Anqi Liu[178], Guillermo Llorente[127], Matthijs Lof[1], Ariel Lohr[4], Francis Longstaff[124], Alejandro Lopez-Lira[142], Shawn Mankad[27], Nicola Mano[102], Alexis Marchal[35], Charles Martineau[182], Francesco Mazzola[38], Debrah Meloso[115], Michael G. Mi[178], Roxana Mihet[101], Vijay Mohan[99], Sophie Moinas[122], David Moore[72], Liangyi Mu[120], Dmitriy Muravyev[76], Dermot Murphy[147], Gabor Neszveda[59], Christian Neumeier[60], Ulf Nielsson[26], Mahendrarajah Nimalendran[142], Sven Nolte[100], Lars L. Norden[114], Peter W. O'Neill[42], Khaled Obaid[18], Bernt A. Ødegaard[175], Per Östberg[193], Emiliano Pagnotta[110], Marcus Painter[107], Stefan Palan[145], Imon J. Palit[99], Andreas Park[183], Roberto Pascual[194], Paolo Pasquariello[160], Lubos Pastor[139], Vinay Patel[179], Andrew J. Patton[29], Neil D. Pearson[148,19], Loriana Pelizzon[46], Michele Pelli[105], Matthias Pelster[90], Christophe Pérignon[49,21], Cameron Pfiffer[167], Richard Philip[178], Tomáš Plíhal[75], Puneet Prakash[77], Oliver-Alexander Press[26], Tina Prodromou[192], Marcel Prokopczuk[71], Talis Putnins[179], Ya Qian[1], Gaurav Raizada[53], David Rakowski[180], Angelo Ranaldo[174], Luca Regis[181], Stefan Reitz[64], Thomas Renault[196], Rex W. Renjie[200], Roberto Reno[186], Steven J. Riddiough[182], Kalle

Rinne[152], Paul J. Rintamäki[1], Ryan Riordan[97], Thomas Rittmannsberger[149], Iñaki Rodríguez Longarela[114], Dominik Roesch[112], Lavinia Rognone[153], Brian Roseman[88], Ioanid Rosu[49], Saurabh Roy[156], Nicolas Rudolf[151], Stephen R. Rush[15], Khaladdin Rzayev[140,66], Aleksandra A. Rzeźnik[206], Anthony Sanford[155], Harikumar Sankaran[82], Asani Sarkar[41], Lucio Sarno[138], Olivier Scaillet[103], Stefan Scharnowski[154], Klaus R. Schenk-Hoppé[153], Andrea Schertler[145], Michael Schneider[28,70], Florian Schroeder[74], Norman Schürhoff[104], Philipp Schuster[176], Marco A. Schwarz[31,16], Mark S. Seasholes[4], Norman J. Seeger[200], Or Shachar[41], Andriy Shkilko[204], Jessica Shui[40], Mario Sikic[193], Giorgia Simion[201], Lee A. Smales[191], Paul Söderlind[174], Elvira Sojli[165], Konstantin Sokolov[158], Jantje Sönksen[184], Laima Spokeviciute[20], Denitsa Stefanova[152], Marti G. Subrahmanyam[80,79], Barnabas Szaszi[34], Oleksandr Talavera[135], Yuehua Tang[142], Nick Taylor[137], Wing Wah Tham[165], Erik Theissen[154], Julian Thimme[62], Ian Tonks[137], Hai Tran[72], Luca Trapin[136], Anders B. Trolle[26], M. Andreea Vaduva[128], Giorgio Valente[52], Robert A. Van Ness[161], Aurelio Vasquez[55], Thanos Verousis[141], Patrick Verwijmeren[38], Anders Vilhelmsson[73], Grigory Vilkov[43], Vladimir Vladimirov[134], Sebastian Vogel[38], Stefan Voigt[150], Wolf Wagner[38], Thomas Walther[197], Patrick Weiss[198], Michel van der Wel[38], Ingrid M. Werner[119], Joakim Westerholm[178], Christian Westheide[188], Hans C. Wika[84], Evert Wipplinger[200], Michael Wolf[193], Christian C. P. Wolff[152], Leonard Wolk[200], Wing-Keung Wong[5], Jan Wrampelmeyer[200], Zhen-Xing Wu[1], Shuo Xia[50], Dacheng Xiu[139], Ke Xu[187], Caihong Xu[114], Pradeep K. Yadav[166], José Yagüe[163], Cheng Yan[141], Antti Yang[38], Woongsun Yoo[23], Wenjia Yu[1], Yihe Yu[132], Shihao Yu[200], Bart Z. Yueshen[54], Darya Yuferova[86], Marcin Zamojski[144], Abalfazl Zareei[114], Stefan M. Zeisberger[100], Lu Zhang[152], S. Sarah Zhang[153], Xiaoyu Zhang[200], Lu Zhao[111], Zhuo Zhong[157], Zeyang (Ivy) Zhou[192], Chen Zhou[38], Xingyu S. Zhu[113], Marius Zoican[183], and Remco Zwinkels[200]

[1]Aalto University, [2]Aix-Marseille University, [3]American University, [4]Arizona State University, [5]Asia University, [6]Auburn University, [7]Australian National University, [8]BI Norwegian Business School, [9]Babson College, [10]Bank for International Settlements, [11]Bank of England , [12]Bar-Ilan University, [13]Barcelona School of Economics, [14]Bayes Business School, [15]Bowling Green State University, [16]CESifo, [17]CNRS, [18]California State University - East Bay,

[19]Canadian Derivatives Institute, [20]Cardiff University, [21]Cascad, [22]Central Bank of Norway, [23]Central Michigan University , [24]Columbia Business School, [25]Columbia University, [26]Copenhagen Business School, [27]Cornell University, [28]Deutsche Bundesbank, [29]Duke University, [30]Durham University, [31]Düsseldorf Institute for Competition Economics, [32]EDF Energy London, [33]EDHEC Business School, [34]ELTE, Eotvos Lorand University, [35]EPFL, [36]ESADE Business School, Univ. Ramon Llull, [37]Emory University, [38]Erasmus University Rotterdam, [39]European Central Bank, [40]Federal Housing Finance Agency , [41]Federal Reserve Bank of New York, [42]Financial Conduct Authority, [43]Frankfurt School of Finance and Management, [44]Georgetown University, [45]Ghent University, [46]Goethe University Frankfurt, [47]HEC Liège - University of Liège, [48]HEC Montréal, [49]HEC Paris, [50]Halle Institute for Economic Research, [51]Hong Kong Institute for Monetary and Financial Research, [52]Hong Kong Monetary Authority, [53]IIM Ahmedabad, [54]INSEAD, [55]ITAM, [56]Ibn Tofail University, [57]Indiana University, [58]International Monetary Fund, [59]John von Neumann University, [60]Justus-Liebig University, [61]KU Leuven, [62]Karlsruhe Institute of Technology, [63]Kiel University, [64]Kiel university, [65]King's College London, [66]Koç University, [67]LUM University, [68]La Trobe University, [69]Lebanese American University, [70]Leibniz Institute for Financial Research SAFE, [71]Leibniz University Hannover, [72]Loyola Marymount University, [73]Lund University, [74]Macquarie University, [75]Masaryk University, [76]Michigan State University, [77]Missouri State University, [78]Monash University, [79]NYU Shanghai, [80]NYU Stern, [81]Neoma Business School, [82]New Mexico state University, [83]None, [84]Norges Bank, [85]Northwestern University, [86]Norwegian School of Economics (NHH), [87]Nottingham Trent University, [88]Oklahoma State University, [89]Open Universiteit, [90]Paderborn University, [91]Paris Dauphine University, [92]Politecnico di Milano, [93]Pontificia Universidad Católica de Chile, [94]Pontificial University of Chile, [95]Princeton University, [96]Purdue University, [97]Queen's University, [98]Queen's University Belfast, [99]RMIT University, [100]Radboud University, [101]SFI at HEC Lausanne, [102]SFI at USI Lugano, [103]SFI at University of Geneva, [104]SFI at University of Lausanne, [105]SFI at University of Zurich, [106]Saint Joseph University, [107]Saint Louis University, [108]Salisbury University, [109]SandP Global Ratings, [110]Singapore Management University, [111]Southwestern University of Finance and Economics, [112]State University of New York at Buffalo, [113]Stockholm School of Economics, [114]Stockholm University, [115]TBS Education, [116]Technische Universität Berlin, [117]Technische Universität Dresden, [118]Tennessee Technological University, [119]The Ohio State University, [120]The University of Manchester, [121]Tinbergen Institute, [122]Toulouse 1 Capitole University, [123]UC Berkeley, [124]UCLA, [125]UCLouvain, [126]UCSI University, [127]Universidad Autónoma de Madrid, [128]Universidad Carlos III de Madrid, [129]Universidad EAFIT, [130]Universitat Pompeu Fabra, [131]University Paris 1 Pantheon-Sorbonne, [132]University at Buffalo, [133]University of Alicante, [134]University of Amsterdam, [135]University of Birmingham, [136]University of Bologna, [137]University of Bristol, [138]University of Cambridge, [139]University of Chicago Booth School of Business, [140]University of Edinburgh, [141]University of Essex, [142]University of Florida, [143]University of Glasgow , [144]University of Gothenburg, [145]University of Graz, [146]University of Hohenheim, [147]University of

Illinois at Chicago, [148]University of Illinois at Urbana-Champaign, [149]University of Innsbruck, [150]University of København, [151]University of Lausanne, [152]University of Luxembourg, [153]University of Manchester, [154]University of Mannheim, [155]University of Maryland, [156]University of Massachusetts, Amherst, [157]University of Melbourne, [158]University of Memphis, [159]University of Miami, [160]University of Michigan, [161]University of Mississippi, [162]University of Munich (LMU), [163]University of Murcia, [164]University of New Mexico, [165]University of New South Wales, [166]University of Oklahoma, [167]University of Oregon, [168]University of Orléans, [169]University of Oxford, [170]University of Padova, [171]University of Queensland, [172]University of San Francisco, [173]University of Southern California, [174]University of St. Gallen, [175]University of Stavanger, [176]University of Stuttgart, [177]University of Sussex, [178]University of Sydney, [179]University of Technology Sydney, [180]University of Texas at Arlington, [181]University of Torino, [182]University of Toronto, [183]University of Toronto Mississauga, [184]University of Tübingen, [185]University of Utah, [186]University of Verona, [187]University of Victoria, [188]University of Vienna, [189]University of Virginia, [190]University of Warwick, [191]University of Western Australia, [192]University of Wollongong, [193]University of Zurich, [194]University of the Balearic Islands, [195]Università della Svizzera italiana, [196]Université Paris 1 Panthéon-Sorbonne, [197]Utrecht University, [198]Vienna Graduate School of Finance, [199]Vienna University of Economics and Business, [200]Vrije Universiteit Amsterdam, [201]WU Vienna University of Economics and Business, [202]Waseda University, [203]West Virginia University, [204]Wilfrid Laurier University, [205]Xi'an Jiaotong-Liverpool University, [206]York University, [207]Zhongnan University of Economics and Law

November 16, 2022

*The first nine authors in italics are the project coordinators. They conceptualized and designed the project, managed it, conducted the meta-analyses, and wrote the manuscript. Any errors are therefore their sole responsibility. The other authors all significantly contributed to the project by participating either as a member of a research team, or as a peer evaluator. The views expressed here are the authors' and do not represent the views of the Federal Reserve Bank of New York or the Federal Reserve System, or any other of the institutions that the authors are affiliated with or receive financing from. The coordinators thank Andrew Chen, Amit Goyal, Campbell Harvey, Lucas Saru, Eric Uhlmann, and participants at the Microstructure Exchange 2021, Derivatives Forum Frankfurt 2022, Financial Intermediation Research Society (FIRS) 2022, Research in Behavioral Finance Conference (RBFC) 2022, Society for Experimental Finance (SEF) 2022, Society for Financial Econometrics (SoFiE) 2022 where the paper was runner-up for the best-paper prize, Vienna-Copenhagen Conference on Financial Econometrics 2022, and the Western Finance Assocation (WFA) 2022 for valuable comments. They further thank Adam Gill, Eugénie de Jong, Ingrid Löfman, and Elmar Nijkamp for research assistance.

The coordinators 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) an FWF grant P29362, (Huber and Kirchler) FWF SFB F63, (Johannesson) Riksbankens Jubileumsfond grant P21-0168, and (Menkveld) NWO-Vici.

(Forthcoming in Journal of Finance)

Non-Standard Errors

Abstract

In statistics, samples are drawn from a population in a data-generating process (DGP). Standard errors measure the uncertainty in estimates of population parameters. In science, evidence is generated to test hypotheses in an evidence-generating process (EGP). We claim that EGP variation across researchers adds uncertainty: *Non-standard errors* (NSEs). We study NSEs by letting 164 teams test the same hypotheses on the same data. NSEs turn out to be sizable, but smaller for better reproducible or higher rated research. Adding peer-review stages reduces NSEs. We further find that this type of uncertainty is underestimated by participants.

Online appendix with additional results and all forms used in #fincap is available at https://bit.ly/3DIQKrB.

1 Introduction

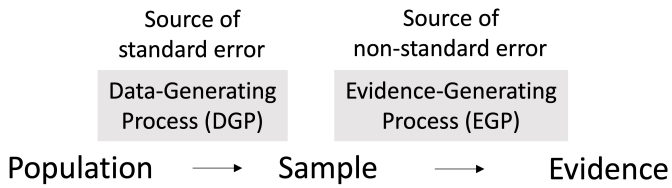
In their recent book, Kahneman, Sibony, and Sunstein (2021) (KSS) discuss variability in human judgment in terms of noise. They illustrate their analysis by judges passing sentence. They decompose total variation in sentencing into two canonical components: *level noise* and *pattern noise* (Ch. 6). Level noise captures the extent to which some judges are more lenient than others. Pattern noise, on the other hand, refers to variation in judgment when the same judge sentences similar cases. In statistical terms, this distinction can be defined as across-judge versus within-judge variation. Variation across judges is also referred as variation in judge *fixed effects*.

There are similarities to empirical science, where researchers analyze samples to test hypotheses. There is within-researcher variation due to sampling error. Re-sampling (or bootstrapping) yields different values of the estimator. The standard deviation (SD) of this distribution is referred to as *standard error* (SE) (Yule, 1897). It is a source of uncertainty that researchers are well aware of when conducting their tests.

Researchers are less aware that there is an additional level of uncertainty due to there not being a *standard* analysis path. Researchers vary in what they deem to be the most reasonable path in the "garden of forking paths" (Gelman and Loken, 2014). Conditional on the path, there is a well-defined estimator and standard error. Conditional on *the sample*, however, estimates may vary across researchers as they might pick different paths.[1] We refer to this additional variation as *non-standard error* (NSE). Note that the adjective, *non-standard*, emphasizes the lack of a standard approach. In other words, if all researchers agree on one path being the most reasonable one, then NSE is zero.

The schema below summarizes the overarching idea of non-standard errors. Statisticians use the term data-generating process (DGP) to convey the idea that samples are random draws from a population. Estimators, therefore, exhibit standard error.

[1] An important source of such variation is that researchers need to translate *conceptual research questions* to *empirical research questions* (Breznau et al., 2022).



Using the same language, one could say that scientists collectively engage in an evidence-generating process (EGP). Researchers potentially pick different analysis paths, which is a source of additional error: Non-standard error. Note that error in this case is to be understood as erratic as opposed to erroneous, in the sense that there simply is no right path in an absolute sense.[2]

Let us illustrate the idea with an example. In microstructure, market efficiency is conceptually defined as the extent to which a price process resembles a random walk. Suppose that one is interested in estimating the trend in market efficiency. To estimate, say, the mean annual change in market efficiency, a researcher faces many forks in the road: How to measure market efficiency, at what frequency to sample the data, how to define outliers, etc. Collectively, we refer to these decisions as the analysis path.

Our objective is to measure and analyze non-standard errors. The four questions that we focus on are:

1. How large are non-standard errors in finance?

2. Can they be "explained" in the cross-section of researchers? Are they smaller

   (a) for papers by higher quality teams?

   (b) for papers with better reproducible results?

   (c) for papers that score higher in peer evaluations?

[2] Variation in estimates reported in meta studies is of both types. The polar cases are the following. Estimates vary because researchers did the analysis in the exact same way, but on different samples (SE). Or, estimates vary because the sample is the same, but the analysis differs (NSE). Mavroeidis, Plagborg-Møller, and Stock (2014) is a special case, because they conduct their meta study by applying all observed analysis paths on all samples. They, unlike us, do not focus on distinguishing the two sources of variation explicitly. For a review of meta studies in finance, see Geyer-Klingeberg, Hang, and Rathgeber (2020).

3. Does peer feedback reduce non-standard errors?

4. Are researchers accurately aware of the size of non-standard errors?

The motivation for these questions is that non-standard errors are undesirable in the sense that they add uncertainty. Such uncertainty becomes particularly worrisome when some estimates are positive, while others are negative. It is reminiscent of the negative result known as the Sonnenschein-Mandel-Debreu "anything goes" theorem (Mas-Colell, Whinston, and Green, 1995, Ch. 17-E). We therefore want to learn if higher quality coincides with tighter NSEs, and if feedback reduces NSEs.

Finding answers to the four questions is extremely costly in terms of human resources. The core structure of an ideal experiment involves two sizable sets of representative researchers. A first set of researchers independently tests the same hypotheses on the same data, and writes a short paper presenting the results. A second, non-overlapping set of researchers obtains these papers, evaluates them, and provides feedback in a single-blind process.

We have run such an experiment under the #fincap tag (FINance Crowd Analysis Project). 164 research teams (RTs) and 34 peer evaluators (PEs) participated, with each PE evaluating about ten papers. The Deutsche Börse kindly made proprietary data available spanning 17 years of trading in Europe's most actively traded instrument: the EuroStoxx 50 index futures. This data enabled researchers to test pre-defined RT-hypotheses[3] on several important market trends. This unique opportunity might explain why participation was exceptionally high (at least double that of similar experiments elsewhere, discussed later in the introduction).[4] A back-of-the-envelope calculation shows that total human resources for #fincap span almost a single academic career: (164×2 months + 34×2 days $\approx$ 27 years).

Statistical framework. We define non-standard error as the interquartile-range (IQR) in estimates across researchers. The reason

[3] We refer to these hypotheses as RT-hypotheses to distinguish them from the hypotheses that we test when analyzing the #fincap results. Our hypotheses are based on the four overarching questions (Section 2.2).

[4] #fincap was presented to all involved by means of a dedicated website (https://fincap.academy) and a short video (https://youtu.be/HPtnus0Yu-o).

for picking a *robust* dispersion measure instead of SD, is that this distribution could exhibit fat tails, and thus be prone to outliers. #fincap itself is a case in point as will become clear. The distribution of estimates across researchers tends to the distribution of researcher fixed effects (RFEs), which could be any distribution. Using a robust dispersion measure, therefore, is a prudent choice.[5]

Statistical inference in #fincap needs to account for multiple hypothesis testing (MHT) (Bonferroni, 1936; Šidák, 1967). The critical values for individual tests need to account for multiple teams testing the same hypothesis. Put simply, if individual tests are performed at a five percent level, then the probability of at least one turning significant for multiple tests, (weakly) exceeds five percent. Harvey, Liu, and Zhu (2016) illustrate how to adjust levels in asset pricing tests. In his presidential address, Harvey (2017) emphasizes that MHT affects all of finance. We follow in his footsteps when applying MHT in #fincap.

Finally, to address the overarching questions, we need to analyze how NSEs co-vary with quality measures, and how they change across stages. Since NSE is defined in terms of quantiles, we will use quantile regression to conduct this analysis (Koenker and Bassett Jr., 1978). Note that ordinary least-squares only models conditional means, and it is therefore unfit for an analysis of dispersion. In addition to the first and the third quartile, we will also model the median, the first decile, and the ninth decile, in order to obtain a more complete view of the distribution, including results on the inter-decile range (IDR).

Summary of our findings. We first show that the group of #fincap participants is representative of the academic community in empirical finance/liquidity. About a third of the 164 research teams have at least one member with publications in the top-three finance, or the top-five

[5]The intuition is as follows. If the number of researchers tends to infinity, then the distribution of estimates tends to the distribution of RFEs, plus sampling errors. If, in addition, the sample size tends to infinity, then the distribution of estimates tends to the distribution of RFEs (because, for each analysis path, the group mean for this path tends to the RFE associated with this path). This distribution can be any distribution and might, therefore, exhibit fat tails. Section 2.3 provides a statistical framework.

economics journals.[6] For the group of peer evaluators, this share is 85%. 52% of RTs consist of at least one associate or full professor. For PEs, this is 88%. On a scale from 1 (low) to 10, the average self-ranked score on experience with empirical finance is 8.1 for RTs, and 8.4 for PEs. For experience with market liquidity, it is 6.9 for RTs, and 7.8 for PEs.

The evidence on the four overarching questions is as follows. First, the dispersion in estimates across RTs is sizable. All six RT-hypotheses had to be tested by proposing a measure and computing the average per-year percentage change. The first RT-hypothesis, for example, was "Market efficiency has not changed over time." The median estimate across RTs is -1.1% with a non-standard error (IQR) of 6.7%. The IDR is 27.3%.[7] The dispersion for the other RT-hypotheses is similar in magnitude, albeit smaller for RT-hypotheses that arguably involve fewer decisions on the analysis path (e.g., testing for a trend in market share).

Statistical tests show that, for all RT-hypotheses, at least a few estimates are significant (at a family level of 0.5%).[8] This number ranges from 6 (out of 164) for RT-H6 to 125 for RT-H3. We further test the null hypothesis of no dispersion in researcher fixed effects. We reject it for all RT-hypotheses. NSEs are therefore statistically significant for all RT-hypotheses.

Finally, it is worth noting that the uncertainty due to non-standard errors is similar in magnitude to that due to standard errors. For RT-H1, for example, the median standard error across RTs is 2.5%. For a Gaussian distribution, this implies an IQR of $1.35 \times 2.5\% = 3.4\%$, which compares to an NSE of 6.7%.

Second, the quantile regressions show that higher quality tends to coincide with smaller NSEs. A one SD increase in reproducibility significantly reduces NSEs by 25.0% and a one SD increase in

[6]Finance: *Journal of Finance*, *Journal of Financial Economics*, and *Review of Financial Studies*. Economics: *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*.

[7]This RT-hypothesis further illustrates the importance of robust statistics. One RT reports an estimate of +74,491%. This extreme outlier causes the mean and standard deviation to be 446.3% and 5,817.5%, respectively.

[8]We use the conservative significance levels advocated by Benjamin et al. (2018): 0.5% for significance and 5% for weak significance. They refer to the latter as "suggestive evidence."

peer-evaluator rating significantly reduces them by 33.3%. A one SD increase in team quality, however, significantly raises NSEs by 2.8%. This effect, however, is small in economic magnitude. If IDR were used instead of IQR, then a one SD increase significantly reduces IDRs for all quality measures: 13.3%, 17.9%, and 11.9%, respectively. Overall, higher quality seems to make extreme values less likely.

Third, peer feedback significantly reduces non-standard errors. The peer-feedback process involves multiple stages. We find that each stage reduces NSEs, albeit insignificantly. The reduction across *all* four stages is significant and amounts to 47.2%. This number for IDRs is also significant, and amounts to an even larger decline: 68.2%.

Fourth, RTs mostly underestimate the dispersion in estimates across RTs, which we tested in an incentivized belief survey. Such underestimation might well be the reason why non-standard errors never attracted much attention, until recently.

Finally, we dig deeper to discover what drives dispersion in estimates. A particularly useful tool for such analysis is a multiverse analysis (Liu et al., 2021). For key forks on the analysis path, the multiverse reveals how sensitive the distribution of estimates is to decisions at each particular fork.

It turns out that many of the key forks in #fincap add substantial noise. For RT-H1 on market efficiency, for example, it matters which frequencies teams choose for their variance ratio calculations. Some teams compare seconds to minutes, others days to months. A comparison of higher frequencies tends to find a decline in market efficiency, whereas for lower frequencies some find an increase in market efficiency.

The multiverse further reveals that Jensen's inequality can cause large dispersion. If a researcher is interested in assessing an N-period (long-term) trend in X_t, and estimates it based on one-period observations, then this could add substantial noise (Blume, 1974). Consider, for example, the expectation of a product of two independent and identically distributed relatives, where a relative is defined as X_t/X_{t-1}. Jensen's inequality implies that the expectation of this product is larger than the product of the expected relatives. The multiverse shows that the noise this adds can become particularly large for teams who sample at a daily frequency to estimate an annual trend, and use

relatives instead of, for example, log-differences or a trend-stationary approach.

Contribution to the literature. The issue of variability in the research process is not new. Leamer (1983), for example, was troubled by the "fumes which leak from our computing centers." He called for studying "fragility in a much more systematic way."

Replication studies echo his concern as they typically find much weaker effects and less statistical strength (Ioannidis, 2005, Open Science Collaboration, 2015, Camerer et al., 2016, 2018). This is potentially the result of *p*-hacking: the process by which researchers try analysis paths until non-significant results turn significant.[9] We caution, however, that poor replication could also be demand-driven instead of supply-driven. This is the case when journals prefer to publish papers with low *p*-values. Munafò et al. (2017) survey the various threats to credible empirical science and propose several fixes.

The literature on replicability in finance is young, but growing rapidly. Examples are: McLean and Pontiff (2016), Hou, Xue, and Zhang (2018), Linnainmaa and Roberts (2018), Chordia, Goyal, and Saretto (2020), Harvey and Liu (2020), Ben-David, Franzoni, and Moussawi (2021), Black et al. (2021), Chen (2021), Mitton (2021), and Jensen, Kelly, and Pedersen (2022).

None of these replication studies focus on explaining the dispersion of estimates in a cross-section of researchers, or study the impact of peer feedback. We are the first to run an experiment, where this can be done in a clean way. Our objective is to study dispersion in estimates, short of a potential bias due to *p*-hacking. By design, there is no need to *p*-hack for #fincap researchers, because anyone who completes all stages of the project had been guaranteed co-authorship. Similarly, peer evaluators are guaranteed co-authorship to ensure clean feedback.

We are the first in finance to run an experiment to study dispersion in estimates, but we are not the first in science. Silberzahn et al. (2018) pioneered the multi-analyst study by letting multiple teams test whether soccer referees are more likely to draw red cards for

[9] The *p*-value is the probability of observing an effect that is at least as large as the estimated effect, under the null hypothesis that there is no effect.

players with a darker skin color. Other examples are Botvinik-Nezer et al. (2020) for neuroscience, Huntington-Klein et al. (2021) for economics, and Breznau et al. (2021) and Schweinsberg et al. (2021) for sociology. We innovate relative to these studies by explaining dispersion in estimates with quality attributes, by adding peer feedback stages, and by soliciting beliefs on dispersion *ex-ante*. A further strength of our study is the large cross-section of research teams: N=164. It is more than twice the size of any of the other multi-analyst samples.

The remainder of the paper is organized as follows. Section 2 provides an in-depth discussion of the project design.[10] It further presents the hypotheses associated with the four overarching questions, and develops an appropriate statistical framework to test them. Section 3 presents our results. Section 4 concludes.

2 Project design and hypotheses

This section first presents the details of the #fincap experiment, then presents hypotheses based on the four overarching questions, and finishes by discussing an appropriate statistical framework.

2.1 Project design

Before starting the #fincap experiment, we had filed a pre-analysis plan (PAP) with the Open Science Foundation (https://osf.io/h82aj/). The original version of *Non-Standard Errors* contains the results of the analysis outlined in the PAP. This original version remains available as Tinbergen Institute Discussion Paper TI 2021-102/IV. Subsequent feedback from various presentations and from reviewers at the *Journal of Finance* have led to the results presented here. Relative to the PAP, we now use robust methods to cope with unanticipated extreme outliers, we account for multiple testing, and we add a multiverse analysis to add deeper insight. Appendix A reconciles the current results with those in the original version.

In a nutshell, the #fincap experiment is about multiple research teams independently testing the same hypotheses on the same sam-

[10] The design of #fincap follows the guidelines for multi-analyst studies proposed by Aczel et al. (2021).

ple. We refer to these hypotheses as RT-hypotheses and to this sample as RT-sample. This is to distinguish them from the hypotheses that *we* will test based on the results generated by RTs and PEs (Section 2.2).[11]

The RT-sample is a plain-vanilla trade sample for the EuroStoxx 50 index futures with, added to it, a principal-agent flag.[12] For each side to a trade (i.e., buy and sell), we therefore know whether the exchange members traded for their own account, or for a client. The sample runs from 2002 through 2018 and contains 720 million trade records. These index futures are among the world's most actively traded index derivatives. They give investors exposure to Europe, or, more precisely, to a basket of euro-area blue-chip equities. With the exception of over-the-counter activity, all trading is done through an electronic limit-order book (see, e.g., Parlour and Seppi, 2008, for details on limit-order book markets).

The RT-hypotheses are all statements about annual trends in the following market characteristics (with the null being no change):

RT-H1 market efficiency

RT-H2 realized bid-ask spread,

RT-H3 share of client volume in total volume,

RT-H4 realized bid-ask spread on client orders,

RT-H5 share of market orders in all client orders, and

RT-H6 gross trading revenue of clients.

[11] RTs and PEs have been recruited mostly by alerting appropriate candidates through suitable channels (e.g., the https://microstructure.exchange/). To inform them about #fincap, we created an online repository: https://fincap.academy. The repository remains largely unaltered (except for, e.g., adding FAQs).

[12] Trade records contain the following fields: Datetime, expiration, buy-sell indicator, size, price, aggressor flag, principal-agent flag, and a full- or partial-execution flag. Note that each side to a trade becomes a record, where the aggressor is the side whose incoming, say, buy order is matched with a resting sell order of the other side. The record is labeled *principal* if the exchange member trades for his own account, and *agent* when he trades for a client. More details on the sample are in Figure OA.6 of the Online Appendix.

The RT-hypotheses are presented only briefly here to conserve space. The full presentation of RT-H1, for example, characterizes informationally efficient prices as a random walk. Appendix B motivates and discusses all RT-hypotheses in detail. For the purpose of our analysis, we like to highlight two points. First, the RT-hypotheses are picked to address first-order questions in the field of empirical-finance/liquidity. These questions were used to market #fincap and convince appropriate candidates to join the project. Second, we ask for trends expressed as percentage changes to make them invariant to choice of unit (e.g., are measures expressed in thousands, or not).

Note that there is, purposefully, considerable variation across RT-hypotheses in the level of abstraction. RT-H1, for example, is on the relatively abstract notion of market efficiency. RT-H3, on the other hand, is on the share of client volume in total volume. Such share should be relatively straightforward to calculate because, in the RT-sample, each buy and sell trade is flagged agent (client) or principal (proprietary).

RTs are asked to test these RT-hypotheses by estimating an average yearly change for a self-proposed measure.[13] They are further asked to report standard errors for these estimates. We compute the ratio of the two, which we refer to as the implied t-value, or t-value for short.

RTs write a short academic paper in which they present and discuss their findings. These papers are evaluated by PEs who were recruited outside the set of researchers who registered as RTs. RT papers were randomly and evenly assigned to PEs in such a way that each paper is evaluated twice, and each PE evaluates nine or ten papers. PEs score the papers by providing an overall rating and a rating per RT-hypothesis. They do so in a single-blind process: PEs see the names of RTs, but not vice versa.[14] The reason for single-blind instead of double-blind is to incentivize RTs to exercise maximum effort.

PEs are asked to motivate their scores in a feedback form where

[13] RTs are asked to express their results in annualized terms. To some, it was not clear. We therefore notified everyone of the following clarification that we added to the FAQ section on https://fincap.academy: "Research teams are asked to report annualized estimates (and the corresponding standard errors); research teams are not required, however, to consider only annualized data."

[14] In our analysis, we remove PE fixed effects by demeaning (see Section 2.2).

they are encouraged to add constructive feedback. RTs receive this feedback unabridged, and are allowed to update their results based on it. Importantly, the design of #fincap was common knowledge to all because it had been available on a dedicated website before registration opened (see footnote 4).

More specifically, #fincap consists of the following four stages:

Stage 1 (January 11 - March 23, 2021.) RTs receive the detailed instructions along with access to the RT-sample. They conduct their analysis and hand in their results (short paper plus code). We emphasized in our emails and on the project website that RTs should work in *absolute secrecy* so as to ensure independence across RTs.

Stage 2 (May 10 - May 28, 2021.) RTs receive feedback from two anonymous PEs and are allowed to update their analysis based on it. They are asked to report their findings in the same way they did in stage 1.

Stage 3 (May 31 - June 18, 2021.) RTs receive the five best papers based on the average raw PE score. The names of the authors of these five papers were removed before distributing the papers.[15] Similar to stage 2, all RTs are allowed to update their analysis and resubmit their results.

Stage 4 (June 20 - June 28, 2021.) RTs report their final results, this time not constrained by delivering code that produces them. In other words, RTs are allowed to Bayesian update their results (i.e., estimates and standard errors) taking in all the information that has become available to them, in particular the five best papers. They could, for example, echo the results of one of these papers, simply because of an econometric approach that they believe is superior but that is beyond their capacity to code. This stage was added to remove all constraints and see how far the RT community can get in terms of reaching consensus.

[15] If two papers were tied in terms of their average score, then, following the pre-analysis plan, we picked the one that had highest reproducibility score provided by Cascad. For more information on Cascad, see the statement of H2 in Section 2.2.

The stages subsequent to the first one mimic the feedback researchers get from various interactions with peer researchers in the research process *before* a first journal submission. Stage 2 mimics, for example, immediate feedback from colleagues over lunch, during seminars, or in coffee breaks at conferences. Stage 3 mimics indirect feedback by means of seeing competitive papers that gain a lot of visibility through endorsements by colleagues, or by being presented in seminars or at conferences. Stage 4 solicits a final estimate whereby researchers are allowed to attach weight to estimates of others whom, for example, they believe implement a superior methodology that they are unable to code themselves. We like to emphasize that all these stages are designed in a way to keep the full dynamics of a refereeing process at a scientific journal out of scope.[16]

2.2 Hypotheses

Before running the experiment, we translated the project's four overarching questions into a set of pre-registered hypotheses. These hypotheses all center on the dispersion in estimates across RTs. Our main measure is the interquartile range, which we refer to as nonstandard error. All hypotheses are stated as null hypotheses and tests will be two-sided.

The first set of three hypotheses focuses on how NSEs relate to various quality measures:

H1 NSE of stage-1 estimates does not co-vary with team quality. Team quality is proxied by the largest common factor in various candidate proxies for team quality. We prefer an appropriately weighted average over simply adding all proxies to maximize statistical power in the regressions. More specifically, we define team quality as the first principal component of the following standardized series:[17]

[16] Studying such dynamics requires a different experiment that involves "publishing" papers, *including* the names of the authors. Note that we do reveal the best five papers (according to PEs) to all RTs in stage 4, but the authors of these papers remain hidden. Our focus is narrowly on the pure findings and beliefs of the RTs, avoiding any possible corruption by "the publication game."

[17] An important advantage of a principal-component analysis (PCA) is that the weighting is data-driven, thus avoiding subjective weights. Note that even the five proxies that enter were picked ex-ante in the pre-analysis plan filed at OSF. The

(a) *Top publications*: The team has at least one top-three publication in finance or one top-five publication in economics (0/1) (see footnote 6).

(b) *Expertise in the field*: Average of self-assessed experience in market liquidity and empirical finance (scale from 0 to 10).

(c) *Experience with big data*: The team has worked with samples at least as large as the sample they analyze in #fincap (0/1).

(d) *Academic seniority*: At least one team member holds an associate or a full professorship (0/1).

(e) *Team size*: The team size attains its maximum of two members (0/1).

H2 NSE of stage-1 estimates does not co-vary with reproducibility score. This score measures the extent to which RT estimates are reproducible from RT code. The scoring was done by the Certification Agency for Scientific Code and Data (Cascad). Cascad is a non-profit certification agency created by academics with the support of the French National Science Foundation (CNRS) and a consortium of French research institutes. The objective of Cascad is to provide researchers with a way to credibly signal the reproducibility of their research (used by, for example, the *American Economic Review*).[18]

H3 NSE of stage-1 estimates does co-vary with the average PE rating (RT-hypothesis level). To remove a possible PE fixed effect, we use demeaned PE ratings in all of our analysis.

The next hypothesis is about convergence in estimates across the four stages.

H4 NSE does not change across all feedback stages.

PCA results will be discussed in Section 3.2.1.

[18]Cascad rates reproducibility on a five-category scale: RRR (perfectly reproducible), RR (practically perfect), R (minor discrepancies), D (potentially serious discrepancies), and DD (serious discrepancies). For #fincap, Cascad converted their standard categorical rating to an equal-distance numeric one: RRR, RR, R, D, and DD become 100, 75, 50, 25, 0, respectively.

The final hypothesis focuses on RT *beliefs* about the dispersion in estimates across RTs.

H5 The average belief of RTs on the dispersion in estimates across RTs, is correct. The dispersion predictions were solicited in terms of the SD measure.

2.3 Statistical framework

To formalize the analysis of non-standard errors in a statistical sense, consider a set of researchers indexed by $j \in \{1, \ldots, J\}$. All researchers are given the same sample of size K. Researchers are asked to estimate the mean of a particular object (e.g., the per-year change in market efficiency). All researchers independently decide on the optimal analysis path and estimate the mean accordingly. Collectively, let these estimates, $X_1, \ldots, X_J$, be distributed as:

$$X_j = e_j + \varepsilon_j, \tag{1}$$

where e_j is a researcher-specific mean, henceforth referred to as a researcher fixed effect (RFE), and ε_j is a sampling error. The Central Limit Theorem (CLT) implies that, for large K, ε_j is approximately normal with mean zero and variance $\sigma^2_{j,K} = \sigma^2_j/K$, where σ^2_j is the path-specific variance of residuals.

Note that sampling errors are likely to correlate across researchers so that, collectively, the estimates are approximately distributed as:

$$\underset{(J\times 1)}{X} = \underset{(J\times 1)}{e} + \underset{(J\times 1)}{\varepsilon}, \text{ where } \underset{(J\times 1)}{\varepsilon} \sim N\left(\underset{(J\times 1)}{0}, \underset{(J\times J)}{\Sigma}\right), \tag{2}$$

where Σ is a positive semidefinite matrix. The off-diagonal elements of Σ are expected to be mostly positive since, if for a particular sample draw, X_i is above its (unconditional) mean e_i, then X_j is, most likely, also above its mean e_i.[19]

[19]For example, consider the case of estimating the mean of a distribution. If two researchers estimate this mean by taking the sample average, but one winsorizes the sample and the other does not, then a particular sample draw with unusually high values likely yields above-mean estimates for both researchers.

Non-standard error. Non-standard error is defined as the interquartile range in estimates:

$$\text{NSE} := Q_{0.75}(x) - Q_{0.25}(x), \tag{3}$$

where x denotes a realization of the random vector X, and $Q_\alpha(x)$ is the αth quantile of x. Note that NSE tends to the IQR of RFEs when J and K both tend to infinity:

$$\text{NSE} \overset{J,K\to\infty}{\to} Q_{0.75}(e) - Q_{0.25}(e). \tag{4}$$

We reiterate that for the distribution of RFEs (i.e., the distribution of e) could be any distribution. It is, therefore, prudent to pick a robust dispersion measure, which is why we use IQR instead of SD. The latter tends to get dominated by the size of extreme outliers.[20]

Testing for non-standard error. We test for "significance of non-standard errors" by testing whether or not there is any dispersion in RFEs. We do this by testing the following set of null hypotheses:

$$H_0 : e_j = \nu, \quad \forall j \in \{1, \ldots, J\}, \tag{5}$$

where ν is the median RFE. Since X_j is an estimator of e_j, these hypotheses can be tested by verifying, for each $j \in \{1, \ldots, J\}$, whether X_j is statistically different from ν. In the implementation, we set ν equal to the median estimate. If any of these tests rejects the null, then dispersion is non-zero, and we consider non-standard errors to be statistically significant.[21]

[20] #fincap is a case in point. For RT-H4, one team reports an estimate of -6,275,383%, whereas the estimates of other teams range from -2,897% to 870%. The SD based on all estimates is 490,024%, but it is only 245% if one leaves out the outlier.

[21] Two more technical points merit discussion. First, we prefer the median over the mean to have a robust location parameter. The asymptotic variance of the mean is smaller than that of the median for Gaussian distributions, but typically not for distributions with fat tails. The reason is that the former depends on variance and thus on extreme outliers, whereas the latter does not: σ^2/N and $1/(4Nf(m))$, respectively, where N is the sample size, f is the density function, and m is the median. Figure OA.1 in the Online Appendix shows that, in #fincap, the variance of the median is an order of magnitude smaller than the variance of the mean. Second, the proposed test assumes that sampling error is negligible for the median estimate as an estimator for the median RFE, because randomness in the median estimate is ignored. Figure OA.1 illustrates that, indeed, the variance of the median estimate is negligible for #fincap.

Conceptually, the distribution of X could be obtained by bootstrapping. Such procedure, however, is infeasible because it requires that researchers redo their analysis for every new draw of the sample. Instead, we use multiple hypothesis testing (MHT) results to develop a feasible testing procedure.

Before turning to MHT, let us pause for a moment and take stock of what is available to us. The #fincap sample consists of estimates x_j, along with their standard error s_j. This is useful, but misses information on the covariance among all possible pairs of estimates across researchers.

To account for multiple testing, we rely on well developed statistical theory. If one aims to test at a level of 5% for a family of N tests, then individual tests should be performed with a $(5/N)\%$ critical value, if the test statistics are mutually independent (Bonferroni, 1936; Šidák, 1967; Harvey, Liu, and Zhu, 2016).[22]

In summary, we propose an NSE test where the null hypothesis is that there is no dispersion in RFEs. We use a Bonferroni adjustment of significance levels to account for multiple testing. The test is conservative, because Bonferroni assumes independence. As pointed out in footnote 19, estimates are likely to correlate across researchers, in which case the *effective* number of tests is likely to be smaller than the actual number of tests. In the implementation, we add a trivial extension where correlations between estimates are calibrated based on our multiverse analysis (Section 3.3). We close the section by discussing an alternative test and pointing out a caveat.

Alternative test. Note that a natural alternative to the proposed test is to simply test if IQR is statistically different from zero. We did not pick this shortcut, because our focus is on whether there is any dispersion at all in estimates across researchers. Although we pick IQR to express dispersion in a single number, the deeper interest is whether the distribution in estimates is non-degenerate.

[22] If the N tests statistics are independent, then the probability of *at least* one significant result is $(1 - (1 - \alpha)^N)$. For example, for $\alpha = 0.05$ and $N = 10$, this probability is 40 percent. Šidák (1967) proposes to adjust the significance level for the individual tests to $\alpha' = 1 - (1 - \alpha)^{1/N}$. A Taylor expansion of α' around zero yields $\alpha' \approx \alpha/N$, which is known as the Bonferroni correction (Bonferroni, 1936).

Caveat. We like to point out one potential caveat. The procedure to obtain a conservative test on RFEs implicitly assumes that SEs reported by researchers are consistent estimators of the true SEs. This might not be true if (some) researchers report non-robust SEs. Non-robust SEs tend to be smaller, because they ignore commonalities. If true, then NSE tests tend to turn significant more often. NSEs themselves, however, remain consistent estimators.[23]

3 Results

This section presents all our findings. They are based on a balanced sample of 164 research teams who completed all stages of the project (out of 168 research teams). The first subsection presents various summary statistics and tests whether non-standard errors are statistically significant. The second subsection tests our hypotheses. The third subsection digs deeper by means of a multiverse analysis. The fourth and final subsection discusses alternative explanations.

3.1 Summary statistics

(Insert Table 1 about here.)

Table 1 summarizes our stage-1 sample by means of three sets of statistics, organized in three panels.[24] Panel (a) summarizes the qualities of the #fincap community. It consists of 164 research teams and 34 peer evaluators. Maximum RT size is two members, which is the size of 79% of RTs.

The statistics testify to the high quality of the #fincap community. 31% of RTs have at least one top publication in finance or economics (see footnote 6 for the list of journals). For PEs, this is 85%. The percentage of RTs who have at least one member who is tenured at the associate or full professor level is 52% for RTs. For PEs, this is

[23] Unfortunately, we do not have precise information on the SEs reported in #fincap, because not all RTs provide detailed information on how they calculate SEs.

[24] Table OA.1 through OA.3 in the Online Appendix repeat panel (c) of Table 1 for the other stages. Panel (a) is the same for all stages, and panel (b) is only available for stage-1 results, since only these results are evaluated by peers and scored by Cascad on reproducibility.

88%. Feedback seems to come from more established scholars, which likely mirrors reality.

RT members and PEs cover the global academic-finance community reasonably well (see Figure OA.2 in the Online Appendix). RT members reside in 34 countries with most residing in the US (51 out of 293). PEs reside in 13 countries with, again, most residing in the US (13 out of 34). The strong skew towards the US is not surprising given that the more senior, well-published finance scholars are predominantly affiliated with US universities.

Most RTs and PEs seem to have the appropriate background for testing the RT-hypotheses on the RT-sample. Their average self-reported scores on having experience in the field of empirical finance is 8.1 for RTs and 8.4 for PEs on a scale from 0 (low) to 10. For experience with market liquidity, these average scores are 6.9 for RTs and 7.8 for PEs. There is considerable variation around these averages as the SDs range from 1.7 to 2.4. When it comes to working with samples as large as the RT-sample, 720 million trade records, most RTs and PEs seem up to it. 65% of RTs have worked with samples at least as large. For PEs, this percentage is 88%.

Panel (b) of Table 1 shows that the average quality of the RT analysis is solid, but the dispersion is large. The average reproducibility score is 64.5 on a scale from 0 (low) to 100 (see footnote 18). This is high when benchmarked against other studies on reproducibility (Colliard, Hurlin, and Pérignon, 2021). The accompanying SD is 43.7, which shows that there is large variation across RTs: Most code either reproduces close to perfectly or barely at all. The paper-quality scores provided by PEs show a similar pattern, albeit with less dispersion. The average score across RTs is 6.2 on a scale from 0 (low) to 10, with an SD of 2.0.

Panel (c) provides descriptive statistics on the distribution of results across RTs. It does so by RT-hypothesis, and by type of result: Estimate, standard error, and t-value. Since our focus is on dispersion in estimates across RTs, we relegate a discussion of RT medians to Appendix B. More specifically, this appendix discusses the RT-hypotheses in-depth and summarizes what RTs, as a group, seem to find with a focus on the across-RT median instead of the across-RT IQR (i.e., the non-standard error).

(Insert Figure 1 about here.)

Perhaps the most salient feature of the extensive panel (c) is that there is substantial variation across RTs for all RT-hypotheses, and for all types of results. Panel (a) in Figure 1 illustrates this result for estimates. For RT-H1 on market efficiency, for example, the median estimate across RTs is -1.1% with an IQR of 6.7%. Even for RT-H3, which is a seemingly straightforward calculation of a market share, the dispersion is sizable: an IQR of 1.2% around a median of -3.3%. The figure further illustrates that there are extreme outliers for all RT-hypotheses, which motivates our analysis in terms of robust statistics.

(Insert Table 2 about here.)

NSE test results. Is the dispersion in estimates statistically significant? Table 2 presents the non-standard error test results. The null of no dispersion in researcher fixed effects is rejected for all RT-hypotheses at a 0.5% (family) significance level. The conservative Bonferroni adjustment in panel (a) yields at least 11 estimates that are individually significantly different from the median (RT-H6), and at most 38 significant differences (RT-H3). There are significant estimates both above and below the median for all RT-hypotheses.

If, instead of assuming zero correlation across test statistics as in Bonferroni, one calibrates them based on bootstrapping from the multiverse analysis (Section 3.3), results change to the ones presented in panel (b). The implied "effective" number of tests is much lower than the 164 tests used in Bonferroni. It ranges from 21 (RT-H3) to 86 (RT-H6). The factor by which significance levels are adjusted is, therefore, up to almost seven times smaller than what Bonferroni suggests (i.e., 164/24=6.8). The result is that, indeed, more differences become significant. The increases are moderate, though, with at most two more differences becoming significant.

In sum, the statistics presented thus far show that there is substantial dispersion across research teams, in terms of their estimates, but also in team quality, in reproducibility score, and in peer-evaluator rating. In the next subsection, we use this dispersion to test the first three hypotheses. Is there more dispersion in estimates for lower quality teams, for results that are harder to reproduce, or for lower quality papers?

3.2 Hypotheses tests

The results on the three sets of hypotheses are discussed in the next three subsections. Standard errors in the quantile regressions account for correlation in residuals by adding RT-hypothesis fixed effects, and by clustering per RT across all stages.

3.2.1 Co-variates for stage-1 dispersion (H1-3)

The first set of hypotheses relates NSEs to various quality variables. One of these is team quality, which we measure by picking the first principal component (PC1) of five standardized quality proxies (see H1 in Section 2.2). PC1 explains 38.3% of total variance, and loads positively on all quality proxies. It loads strongest on publications and weakest on big-data but, importantly, it loads positively on all of them. Table OA.4 in the Online Appendix provides detailed results on the PCA.

(Insert Table 3 and Figure 2 about here.)

Table 3 summarizes the results of the stage-1 quantile regressions, with as dependent variables, the 10th, 25th, 50th, 75th, and 90th percentile of the distribution in estimates across RTs. Figure 2 illustrates the results by showing how a one SD increase in each co-variate affects IQR (i.e., NSE) and IDR. Taken together, these results allow us to test the first three hypotheses that relate quality variables to dispersion in estimates.

First, we find that higher team quality coincides with somewhat larger IQR, but with smaller IDR. The effect of team quality on the 25th percentile is not significant, but for the 75th percentile, it is significantly positive. The economic magnitude is small, though, as can be seen in Figure 2. A one SD increase in quality raises IQR by only $(0.032 - 0.004) \times 7.2 = 0.2$ percentage points (pps), where 7.2 is the average IQR across hypotheses (see panel (c) of Table 1). This increase of 0.2 pps implies a relative increase of 2.8%.[25] In contrast,

[25] A direct test on IQR, instead of separate tests on the 25th and 75th percentiles, requires jointly modeling these percentiles. Such multivariate modeling, combined with clustering on errors, is a non-trivial econometric challenge. Univariate modeling with clustering, on the other hand, is relatively standard. We use a python package to run these regressions: pyqreg.

a one SD increase in team quality, *reduces* IDR by 6.7 pps (-11.9%, since average IDR is 56.3). This is the result of a significant increase in the first decile and a significant reduction in the ninth decile. These findings suggest that higher quality teams are less likely to report extreme estimates.

If one replaces team quality by the five quality variables on which it is based, then a more nuanced picture emerges (Table OA.5 in the Online Appendix). The statistically significant and sizable relationships are the following. A one SD increase in academic seniority (i.e., an associate/full professor in the team), reduces IQR by 1.4 pps (-19.4%). A one SD increase in team size reduces it by 0.9 pps (-12.5%).

A one SD increase in top publications, however, *increases* IQR by 1.9 pps (+26.4%). These three variables are positively correlated which explains why we find that the (aggregate) team variable has a relatively small effect on IQR. For IDR, the effects are of the same sign, but larger in magnitude: -19.4, -7.0, and +6.1 pps , respectively (-34.4%, -12.4%, and +10.8%). Note that now the negative effects really dominate, which explains that IDR co-varies negatively with team quality. In sum, these findings suggest that well published scholars seem to disagree more, but such effect is offset by the presence of a senior scholar or a second team member.

Second, all percentiles co-vary significantly with reproducibility, except for the median. The 10th and the 25th percentile co-vary positively and the 75th and the 90th percentile co-vary negatively. The figure shows that these changes are sizable. A one SD increase in reproducibility reduces IQR by 1.8 pps (-25.0%) and IDR by 7.5 pps (-13.3%). In sum, better reproducibility lowers overall dispersion.

Third, the results for paper quality mirror those of reproducibility, albeit a bit stronger in magnitude. The 10th and 25th percentile co-vary significantly positively, the 75th and 90th percentile co-vary significantly negatively. A one SD increase in paper quality reduces IQR by 2.4 pps (-33.3%) and IDR by 13.6 pps (-17.9%). Higher rated papers exhibit less dispersion in estimates.

In summary, the evidence on the first three hypotheses is such that the null of no co-variation is rejected for all three. Generally, higher quality is associated with less dispersion in estimates.

3.2.2 Convergence across stages? (H4)

The analysis of first-stage results has shown that dispersion in estimates is sizable and statistically significant. Does peer feedback create convergence? In other words, does dispersion in estimates decline in the three subsequent stages where teams get feedback from peers. This is the focus of the fourth hypothesis.

(Insert Table 4 and Figure 3 about here.)

Table 4 presents the results of quantile regressions to explain the dispersion in estimates in all four stages (thus far, only stage 1 has been analyzed). To account for heterogeneity in dispersion across RT-hypotheses, the explanatory variables are stage dummies that are multiplied by stage-1 (estimate) IQR *per* RT-hypothesis. The coefficients, therefore, measure a stage effect, expressed in IQR units. Figure 3 presents the results graphically.

The evidence makes us reject the null hypothesis of no convergence across all stages. All changes across consecutive stages are positive for the 10th and 25th percentile, and negative for the 75th and 90th percentile. The majority, however, is insignificant. However, the *total* change across stages is significant for all these percentiles at the 5% level, and, for all but one at the 0.5% level. Taken together, these results show that there is significant convergence from the first to the last stage, but a decomposition across the various stages lacks significance.

Figure 3 illustrates that the convergence is sizable. Panel (a) shows that the total decline in IQR is 3.4 pps (-47.2%). The decline seems evenly distributed across the stages, although this decomposition is mostly insignificant. Panel (b) shows that the total decline in IDR is even larger: 38.4 pps (-68.2%). More than half of it seems to happen from the first to the second stage, where RTs receive anonymized feedback from two PEs. However, this result is only weakly significant, since only the increase in the first decile is weakly significant (i.e., at a 5% level, not at a 0.5% level).

3.2.3 Are RT-beliefs on dispersion in estimates accurate? (H5)

The fifth and final hypothesis focuses on whether RTs are accurately aware of the dispersion in estimates across teams. Beliefs have been

solicited in an incentivized way. All teams were asked to predict SDs in estimates across teams.[26] We randomly selected 20% of all RTs and paid each of them $300 if one of their predictions (randomly drawn) was within 50% of the realized SD. Details on the reward scheme are in the instruction sheet they received before reporting their beliefs (Figure OA.15 in the Online Appendix). The hypothesis pertains to stage-1 estimates, because beliefs are solicited for this stage only.

As H5 is stated in terms of the average belief being correct, testing it requires a test on the equality of means: the mean belief about SDs in estimates across teams, and the SDs of these estimates in the population. Let us define a test statistic D that measures the relative distance between beliefs and realizations:

$$D = \frac{1}{6n} \sum_{i,j} \left(\frac{BeliefOnSD_{ij} - RealizationOfSD_j}{RealizationOfSD_j} \right), \tag{6}$$

where $BeliefOnSD_{ij}$ is the belief of team i on the SD in estimates across teams for RT-hypothesis j and $RealizationOfSD_j$ is the realized SD for this RT-hypothesis in the raw sample.[27] The distribution of D under the null of equal means is obtained by bootstrapping. For details on the bootstrap procedure, we refer to Appendix D.

(Insert Figure 4 about here.)

Figure 4 plots the distribution of beliefs on SDs, along with realized SDs depicted by red dots. It illustrates that the vast majority of teams underestimate dispersion in estimates. The interquartile range denoted by the boxes is consistently below the red dot, which implies that at least 75% of the teams underestimate the dispersion.

One might think that teams simply overlook the extreme values that make realized SDs explode. This, however, does not seem to be

[26] In retrospect, we should have (also) asked for an IQR prediction, because SD is very sensitive to extreme outliers (see footnote 20). To assess whether RTs might have overlooked such outliers, we will compare their SD predictions with realized SDs, both on the full sample and on a trimmed sample.

[27] The benefit of a relative measure as opposed to an absolute one is that (i) it is easy to interpret as it allows for statements of RTs over- or underestimating by some percentage and (ii) it accounts for level differences across hypotheses (e.g., under the null of accurate beliefs, a uniform distribution of beliefs on the support 0.09 to 0.11 will exhibit the same dispersion as a uniform distribution of beliefs on 900 to 1100).

the case, because even if one trims the estimates by removing the top and bottom 2.5%, the IQR box stays below these "trimmed" realized SDs, depicted by orange dots. The only exception is RT-H3, for which the orange dot is just within the top of the box.

The formal test results are in Table OA.6 of the Online Appendix. Pooling across all RT-hypothesis, the test statistic shows that the predicted SD is 71.7% below the realized SD. This underestimation is significant at a 0.5% level. Similar results holds for all RT-hypotheses individually, except for RT-H3, for which the underestimation is insignificant. Its value was also lowest of all, only 9.0% underestimation. RT-H3 is an hypothesis on market shares that, arguably, is relatively straightforward to test. In summary, the vast majority of tests show significant underestimation and we therefore firmly reject the null that beliefs on the dispersion in estimates are accurate.

3.3 Digging deeper: A multiverse analysis

Non-standard errors in #fincap are significant and sizable. Why? Can we somehow identify which forks on the analysis paths cause most of the dispersion? More specifically, can we rank key forks on the path according to the degree of refraction they cause in the light the sample sheds on the research question at hand? We turn to a multiverse analysis to address these questions.

Steegen et al. (2016) coined the term multiverse analysis to emphasize that data *construction* involves multiple decisions. The sample that enters the analysis, therefore, is a function of the set of reasonable choices. The sample becomes a (p. 702) "many worlds or *multiverse* of data sets." A particular result of an analysis then becomes a distribution of results (because samples vary). We generalize this approach by adding decision forks for the part of the analysis that follows the sample construction (e.g., the choice of econometric model).

The strength of a multiverse analysis is that it reveals how sensitive an estimate is to a particular fork on the analysis path. It does so by studying how much the estimates refract when varying across all reasonable alternatives at the fork. For example, let there be N reasonable analysis paths. Now suppose there are $k \leq N$ reasonable alternatives at the jth fork. Then the N estimates associated with the N paths are sorted into k sets, depending on the alternative picked at

the fork. The degree to which the results differ across the k sets determines how sensitive results are to the jth fork. We measure the degree to which k distributions differ by a k-sample Anderson-Darlin (AD) test. Appendix E discusses the AD test in detail, including why it fits our application particularly well. AD is a standard option in the Boba software that we use (Liu et al., 2021).[28]

(Insert Table 5 about here.)

To make the multiverse feasible, we identify key forks on the analysis path and, for each fork, we ask RTs to select the alternative they picked among a set of pre-defined alternatives. This was done by means of a questionnaire that all filled out after the experiment. The choice of forks and the alternatives at each fork is informed by the short papers RTs wrote for #fincap. The discretization of the decision space enables us to project the large space of realized analysis paths, onto a manageable space of "representative" paths. Table 5 provides an overview of all forks for the six RT-hypotheses. It lists the alternatives at each fork, along with the fraction of RTs that picked them (depicted in Figure OA.5 of the Online Appendix).

For each fork, we also asked RTs to rate the fit between the alternative they picked from the set, and what they actually did in #fincap. Their average rating ranges between 4.0 for RT-H6 and 4.4 for RT-H3 on a scale from 1 "Far from what we did" to 5 "Very close to what we did" (see Figure OA.4 in the Online Appendix). We, therefore, believe that the multiverse analysis is representative of the #fincap analysis itself.

A multiverse analysis is powerful, but resource intensive. The table illustrates that the analysis becomes very large very quickly. For RT-H6, for example, the nine forks generate $2 \times 2 \times 3 \times 4 \times 3 \times 4 \times 2 \times 3 \times 2 = 6,912$ possible paths. Not all *possible* paths are equally reasonable, and the #fincap data help us select the most reasonable ones. The result is a weighted multiverse, where untraveled paths get zero weight. The other ones get weights proportional to the number of teams who picked the path. The vast majority of paths, however, was picked by only one team so the size of the multiverse is slightly less than 164 (the actual number varies across RT-hypotheses).

[28]The Boba software is available at https://github.com/uwdata/boba.

The analysis is done for the original sample as well as for 1000 bootstrapped samples. These additional samples are needed to estimate the correlations in test statistics across paths. These correlations are used to adjust significance levels when accounting for MHT. This is used in assessing whether NSEs are statistically significant, and whether individual estimates are statistically significant (see panels (b) in Table 2 and Table OA.7, where the latter is in the Online Appendix, respectively). Each RT-hypothesis, therefore, requires processing the 720 million trade records almost 164,000 times.[29]

(Insert Figure 5 about here.)

Results. Figure 5 illustrates that the multiverse is able to generate dispersion in estimates that is on par with the dispersion in reported estimates. The box plots for reported estimates are drawn in gray, overlaid by the multiverse box plots in color. The large dispersion in multiverse is remarkable, since they are based on a few decisions only.[30]

(Insert Figure 6 and Figure 7 about here.)

Figure 6 illustrates how sensitive the distribution of estimates is to variation across alternatives at the various forks. The plots reveal that two common strong refractors are the (econometric) model choice and the sampling frequency. A well-known force that drives a wedge

[29]To keep the multiverse analysis feasible, we optimized the code by identifying commonalities across paths and use these to economize on loops. For example, for a particular day, realized spread calculations can iterate once over all trades to obtain realized spreads both for the path that retains all trading and the path that excludes the first and last 30 minutes of trading. Efficient coding further involves identifying opportunities for parallel processing. The multiverse analysis has been implemented on Snellius, a national supercomputer available to Dutch scientists (128 cores and 200 GB internal memory). With all this help, the code took a few days instead of a few months to run for each RT-hypothesis.

[30]The multiverse models only a few forks and its estimates, therefore, are unlikely to accurately predict reported estimates. The explanatory power of regressions with reported #fincap estimates as dependent variables and multiverse estimates as explanatory variables is low. The larger point of the multiverse is to illustrate that, for a subset of forks, variation across paths can generate large non-standard errors. It further allows researchers to drill down and identify the forks that generate most of the dispersion in estimates.

between high- and low-frequency relatives is Jensen's inequality (Blume, 1974):

$$\Pi_{t=1}^{T} \underbrace{E\left(M_t\right)}_{\substack{\text{Expected}\\ \text{high}\\ \text{frequency}\\ \text{relative}}} < \underbrace{E\left(\Pi_{t=1}^{T} M_t\right)}_{\substack{\text{Expected low}\\ \text{frequency relative}}}, \tag{7}$$

if $M_t \in \mathbb{R}^+$ are identical independently distributed random variables, since $f(x) = x^T$ is a convex function. First-order Taylor expanding the left-hand side around one, and then subtracting one from both sides, yields:

$$T\left(E\left(M_t\right) - 1\right) \lesssim E\left(\Pi_{t=1}^{T} M_t\right) - 1. \tag{8}$$

If there are T high-frequency periods in a low-frequency period, then T times the average high-frequency return is expected to be lower than the average low-frequency return. Figure 7 illustrates the effect of this inequality. The three right-most bars illustrate how, for the relative-change model, the median annualized return is -23,000% for data sampled at the daily frequency, -200% for the monthly frequency, and only -4.56% for the yearly frequency. The left-most six bars that correspond to the trend-stationary or log-difference model do not show such discrepancy across frequencies. The reason is that both these models are linear and, therefore, do not suffer from Jensen's inequality. The trend-stationary model features a linear trend and in a log-difference model, the log of a product of relatives becomes a sum of log relatives.

Figure 6 further highlights some idiosyncratic sensitivities. For RT-H1, for example, the second-most sensitive fork is the frequencies that are picked to assess the deviation from a random walk. Further analysis reveals that when comparing high frequencies, such as one-second returns to one-minute returns, then almost all analyses exhibit a decline in market efficiency. But, when comparing low frequencies, such as daily returns to monthly returns, then about half of the analyses show an increase in market efficiency whereas the other half show a decline.

Another example is the retain-negative-sign fork, which is the most sensitive one for RT-H6. The decision each team had to make is whether a negative number that becomes more negative yields a positive percentage change, or a negative percentage change. The first one emphasizes that a (negative) number becomes magnified, whereas the

second one emphasizes a negative trend (i.e., "retain a negative sign"). 21% of the teams picked the first option, 79% picked the second one. It is not surprising that mapping an estimate from the positive to the negative domain causes strong refraction in estimates. This is an example of how a decision that each team might have thought was a trivial one (in sense that there is only one option) can generate non-standard error.

3.4 Alternative explanations

After having presented all our results, it is useful to discuss alternative explanations. Might the sizable non-standard errors be due to the presence of inexperienced researchers testing unsuitable hypotheses with little effort? We believe this is unlikely to be the case for the following reasons.

Experience. Aware of this potential pitfall, we selectively approached researchers (for RTs and PEs), whom we knew were sufficiently experienced in the field. When signing up, they ticked a box that they understood that participating in #fincap requires research expertise and experience in empirical finance/liquidity and the analysis of large datasets. Ticking the box further meant that they acknowledge that one of the team members held a PhD in finance or economics. After ticking the box, researchers had to motivate in an open text box why they believe they meet these requirements. We parsed the content of this box to make sure that the team qualifies before accepting them into #fincap (see Figure OA.7 in the Online Appendix for the sign-up sheet).

Hypotheses. We proceeded with care when designing RT-hypotheses. Early versions were shared with senior scholars, and their feedback helped us fine-tune RT-hypotheses. We, therefore, feel comfortable that the RT-hypotheses are suitable and well motivated hypotheses to test with the RT-sample (see Figure OA.11 in the Online Appendix for the RT instruction sheet, which shows how RT-hypotheses were presented to RTs).

Related to the suitability question, one might wonder whether vagueness of an RT-hypothesis might be a viable alternative expla-

nation for sizable NSEs. To address this concern, we included a very precise RT-hypothesis: RT-H3 on client volume share. The results for RT-H3 show that NSEs can be sizable, even for relatively precise hypotheses. It is true, however, that NSEs tend to be lower for the more precise RT-hypotheses.

Effort. We incentivized research teams to exert effort by providing them with the following information (before they sign up): the deadlines of the various stages so that they could plan for it; their *non-anonymized* paper would be evaluated by senior peer reviewers; the top-five (anonymized) papers would be announced to all others;[31] and, only those who complete all stages become co-authors. In addition to these incentives, we believe that most scientists are propelled by an intrinsic motivation to do good research.

Looking back, we have various reasons to believe that researchers did indeed exert serious effort. First, only four out of 168 research teams failed to complete all stages. 123 out of 168 teams (73.2%) handed in their stage-1 report at least a day early, and none of the teams seriously breached any deadline. The average reproducibility score was 64.5 on a scale from 0 (low) to 100, which is high in comparison to what has been reported in other reproducibility studies (Colliard, Hurlin, and Pérignon, 2021). Finally, the average paper quality was 6.2 on a scale from zero (low) to 10. As for peer evaluators, we also believe they exerted serious effort, because all who signed up as a PE completed their reviews on time.

4 Conclusion

Researchers need to take many decisions when testing hypotheses on a particular sample: pick an appropriate measure, treat outliers, select a statistical model, etc. If researchers are not perfectly aligned on these decisions, their estimates likely differ. This potential dispersion in estimates therefore adds uncertainty to an estimate reported by a

[31] Individuals obtain "ego utility" from positive views about their ability to do well and they exert more effort (or take more risks) when they are informed about their rank in non-incentivized competitions (Köszegi, 2006; Tran and Zeckhauser, 2012; Kirchler, Lindner, and Weitzel, 2018).

single team. Other teams might have reported other estimates based on the same data.

We measure dispersion in estimates across researchers robustly with an inter-quartile range, and refer to it as non-standard error. We study NSEs in an experiment where 164 teams test the same six RT-hypotheses on the same sample. We find NSEs to be substantial, even for a relatively straightforward market-share hypothesis. For this RT-hypothesis, we find it to be 1.2% around a median of -3.3%. A more opaque RT-hypothesis on market-efficiency yields larger variation with an NSE of 6.7% around a median of 1.1%. We further find that NSEs are smaller for better reproducibility and higher quality papers as rated by peers.

A multiverse analysis based on key forks sheds light on how important each fork is in generating dispersion in estimates. It turns out that many forks add substantial dispersion in estimates. Two particularly powerful ones are sampling frequency and the statistical model. Using a non-linear model at high frequency to estimate a low frequency trend can add substantial noise (Jensen's inequality).

NSEs being substantial is worrisome. An encouraging result, however, is that peer feedback reduces NSEs by half. In the real-world, published papers likely have gone through more stages of feedback, which makes #fincap NSEs an upper bound for real-world dispersion in estimates. Published results might further be affected by *p*-hacking (scoped out in #fincap), which is a selective process and thus likely further reduces dispersion, and potentially introduces bias. Overall, we believe the full process towards published empirical research deserves further scrutiny.

Finally, our multiverse analysis provides guidance on what threshold to use in individual tests when accounting for multiple testing. Bonferroni assumes independence among test statistics and adjusts significance levels by the number of tests: 164 in the case of #fincap. Bootstrapped multiverse results show that there is substantial correlation among test statistics and finds adjustment factors that range between 13 and 91 (depending on RT-hypothesis). The threshold for two-sided testing at 5% therefore should be at least $\Phi(1-0.025/13) = 2.9$. This is in line with the 3.0 lower bound recommended by Harvey, Liu, and Zhu (2016) for factor tests in asset pricing.

Appendices

A Reconciliation with pre-analysis-plan results

The original version of *Non-Standard Errors* contains the results of the analysis outlined in the pre-analysis plan. This original version is available as Tinbergen Institute Discussion Paper TI 2021-102/IV. Most tables and figures have not changed.[32]

The only two tables that have changed are Table 3 and 4. The reason is that these are the only two regression tables. In the original version, we estimate a heteroskedasticity model with ordinary least-squares (OLS). The dependent variable is log squared error. However, OLS estimates are notoriously sensitive to extreme outliers, which turn out to be a feature of the #finap sample (see footnote 20 or Figure 1). Quantile regressions are robust to the presence of extreme outliers and are, therefore, more appropriate for the analysis of our sample. Moreover, they model the entire distribution instead of just a conditional mean (as emphasized in the introduction). In the remainder, we compare results across the two tables in the original version and the current version to reconcile previous findings with current ones.

Table 3 in the original version has become Table 3 in the current version. These tables both relate dispersion in estimates to quality variables in order to test the first hypothesis. In the original version, most results are insignificant. The only significance is for reproducibility when using a 2.5%-97.5% winsorized sample. The coefficient of -0.24 implies that a 10% increase in reproducibility coincides with a reduction in the standard deviation of estimates by $1/2 \times 0.24 \times 10\% = 1.2\%$ (the coefficient 1/2 converts variance to SD, see footnote 21 in original paper). In the current version, the first quartile (Q1) co-varies significantly *positively* with reproducibility and paper quality, whereas the third quartile co-varies significantly *negatively* with them. They, therefore, co-vary significantly nega-

[32] More specifically, Table 1, 2, and 5, and Figure 1, 2, 3, 4, and 5 have not changed. In the current version, they appear as Table 1, OA.4, and OA.6, and Figure OA.2, 1a, 1b, OA.3, and 4, respectively, where the OA prefix indicates that they are in the Online Appendix.

tively with IQR. A 10% increase in reproducibility coincides with a reduction in IQR by $10\% \times (0.109 + 0.142) \times 0.44 = 1.1\%$.[33] Note that this effect is in the same ballpark as the 1.2% in the original paper.

Table 4 in the original version has become Table 4 in the current version. In the original version, the unwinsorized sample shows a weakly significant decline in dispersion of estimates across all stages. The effect is also relatively small in magnitude since the SD decline is only 9%. With extreme outliers removed in the 2.5%-97.5% winsorized sample, the decline becomes both significant and larger in magnitude. The SD now declines by 53.5% across all stages. The results in the current version show that Q1 of the estimate distribution increases significantly across all stages and Q3 declines significantly. The result is a decline of 47.2% (depicted in Figure 3). Again, the numbers in both versions are in the same ballpark.

B RT-sample, RT-hypotheses, and results

This appendix presents the RT-hypotheses in detail and the test results of #fincap RTs as a group. The instruction sheet itself is available as Figure OA.11 in the Online Appendix. We start by providing the context that motivates the RT-hypotheses.

B.1 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 dealer ask prices when buying, and accepting dealer bid prices when selling. The wedge between these 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 can now also quote bid and ask prices themselves by submitting (non-marketable)

[33] The square root of the average variance of reproducibility (de-meaned by RT-hypothesis) is 0.44.

standing limit orders. And, investors increasingly 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.2 RT-hypotheses and test results

The RT-hypotheses and results are discussed based on estimates in the final stage of the project (available as Table OA.3 in the Online Appendix). We therefore base our discussion on the results that RTs settle on after receiving all feedback. What do RTs find after having shown some convergence across the stages? And, consistent with the main text, we base our discussion on robust location and dispersion statistics: the median and IQR, respectively. Finally, we note that such discussion is meaningful, because Table OA.7 in the Online Appendix shows that, for all RT-hypotheses, the null of a zero trend is rejected at a 0.5% significance level. This significance level is used for all tests in the remainder of the subsection.

(The first two hypotheses focus on all trades.)

RT-H1. Assuming that informationally-efficient prices follow a random walk, did market efficiency change over time?
Null hypothesis: Market efficiency has not changed over time.

Findings. The median estimate is -1.1% with an IQR of 2.6%. The third quartile is -0.2% and the vast majority therefore finds a negative trend in efficiency. The Bonferroni tests show that 31 RTs find a significant negative trend against only four who find a significant positive trend. The decline seems modest as the across-RT median[34] is -1.1% per year. The small changes add up, though, to a total change in the 2002-2018 sample of approximately $(0.989^{17} - 1) = -17.1\%$. This might reflect a trend of declining depth in the market, possibly due to new regulation in the aftermath of the global financial crisis of 2007-2008. Post-crisis regulation constrains the supply of liquidity by sell-side banks (e.g., Bao, O'Hara, and Zhou, 2018; Jovanovic

[34]The across-RT median includes all RTs, thus also those who report insignificant results.

and Menkveld, 2021). If these banks incur higher inventory costs as a result, then, in equilibrium, one observes larger transitory price pressures thus reducing market efficiency (e.g., Pastor and Stambaugh, 2003; Hendershott and Menkveld, 2014). In the interest of brevity, we discuss all remaining hypotheses in the same way.

RT-H2. 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: The realized spread on market orders has not changed over time.

Findings. The median estimate is -2.3% with an IQR of 4.3%. The third quartile is -0.1% and the vast majority therefore finds a negative trend in realized spread. The tests show that 38 RTs find a significant negative trend, whereas only three RTs find a significant positive trend. The median decline of 2.3% per year implies a 32.7% decline over the full sample. This trend might be due to the arrival of high-frequency market makers who operate at low costs. They do not have the deep pockets that sell-side banks have, but they will offer liquidity for regular small trades by posting near the inside of the market. Their arrival is typically associated with a tighter bid-ask spread, but not necessarily with better liquidity supply for large orders (e.g., Jones, 2013; Angel, Harris, and Spatt, 2015; Menkveld, 2016).

(The remaining hypotheses focus on agency trades only.)

RT-H3. Did the share of client volume in total volume change over time?
Null hypothesis: Client share volume as a fraction of total volume has not changed over time.

Findings. The median estimate is -2.9% with an IQR of 1.7%. The ninth decile is -1.1%, which shows that almost all RTs report a negative trend. The tests show that 123 RTs find a significant negative trend against only two RTs documenting a significant positive trend. An median decline of 2.9% per year implies a total decline of 39.4%

for the full sample. Intermediation, therefore, seems to have increased which should surprise those who believe that the arrival of agency algorithms enables investors to execute optimally themselves, thus reducing the need for intermediation.[35]

RT-H4. On their market orders and marketable limit orders, did the realized bid-ask spread that clients paid, change over time?
Null hypothesis: Client realized spreads have not changed over time.

Findings. The median estimate is -0.2% with an IQR of 2.4%. The third quartile, however, is positive suggesting that a modest majority finds a negative trend. The tests show a bit stronger evidence for a negative trend, because 15 RTs find it to be significantly negative against only eight who find a significant positive trend. The median decline of 0.2% per year translates to a 3.3% decline for the full sample. The decline in client realized spread is therefore only about a tenth of the total realized spread decline, which suggests that market orders of intermediaries benefited most from the general realized-spread decline.

RT-H5. 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: The fraction of client trades executed via market orders and marketable limit orders has not changed over time.

Findings. The median estimate is 0.0% with an IQR of 0.6%. 13 RTs find a significantly negative trend, whereas nine find a significantly positive trend. The results seem rather balanced between a negative and a positive trend. The results therefore seem to suggest that clients neither increased their share of market orders, nor did they decrease it. One might have expected the latter because an increased use of agency algorithms should allow them to execute more through non-marketable limit orders as opposed to market orders or marketable

[35] We verified with Deutsche Börse that this change is not purely mechanical in the sense that, in the sample period, many institutions became an exchange member and, with it, the status of their volume changes from agency to principal.

limit orders. The benefit of execution via a non-marketable limit order is that one earns half the bid-ask spread as opposed to paying it.

RT-H6. 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: Relative gross trading revenue (GTR) for clients has not changed over time.

Findings. The median estimate is 0.0% with an IQR of 1.1%. Three RTs find a significantly positive trend and another three find a significantly negative one. The significance, therefore, is rather weak and balanced. We cautiously conclude that GTR has stayed mostly at the same level throughout the sample.

C Explanatory variables for error variance

C.1 Team quality

The quality measures for research teams are based on the survey that participants filled out upon registration (see Figure OA.7 in the Online Appendix). To keep the regression model both concise and meaningful, we reduce the ordinal variable "current position" and the logarithmic interval-based variable "size of largest dataset worked with" to binary variables. The academic position variable is one if a researcher is either associate or full professor. The dataset variable is one if the researcher has worked with datasets that are contained at least 100 million observations, because the #fincap sample contains

720 million observations. We aggregate these binary variables to research team level by taking the maximum across the team members.

As for self-assessed experience, we asked for both empirical finance and market liquidity, which we deem equally relevant for testing the RT-hypotheses. Thus, and because of the anticipated high correlation, we use the average of these two measures to obtain the individual score. And, in the interest of consistency, we again aggregate to the team level by taking the maximum across the team members.

C.2 Workflow quality

We proxy for workflow quality with an objectively obtained score of code quality provided by Cascad (see footnote 18). The scale ranges from 0 (serious discrepancies) to 100 (perfect reproducibility).

C.3 Paper quality

Papers are rated by an external group of peer evaluators. They rate the analyses associated with each RT-hypothesis individually, but also the paper in its entirety (see Figure OA.16 in the Online Appendix). The ratings range from from 0 (very weak) to 10 (excellent). Each paper is rated by two PEs and the paper rating is the average of the two (after removing a PE fixed effect as discussed in Section 2.1).

D Bootstrap procedure for belief statistic D

The distribution of D under the null of equal means is obtained by bootstrapping as follows. For each RT-hypothesis, we subtract the difference between the average belief on standard deviation and the observed standard deviation, from the beliefs:

$$AdjBeliefOnSD_{ij} = BeliefOnSD_{ij} - \left[\left(\frac{1}{n}\sum_{i} BeliefOnSD_{ij}\right) - RealizationOfSD_{j}\right] \quad (9)$$

In this new sample with adjusted beliefs, the average belief about dispersion equals the observed dispersion, by construction. This sample is input to the bootstrapping procedure which iterates through the following steps 10,000 times:

1. As we have n RTs, in each iteration we draw n times from the new sample, with replacement. Each draw picks a particular RT and stores its beliefs and its results for all of the six RT-hypotheses. The result of these n draws therefore is a simulated sample that has the same size as the original sample.

2. The simulated sample is used to compute the test statistic D in (6). This statistic for iteration k, a scalar, is stored as D_k.

The bootstrap procedure yields 10,000 observations of the test statistic under the null. For a significance level of 0.005, the statistic observed in the #fincap sample is statistically significant if it lands below the 25th lowest simulated statistic or above the 25th highest simulated statistic. Its p-value is:[36]

$$2\min(\textit{EmpiricalQuantileFincapStatistic}, \\ 1 - \textit{EmpiricalQuantileFincapStatistic}). \tag{10}$$

E Anderson-Darlin test

The sensitivity of dispersion to a particular fork is measured by a k-sample Anderson-Darling test (Scholz and Stephens, 1987). This test was designed to verify whether k separate samples are drawn from the same distribution. The AD test statistic T_{k-1} measures the distance between the empirical distribution functions of k separate samples. It does not rely on parametric assumptions. It is, therefore, particularly attractive for our application as distributions are unknown ex-ante. In case of independence, the percentiles of the asymptotic distributions are known (Scholz and Stephens, 1987, Table 1 with $m = k - 1$). T_{k-1} converges to a standard normal for k tending to infinity.

The AD approach builds on tests previously proposed by Kolmogorov, Smirnov, Cramér, and von Mises. It adds a weight function to allow the researcher to attach differential importance to various

[36]Note that the procedure accounts for within-RT correlations (i.e., including possible non-zero correlations among a particular RT's results and the beliefs that it reports). The reason the procedure accounts for these correlations is that the bootstrap uses block-sampling where, when an RT is drawn, all of its beliefs and all of its estimates are drawn. One therefore only assumes independence across RTs which holds by construction given the design of #fincap.

portions of the distribution function (Anderson and Darling, 1964a). It nests the Cramér-von Mises ω^2 statistic which is based on equal weighting. The AD default weighting is one that equalizes the sampling error across the (empirical) support of the distribution function (Anderson and Darling, 1964b, p. 767). It effectively attaches more weight to the tails of the distribution. Scholz and Stephens (1987, p. 919) argue that among alternatives, the AD test statistic has attractive small sample (i.e., small k) properties.

References

Aczel, Balazs, Barnabas Szaszi, Gustav Nilsonne, Olmo R. van den Akker, Casper J. Albers, Marcel A.L.M. van Assen, Jojanneke A. Bastiaansen, Dan Benjamin, Udo Boehm, Rotem Botvinik-Nezer, Laura F. Bringmann, Niko A. Busch, Emmanuel Caruyer, Andrea M. Cataldo, Nelson Cowan, Andrew Delios, Noah N.N. van Dongen, Chris Donkin, Johnny B. van Doorn, Anna Dreber, Gilles Dutilh, Gary F. Egan, Morton Ann Gernsbacher, Rink Hoekstra, Sabine Hoffmann, Felix Holzmeister, Juergen Huber, Magnus Johannesson, Kai J. Jonas, Alexander T. Kindel, Michael Kirchler, Yoram K. Kunkels, D. Stephen Lindsay, Jean-Francois Mangin, Dora Matzke, Marcus R Munafò, Ben R. Newell, Brian A. Nosek, Russell A Poldrack, Don van Ravenzwaaij5, Jörg Rieskamp, Matthew J. Salganik, Alexandra Sarafoglou, Tom Schonberg, Martin Schweinsberg, David Shanks, Raphael Silberzahn, Daniel J. Simons, Barbara A. Spellman, Samuel St-Jean, Jeffrey J. Starns, Eric L. Uhlmann, Jelte Wicherts, and Eric-Jan Wagenmakers. 2021. "Consensus-Based Guidance for Conducting and Reporting Multi-Analyst Studies." *eLife* 10:1–13.

Anderson, Theodore W. and Donald A. Darling. 1964a. "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes." *The Annals of Mathematical Statistics* 23:193–212.

———. 1964b. "A Test of Goodness of Fit." *Journal of the American Statistical Association* 49:765–769.

Angel, James J., Lawrence E. Harris, and Chester S. Spatt. 2015. "Eq-

uity Trading in the 21st Century: An Update." *Quarterly Journal of Finance* 5:1–39.

Bao, Jack, Maureen O'Hara, and Xing (Alex) Zhou. 2018. "The Volcker Rule and Corporate Bond Market Making in Times of Stress." *Journal of Financial Economics* 130:95–113.

Ben-David, Itzhak, Francesco Franzoni, and Byungwook Kim Rabih Moussawi. 2021. "Competition for Attention in the ETF Space." Manuscript, Ohio State University.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony G. Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P.A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don A. Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E. Johnson. 2018. "Redefine Statistical Significance." *Nature Human Behavior* 2:6–10.

Black, Bernard S., Hemang Desai, Kate Litvak, Woongsun Yoo, and Jeff Jiewei Yu. 2021. "Specification Choice in Randomized and Natural Experiments: Lessons from the Regulation SHO Experiment." Manuscript, Northwestern University.

Blume, Marshall E. 1974. "Unbiased Estimators of Long-Run Expected Rates of Return." *Journal of the American Statistical Association* 69:634–638.

Bonferroni, Carlo E. 1936. "Teoria Statistica delle Classi e Calcolo delle Probabilità." *Pubblicazioni del R Istuto Superiore di Scienze Economiche e Commerciali di Firenze* 8:1–62.

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A. Mumford, R. Alison Adcock, Paolo Avesani, Blazej M. Baczkowski, Aahana Bajracharya, Leah Bakst, Sheryl Ball, Marco Barilari, Nadège Bault, Derek Beaton, Julia Beitner, Roland G. Benoit, Ruud M.W.J. Berkers, Jamil P. Bhanji, Bharat B. Biswal, Sebastian Bobadilla-Suarez, Tiago Bortolin, Katherine L. Bottenhorn, Alexander Bowring, Senne Braem, Hayley R. Brooks, Emily G. Brudner, Cristian B. Calderon, Julia A. Camilleri, Jaime J. Castrellon, Luca Cecchetti, Edna C. Cieslik, Zachary J. Cole, Olivier Collignon, Robert W. Cox, William A. Cunningham, Stefan Czoschke, Kamalaker Dadi, Charles P. Davis, Alberto De Lucas, Mauricio R. Delgado, Lysia Demetriou, Jeffrey B. Dennison, Xin Di, Erin W. Dickie, Ekaterina Dobryakova, Claire L. Donnat, Juergen Dukart, Niall W. Duncan, Joke Durnez, Amr Eed, Simon B. Eickhoff, Andrew Erhart, Laura Fontanesi, G. Matthew Fricke, Shiguang Fu, Adriana Galván, Remi Gau, Sarah Genon, Tristan Glatard, Enrico Glerean, Jelle J. Goeman, Sergej A.E. Golowin, Carlos Gonzáez-García, Krzysztof J. Gorgolewski, Cheryl L. Grady, Mikella A. Green, Joao F. Guassi Moreira, Olivia Guest, Shabnam Hakimi, J. Paul Hamilton, Roeland Hancock, Giacomo Handjaras, Bronson B. Harry, Colin Hawco, Peer Herholz, Gabrielle Herman, Stephan Heunis, Felix Hoffstaedter, Jeremy Hogeveen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacovella, Alexandru D. Iordan, Peder M. Isager, Ayse I. Isik, Andrew Jahn, Matthew R. Johnson, Tom Johnstone, Michael J.E. Joseph, Anthony C. Juliano, Joseph W. Kable, Michalis Kassinopoulos, Cemal Koba, Xiang-Zhen Kong, Timothy R. Koscik, Nuri Erkut Kucukboyaci, Brice A. Kuhl, Sebastian Kupek, Angela R. Laird, Claus Lamm, Robert Langner, Nina Lauharatanahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Andrea Leo, Elise Lesage, Flora Li, Monica Y.C. Li, Phui Cheng Lim, Evan N. Lintz, Schuyler W. Liphardt, Annabel B. Losecaat Vermeer, Bradley C. Love, Michael L. Mack, Norberto Malpica, Theo Marins, Camille

Maumet, Kelsey McDonald, Joseph T. McGuire, Helena Melero, Adriana S. Méndez Leal, Benjamin Meyer, Kristin N. Meyer, Glad Mihai, Georgios D. Mitsis, Jorge Moll, Dylan M. Nielson, Gustav Nilsonne, Michael P. Notter, Emanuele Olivetti, Adrian I. Onicas, Paolo Papale, Kaustubh R. Patil, Jonathan E. Peelle, Alexandre Pérez, Doris Pischedda, Jean-Baptiste Poline, Yanina Prystauka, Shruti Ray, Patricia A. Reuter-Lorenz, Richard C. Reynolds, Emiliano Ricciardi, Jenny R. Rieck, Anais M. Rodriguez-Thompson, Anthony Romyn, Taylor Salo, Gregory R. Samanez-Larkin, Emilio Sanz-Morales, Margaret L. Schlichting, Douglas H. Schultz, Qiang Shen, Margaret A. Sheridan, Jennifer A. Silvers, Kenny Skagerlund, Alec Smith, David V. Smith, Peter Sokol-Hessner, Simon R. Steinkamp, Sarah M. Tashjian, Bertrand Thirion, John N. Thorp, Gustav Tinghög, Loreen Tisdall, Steven H. Tompson, Claudio Toro-Serey, Juan Jesus Torre Tresols, Leonardo Tozzi, Vuong Truong, Luca Turella, Anna van â€˜t Veer, Tom Verguts, Jean M. Vettel, Sagana Vijayarajah, Khoi Vo, Matthew B. Wall, Wouter D. Weeda, Susanne Weis, David J. White, David Wisniewski, Alba Xifra-Porxas, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S. L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E. Nichols, Russell A. Poldrack, and Tom Schonberg. 2020. "Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams." *Nature* 582:84–88.

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Henrik K. Andersen, Daniel Auer, Flavio Azevedo, Oke Bahnsen, Dave Balzer, Gerrit Bauer, Paul C. Bauer, Markus Baumann, Sharon Baute, Verena Benoit, Julian Bernauer, Carl Berning, Anna Berthold, Felix S. Bethke, Thomas Biegert, Katharina Blinzler, Johannes N. Blumenberg, Licia Bobzien, Andrea Bohman, Thijs Bol, Amie Bostic, Zuzanna Brzozowska, Katharina Burgdorf, Kaspar Burger, Kathrin Busch, Juan Carlos-Castillo, Nathan Chan, Pablo Christmann, Roxanne Connelly, Christian S. Czymara, Elena Damian, Alejandro Ecker, Achim Edelmann, Maureen A. Eger, Simon Ellerbrock, Anna Forke, Andrea Forster, Chris Gaasendam, Konstantin Gavras, Vernon Gayle, Theresa Gessler, Timo Gnambs, Amélie Godefroidt, Max Grömping, Martin Groß, Stefan Gruber, Tobias Gummer, Andreas Hadjar, Jan Paul Heisig, Sebas-

tian Hellmeier, Stefanie Heyne, Magdalena Hirsch, Mikael Hjerm, Oshrat Hochman, Andreas Hövermann, Sophia Hunger, Christian Hunkler, Nora Huth, Zsófia S. Ignácz, Laura Jacobs, Jannes Jacobsen, Bastian Jaeger, Sebastian Jungkunz, Nils Jungmann, Mathias Kauff, Manuel Kleinert, Julia Klinger, Jan-Philipp Kolb, Marta Kołczyńska, John Kuk, Katharina Kunißen, Dafina Kurti Sinatra, Alexander Langenkamp, Philipp M. Lersch, Lea-Maria Löbel, Philipp Lutscher, Matthias Mader, Joan E. Madia, Natalia Malancu, Luis Maldonado, Helge-Johannes Marahrens, Nicole Martin, Paul Martinez, Jochen Mayerl, Oscar J. Mayorga, Patricia McManus, Kyle McWagner, Cecil Meeusen, Daniel Meierrieks, Jonathan Mellon, Friedolin Merhout, Samuel Merk, Daniel Meyer, Leticia Micheli, Jonathan Mijs, Cristóbal Moya, Marcel Neunhoeffer, Daniel Nüst, Olav Nygård, Fabian Ochsenfeld, Gunnar Otte, Anna Pechenkina, Christopher Prosser, Louis Raes, Kevin Ralston, Miguel Ramos, Arne Roets, Jonathan Rogers, Guido Ropers, Robin Samuel, Gregor Sand, Ariela Schachter, Merlin Schaeffer, David Schieferdecker, Elmar Schlueter, Regine Schmidt, Katja M. Schmidt, Alexander Schmidt-Catran, Claudia Schmiedeberg, Jürgen Schneider, Martijn Schoonvelde, Julia Schulte-Cloos, Sandy Schumann, Reinhard Schunck, Jürgen Schupp, Julian Seuring, Henning Silber, Willem Sleegers, Nico Sonntag, Alexander Staudt, Nadia Steiber, Nils Steiner, Sebastian Sternberg, Dieter Stiers, Dragana Stojmenovska, Nora Storz, Erich Striessnig, Anne-Kathrin Stroppe, Janna Teltemann, Andrey Tibajev, Brian Tung, Giacomo Vagni, Jasper Van Assche, Meta van der Linden, Jolanda van der Noll, Arno Van Hootegem, Stefan Vogtenhuber, Bogdan Voicu, Fieke Wagemans, Nadja Wehl, Hannah Werner, Brenton M. Wiernik, Fabian Winter, Christof Wolf, Yuki Yamada, Nan Zhang, Conrad Ziller, Stefan Zins, Tomasz Żółtak, and Hung H.V. Nguyen. 2021. "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty." Manuscript, University of Bremen.

Breznau, Nate, Deadric T. Williams, Katrin Auspurg, Josef Bruederl, Felix Holzmeister, Gustav Nilsonne, Balázs Aczel, Cory J. Clark, and Eric Uhlmann. 2022. "Is the Variability in Social Science Research Results Produced by Different Researchers Explicable? An Adversarial Collaboration and Joint Effort to Parse the Dispersion

in Estimates." Manuscript, University of Bremen.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351:1433–1436.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science." *Nature Human Behaviour* 2:637–644.

Chen, Andrew Y. 2021. "The Limits of *p*-Hacking: Some Thought Experiments." *Journal of Finance* 76:2447–2480.

Chordia, Tarun, Amit Goyal, and Alessio Saretto. 2020. "Anomalies and False Rejections." *Review of Financial Studies* 33:2134–2179.

Colliard, Jean-Edouard, Christophe Hurlin, and Christophe Pérignon. 2021. "The Economics of Research Reproducibility." Manuscript, HEC Paris.

Gelman, Andrew and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102:460–465.

Geyer-Klingeberg, Jerome, Markus Hang, and Andreas Rathgeber. 2020. "Meta-analysis in finance research: Opportunities, challenges, and contemporary applications." *International Review of Financial Analysis* 7:1–15.

Harvey, Campbell R. 2017. "Presidential Address: The Scientific Outlook in Financial Economics." *Journal of Finance* 72:1399–1440.

Harvey, Campbell R. and Yan Liu. 2020. "False (and Missed) Discoveries in Financial Economics." *Journal of Finance* 75:2503–2553.

Harvey, Campbell R., Yan Liu, and Heqing Zhu. 2016. "...and the Cross-Section of Expected Returns." *Review of Financial Studies* 29:5–68.

Hendershott, Terrence and Albert J. Menkveld. 2014. "Price Pressures." *Journal of Financial Economics* 114:405–423.

Hou, Kewei, Chen Xue, and Lu Zhang. 2018. "Replicating Anomalies." *Review of Financial Studies* 33:2019–2133.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, , Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yani Stopnitzky. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry* 59:944–960.

Ioannidis, John P.A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2:696–701.

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Pedersen. 2022. "Is There a Replication Crisis in Finance." *Journal of Finance (forthcoming)* .

Jones, Charles M. 2013. "What Do We Know About High-Frequency Trading?" Manuscript, Columbia University.

Jovanovic, Boyan and Albert J. Menkveld. 2021. "Equilibrium Bid-Price Dispersion." *Journal of Political Economy (forthcoming)* .

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment*. 1 London Bridge Street, London: William Collins.

Kirchler, Michael, Florian Lindner, and Utz Weitzel. 2018. "Rankings and Risk-Taking in the Finance Industry." *Journal of Finance* 73:2271–2302.

Koenker, Roger and Gilbert Bassett Jr. 1978. "Regression Quantiles." *Econometrica* 46:33–50.

Köszegi, Botond. 2006. "Ego Utility, Overconfidence, and Task Choice." *Journal of the European Economic Association* 4:673–707.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73:31–43.

Linnainmaa, Juhani T. and Michael R. Roberts. 2018. "The History of the Cross-Section of Stock Returns." *Review of Financial Studies* 31:2606–2649.

Liu, Yang, Alex Kale, Tim Althoff, and Jeffrey Heer. 2021. "Boba: Authoring and Visualizing Multiverse Analyses." *IEEE Transactions on Visualization and Computer Graphics* 27:1753–1763.

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*. New York: Oxford University Press.

Mavroeidis, Sophocles, Mikkel Plagborg-Møller, and James H. Stock. 2014. "Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve." *Journal of Economic Literature* 52:124–188.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251).

McLean, R. David and Jeffrey Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability." *Journal of Finance* 71:5–31.

Menkveld, Albert J. 2016. "The Economics of High-Frequency Trading: Taking Stock." *Annual Review of Financial Economics* 8:1–24.

Mitton, Todd. 2021. "Methodological Variation in Empirical Corporate Finance." *Review of Financial Studies (forthcoming)* .

Munafò, Marcus R., Brian A. Nosek, Dorothy V.M. Bishop, Katerine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P.A. Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1:1–9.

Parlour, Christine A. and Duane J. Seppi. 2008. "Limit Order Markets: A Survey." In *Handbook of Financial Intermediation and Banking*, edited by Arnoud W.A. Boot and Anjan V. Thakor. Amsterdam, Netherlands: Elsevier Publishing.

Pastor, L. and R.F. Stambaugh. 2003. "Liquidity Risk and Expected Returns." *Journal of Political Economy* 111:642–685.

Scholz, Fritz W. and Michael A. Stephens. 1987. "K-Sample Anderson-Darling Tests." *Journal of the American Statistical Association* 82:918–924.

Schweinsberg, Martin, Michael Feldman, Nicola Staub, Olmo R. van den Akker, Robbie C.M. van Aert, Marcel A.L.M. van Assen, Yang Liu, Tim Althoff, Jeffrey Heer, Alex Kale, Zainab Mohamed, Hashem Amireh, Vaishali Venkatesh Prasad, Abraham Bernstein, Emily Robinson, Kaisa Snellman, S. Amy Sommer, Sarah M.G. Otner, David Robinson, Nikhil Madan, Raphael Silberzahn, Pavel Goldstein, Warren Tierney, Toshio Murase, Benjamin Mandl, Domenico Viganola, Carolin Strobl, Catherine B.C. Schaumans, Stijn Kelchtermans, Chan Naseeb, S. Mason Garrison, Tal Yarkoni, C.S. Richard Chan, Prestone Adie, Paulius Alaburda, Casper Albers, Sara Alspaugh, Jeff Alstott, Andrew A. Nelson, Eduardo Ariño de la Rubia, Adbi Arzi, Štěpán Bahník, Jason Baik, Laura Winther Balling, Sachin Banker, David AA Baranger, Dale J. Barr, Brenda Barros-Rivera, Matt Bauer, Enuh Blaise, Lisa Boelen, Katerina Bohle Carbonell, Robert A. Briers, Oliver Burkhard, Miguel-Angel Canela, Laura Castrillo, Timothy Catlett, Olivia Chen, Michael Clark, Brent Cohn, Alex Coppock, Natàlia Cugueró-Escofet, Paul G. Curran, Wilson Cyrus-Lai, David Dai, Giulio Valentino Dalla Riva, Henrik Danielsson, Rosaria de F.S.M. Russo, Niko de Silva, Curdin Derungs, Frank Dondelinger, Carolina Duarte de Souza, B. Tyson Dube, Marina Dubova, Ben Mark Dunn, Peter Adriaan Edelsbrunner, Sara Finley, Nick Fox, Timo Gnambs, Yuanyuan Gong, Erin Grand, Brandon Greenawalt, Dan Han, Paul H.P. Hanel, Antony B. Hong, David Hood, Justin Hsueh, Lilian Huang, Kent N. Hui, Keith A. Hultman, Azka Javaid, Lily Ji Jiang, Jonathan Jong, Jash Kamdar, David Kane, Gregor Kappler, Erikson Kaszubowski, Christopher M. Kavanagh, Madian Khabsa, Bennett Kleinberg, Jens Kouros, Heather Krause, Angelos-Miltiadis Krypotos, Dejan Lavbič, Rui Ling Lee, Timothy Leffel, Wei Yang Lim, Silvia Liverani, Bianca Loh, Dorte Lønsmann, Jia Wei Low, Alton Lu, Kyle MacDonald, Christopher R. Madan, Lasse Hjorth Madsen, Christina Maimone, Alexan-

dra Mangold, Adrienne Marshall, Helena Ester Matskewich, Kimia Mavon, Katherine L. McLain, Amelia A. McNamara, Mhairi McNeill, Ulf Mertens, David Miller, Ben Moore, Andrew Moore, Eric Nantz, Ziauddin Nasrullah, Valentina Nejkovic, Colleen S Nell, Andrew Arthur Nelson, Gustav Nilsonne, Rory Nolan, Christopher E. O'Brien, Patrick O'Neill, Kieran O'Shea, Toto Olita, Jahna Otterbacher, Diana Palsetia, Bianca Pereira, Ivan Pozdniakov, John Protzko, Jean-Nicolas Reyt, Travis Riddle, Amal (Akmal) Ridhwan Omar Ali, Ivan Ropovik, Joshua M. Rosenberg, Stephane Rothen, Michael Schulte-Mecklenbeck, Nirek Sharma, Gordon Shotwell, Martin Skarzynski, William Stedden, Victoria Stodden, Martin A. Stoffel, Scott Stoltzman, Subashini Subbaiah, Rachael Tatman, Paul H. Thibodeau, Sabina Tomkins, Ana Valdivia, Gerrieke B. Druijff-van de Woestijne, Laura Viana, Florence Villesèche, W. Duncan Wadsworth, Florian Wanders, Krista Watts, Jason D Wells, Christopher E. Whelpley, Andy Won, Lawrence Wu, Arthur Yip, Casey Youngflesh, Ju-Chi Yu, Arash Zandian, Leilei Zhang, Chava Zibman, and Eric Luis Uhlmann. 2021. "Same Data, Different Conclusions: Radical Dispersion in Empirical Results when Independent Analysts Operationalize and Test the Same Hypothesis." *Organizational Behavior and Human Decision Processes* 165:228–249.

Šidák, Zbyně. 1967. "Rectangular confidence regions for the means of multivariate normal distributions." *Journal of the American Statistical Association* 62 (318):626–633.

Silberzahn, Raphael, Eric L. Uhlmann, Dan P. Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štepán Bahník, Feng Bai, Colin Bannard, Evelina Bonnier, Rickard Carlsson, Felix Cheung, Garret Christensen, Russ Clay, Maureen A. Craig, Anna Dalla Rosa, Lammertjan Dam, Mathew H. Evans, Ismael Flores Cervantes, Nathan Fong, Monica Gamez-Djokic, Andreas Glenz, Shauna Gordon-McKeon, Tim J. Heaton, Karin Hederos, Moritz Heene, Alicia J. Hofelich Mohr, Fabia Högden, Kent Hui, Magnus Johannesson, Jonathan Kalodimos, Erikson Kaszubowski, Deanna M. Kennedy, Ryan Lei, Thomas A. Lindsay, Silvia Liverani, Christopher R. Madan, Daniel Molden, Eric Molleman, Richard D. Morey, Laetitia B. Mulder, Bernard R. Nijs-

tad, Nolan G. Pope, Bryson Pope, Jason M. Prenoveau, Floor Rink, Egidio Robusto, Hadiya Roderique, Anna Sandberg, Elmar Schlüter, Felix D. Schönbrodt, Martin F. Sherman, S. Amy Sommer, Kristin Sotak, Seth Spain, Christoph Spörlein, Tom Stafford, Luca Stefanutti, Susanne Tauber, Johannes Ullrich, Michelangelo Vianello, Eric-Jan Wagenmakers, Maciej Witkowiak, Sangsuk Yoon, and Brian A. Nosek. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1:337–356.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency Through a Multiverse Analysis." *Psychological Science* 11:702–712.

Tran, Anh and Richard Zeckhauser. 2012. "Rank as an Inherent Incentive: Evidence from a Field Experiment." *Journal of Public Economics* 96:645–650.

Yule, G. Udny. 1897. "On the Significance of Bravais' Formulae for Regression." *Proceedings of the Royal Society of London* 60:477–489.

Table 1: Summary statistics

This table presents summary statistics. Standard deviations are in parentheses.

Panel (a): Quality of the #fincap community

	Research teams	Peer evaluators
Fraction with top finance/econ publications (see footnote 6)	0.31	0.85
Fraction including at least associate/full professor	0.52	0.88
Experience empirical-finance research (low-high, 1-10)	8.1 (1.7)	8.4 (1.8)
Experience market-liquidity research (low-high, 1-10)	6.9 (2.4)	7.8 (2.3)
Relevant experience (average of the above two items)	7.5 (1.3)	8.1 (1.7)
Fraction with "big data" experience (>#fincap sample)	0.65	0.88
Fraction teams consisting of two members (maximum team size)	0.79	
Number of observations	164	34

Panel (b): Quality of the analysis of research teams

	Research teams
Reproducibility score according to Cascad (low-high, 0-100)	64.5 (43.7)
Paper quality as judged by peer evaluators (low-high, 0-10)	6.2 (2.0)

(continued on next page)

(continued from previous page)

Panel (c): Dispersion across teams of stage-1 results: Estimates, SEs, and *t*-values

	RT-H1 Efficiency	RT-H2 RSpread	RT-H3 Client Volume	RT-H4 Client RSpread	RT-H5 Client MOrders	RT-H6 Client GTR
Estimate (yearly change, %)						
Mean	446.3	-1,093.4	-3.5	-38,276.1	-3.5	-87.1
SD	5,817.5	14,537.2	9.4	490,024.2	37.6	728.5
Min	-171.1	-186,074.5	-117.5	-6,275,383.0	-452.9	-8,254.5
Q(0.10)	-23.7	-6.9	-3.8	-6.7	-1.6	-192.1
Q(0.25)	-6.2	-3.6	-3.5	-2.1	-0.6	-18.2
Median	-1.1	-0.0	-3.3	0.1	-0.0	0.0
Q(0.75)	0.5	3.9	-2.4	3.8	0.2	3.2
Q(0.90)	3.7	21.5	-0.1	20.4	1.0	56.5
IQR (i.e., NSE)	6.7	7.5	1.2	5.9	0.8	21.4
IDR	27.3	28.4	3.7	27.1	2.5	248.5
Max	74,491.1	4,124.0	8.7	870.2	69.5	1,119.0
Standard error						
Mean	468.7	1,195.3	3.7	38,302.0	6.2	148.2
SD	5,810.6	14,711.9	29.5	489,929.5	40.1	526.0
Min	0.0	0.0	0.0	0.0	0.0	0.0
Q(0.10)	0.1	0.2	0.1	0.2	0.1	0.0
Q(0.25)	0.5	1.1	0.3	1.2	0.2	0.7
Median	2.5	5.0	1.4	4.4	1.0	9.7
Q(0.75)	9.3	13.9	2.0	14.3	2.4	77.1
Q(0.90)	44.7	39.6	2.2	31.2	3.1	235.4
IQR	8.8	12.8	1.7	13.1	2.2	76.4
IDR	44.6	39.4	2.1	31.0	3.1	235.4
Max	74,425.5	188,404.1	378.8	6,274,203.0	463.7	4,836.2
t-value						
Mean	-3.6	35.3	-47.1	24.3	-5.7	-2.0
SD	28.4	541.2	269.9	406.0	60.1	21.2
Min	-322.3	-764.6	-2,770.6	-852.6	-631.6	-191.7
Q(0.10)	-4.7	-5.7	-37.4	-3.5	-2.3	-1.7
Q(0.25)	-1.9	-1.5	-11.5	-1.0	-0.6	-1.0
Median	-0.7	-0.1	-1.8	0.1	0.0	0.0
Q(0.75)	0.3	0.8	-1.6	1.0	0.8	0.7
Q(0.90)	1.7	1.5	-0.3	1.6	1.7	1.2
IQR	2.2	2.3	9.9	1.9	1.3	1.7
IDR	6.4	7.2	37.1	5.2	3.9	2.9
Max	51.6	6,880.5	29.5	5,119.5	89.6	100.6

Table 2: Non-standard error test

This table tests for the presence of non-standard errors in stage 1. It does so by testing whether estimates provided by researchers deviate from the median across researchers. Critical values of the individual tests are raised to achieve the desired significance level at the family of tests. The number of significantly negative tests and significantly positive tests is reported in brackets. The reported family p-value is the probability that out of all test statistics, at least one is larger than the reported value, under the null of a multivariate normal with means equal to the realized #fincap medians, and a covariance matrix with squared SEs (reported by the RTs) on the diagonal and off-diagonals that are either zero (Bonferroni) or based on the multiverse analysis (Section 3.3).

Panel (a): Multiple tests (Bonferroni)

	Reject no-NSE at 0.5%?	p-value of family test	Mean (SD) correlation test statistics	Effective number of tests
RT-H1	Yes (8, 25)	<0.0001	0.00 (0.00)	164
RT-H2	Yes (24, 10)	<0.0001	0.00 (0.00)	164
RT-H3	Yes (13, 25)	<0.0001	0.00 (0.00)	164
RT-H4	Yes (22, 4)	<0.0001	0.00 (0.00)	164
RT-H5	Yes (13, 10)	<0.0001	0.00 (0.00)	164
RT-H6	Yes (8, 3)	<0.0001	0.00 (0.00)	164

Panel (b): Multiple tests (based on multiverse analysis)

	Reject no-NSE at 0.5%?	p-value of family test	Mean (SD) correlation test statistics	Effective number of tests
RT-H1	Yes (8, 26)	<0.0001	0.03 (0.21)	77
RT-H2	Yes (24, 10)	<0.0001	0.05 (0.22)	81
RT-H3	Yes (13, 26)	<0.0001	0.22 (0.34)	21
RT-H4	Yes (22, 4)	<0.0001	0.08 (0.24)	67
RT-H5	Yes (13, 10)	<0.0001	0.20 (0.34)	31
RT-H6	Yes (8, 3)	<0.0001	0.02 (0.21)	86

Table 3: Stage-1 quantile regressions

This table presents the results of quantile regressions that characterize how the distribution of stage-1 estimates co-varies with various quality metrics. These metrics are team quality, reproducibility score, and (de-meaned) peer rating. The three quality variables have been standardized and, subsequently, multiplied by the IQR per RT-hypothesis. Their coefficient therefore measures the result of a one-standard deviation change, expressed in terms of interquartile-range units. $^{*}/^{**}$ correspond to significance at the 5/0.5% level, respectively.

	Q(0.10)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.90)
Team quality (standardized/scaled)	0.597**	0.004	0.002	0.032**	-0.325**
	(0.030)	(0.014)	(0.007)	(0.012)	(0.030)
Reproducibility score (standardized/scaled)	0.473**	0.109**	-0.001	-0.142**	-0.555**
	(0.033)	(0.014)	(0.007)	(0.011)	(0.028)
Average rating (standardized/scaled)	0.766**	0.230**	-0.001	-0.097**	-0.626**
	(0.034)	(0.014)	(0.007)	(0.011)	(0.028)
Dummy RT-H1 Efficiency	-29.592**	-6.099**	-1.132**	0.939**	9.057**
	(0.813)	(0.340)	(0.166)	(0.269)	(0.708)
Dummy RT-H2 RSpread	-15.933**	-3.930**	-0.017	3.674**	22.451**
	(0.849)	(0.342)	(0.166)	(0.268)	(0.705)
Dummy RT-H3 Client Volume	-5.629**	-3.789**	-3.319**	-2.386**	0.221
	(0.836)	(0.339)	(0.166)	(0.268)	(0.721)
Dummy RT-H4 Client RSpread	-12.089**	-2.437**	0.162	4.161**	19.619**
	(0.837)	(0.340)	(0.166)	(0.266)	(0.704)
Dummy RT-H5 Client MOrders	-2.479**	-0.744*	-0.001	0.297	1.625*
	(0.837)	(0.339)	(0.166)	(0.268)	(0.721)
Dummy RT-H6 GTR	-194.457**	-21.385**	0.022	5.137**	65.203**
	(0.806)	(0.337)	(0.167)	(0.268)	(0.679)
#Observations	984	984	984	984	984

Table 4: All-stages quantile regressions

This table presents the results of quantile regressions that characterize how the distribution of estimates varies across all stages of the #fincap project. The stage dummies have been multiplied by the (stage-1) IQR per RT-hypothesis. Their coefficient therefore measures the effect in terms of interquartile-range units. Standard errors account for correlation in residuals by adding RT-hypothesis fixed effects and by clustering per RT across all stages. $^{*}/^{**}$ correspond to significance at the 5/0.5% level, respectively.

	Q(0.10)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.90)
Dummy Stage 2 - Dummy Stage 1	2.44*	0.07	-0.00	-0.06	-0.73
	(1.18)	(0.14)	(0.01)	(0.06)	(0.64)
Dummy Stage 3 - Dummy Stage 2	0.94*	0.15	0.00	-0.09	-0.73
	(0.41)	(0.09)	(0.01)	(0.05)	(0.40)
Dummy Stage 4 - Dummy Stage 3	0.21*	0.06*	0.00	-0.04	-0.25*
	(0.09)	(0.03)	(0.01)	(0.03)	(0.11)
Dummy Stage 4 - Dummy Stage 1	3.59**	0.28*	-0.00	-0.19**	-1.71**
	(1.23)	(0.14)	(0.01)	(0.05)	(0.50)
RT-hypotheses dummies	Yes	Yes	Yes	Yes	Yes
#Observations	3,936	3,936	3,936	3,936	3,936

Table 5: Analysis paths

This table summarizes all analysis paths by spelling out all forks and all alternatives at these forks. It further presents the empirical distribution of decisions at all forks.

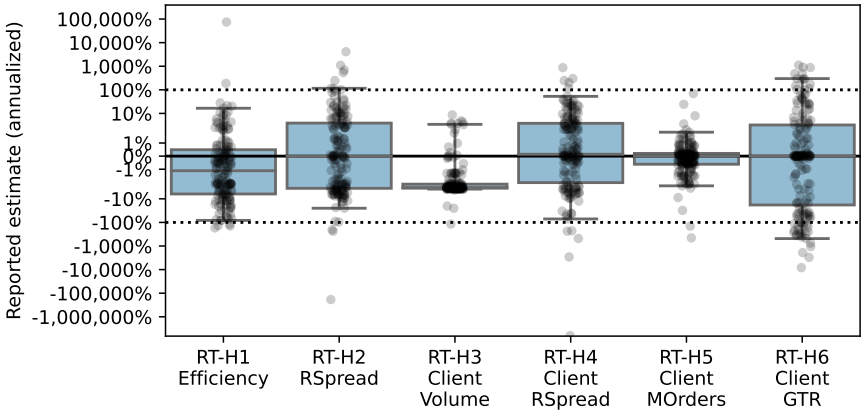
RT-hypothesis	Fork	Fork description	Alternatives	Frequency
All	1	Remove open/close	No	79%
			Yes, 30 minutes	21%
All	2	Days excluded	None	81%
			Settlement weeks	19%
All	3	Outlier treatment	None	65%
			Winsorize measure at 2.5 and 97.5 percentile[a]	20%
			Trim measure at 2.5 and 97.5 percentile[a]	14%
All	4	Frequency analysis	Daily	37%
			Weekly	1%
			Monthly	21%
			Annual	41%
All	5	Model	Trend stationary (regresion with linear trend)	35%
			Log difference (trivial regression, i.e., intercept only)	5%
			Relative difference (trivial regression)	60%
1	6	Measure	Variance ratio (low-frequency in numerator)	63%
			Autocorrelation (R^2 of AR model for returns)	37%
1	7	Measure frequencies	Second to minute	18%
			One to five minutes	26%
			Five to thirty minutes	34%
			Day to week	13%
			Day to month	10%
2,4,5	6	Tick test or aggressor flag	Aggressor flag (available only for part of the sample)	84%
			Tick test	16%
2,4	7	Post-trade value	Price 5 minutes after trade	81%
			Price 10 minutes after trade	6%
			Price 30 minutes after trade	13%
2,4	8	Aggregation	Equal-weighted average	47%
			Trade-size-weighted average	53%
3	6	Units...	Volume expressed in #contracts	70%
			Volume expressed in euro	30%
6	6	Reference price	Last trade price in the day	62%
			Last trade price one day later	1%
			Volume-weighted-average-price (VWAP) full-day	24%
			VWAP based on last five trades in the day	0%
6	7	Mean or median	Mean	96%
			Median	4%
6	8	Handle non-negatives	Translate and transform ($\varepsilon = 0.001$)	14%
			Translate and transform ($\varepsilon = 1$)	7%
			Set to missing	79%
6	9	Retain negative-trend sign	Yes	79%
			No	21%

[a]: Winsorization is applied at the frequency of analysis (fork 4).

Figure 1: Dispersion of stage-1 estimates across research teams

This plot illustrates the dispersion of stage-1 estimates across research teams. These estimates all focus on a trend in the sample, expressed in terms of a yearly percentage change. The six box plots correspond to the six trends RTs were asked to estimate. The boxes depict the first and third quartile. The horizontal line in the box corresponds to the median. The whiskers depict the 2.5% and 97.5% quantile. All estimates are also plotted individually as gray dots.

Panel (a): Dispersion of all estimates (N=164)



Panel (b): Dispersion of highest quality estimates (N=9)

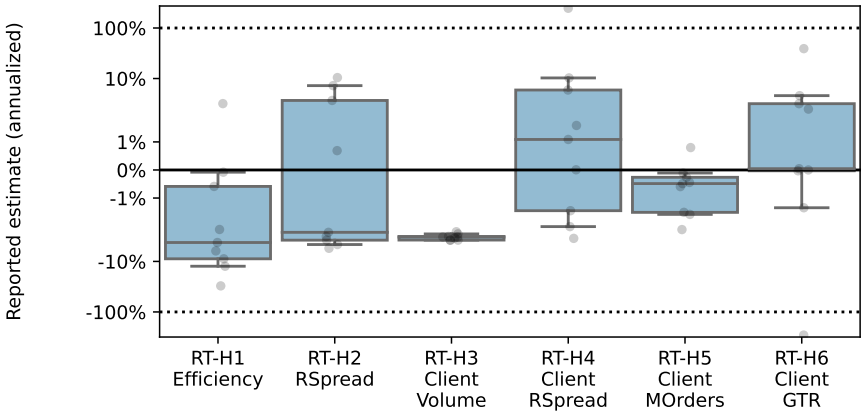


Figure 2: Dispersion in estimates related to quality measures

This figure plots how the dispersion in stage-1 estimates co-varies with various quality measures. The top graph uses the interquartile range (IQR) as a dispersion measure and the bottom graph uses the interdecile range (IDR). The quality variables are team quality, reproducibility score, and the rating by peer evaluators. The IQR and IDR estimates are taken from Table 3, where relative changes are averaged across RT-hypotheses. The baseline level is the average dispersion across RT-hypotheses.

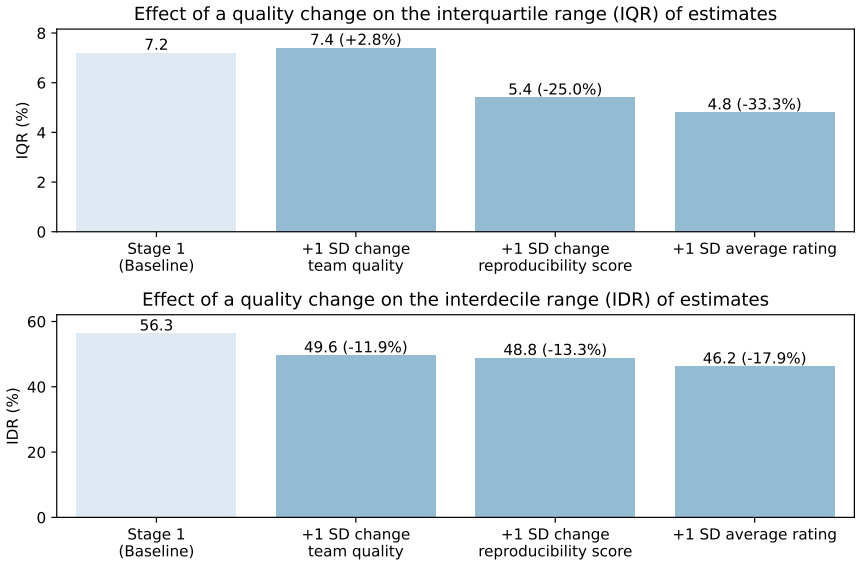


Figure 3: Dispersion in estimates related to feedback stages

This figure plots how the dispersion in estimates changes across feedback stages. Stage 1 is the baseline stage, which is the stage before any feedback. The top graph uses the interquartile range (IQR) as a dispersion measure, whereas bottom graph uses the interdecile range (IDR). The IQR and IDR values are based on the estimates in Table 4, where relative changes are averaged across all RT-hypotheses. The baseline level is the average dispersion in stage-1 estimates across RT-hypotheses.

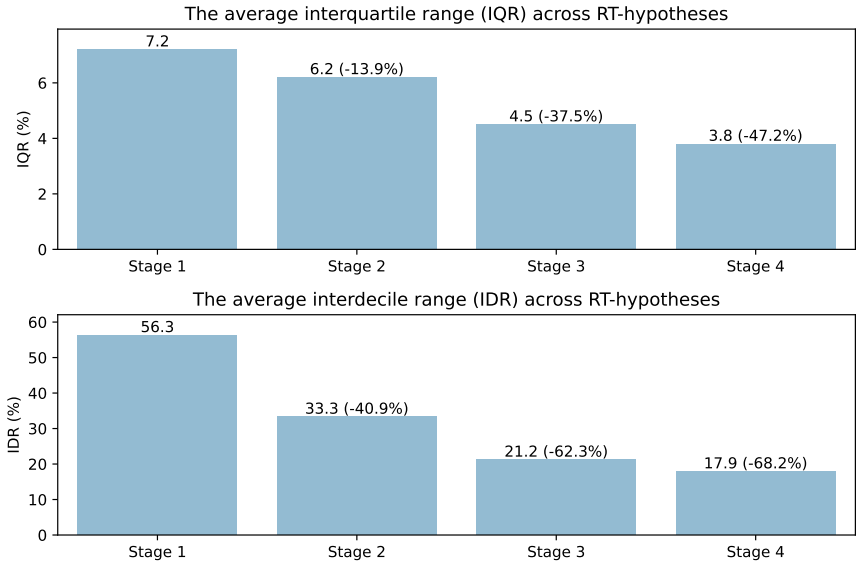


Figure 4: Research team beliefs on dispersion stage-1 estimates

This plot illustrates the dispersion in beliefs across research teams, for all six RT-hypotheses. All teams were asked to predict the SD in estimates across all RTs. The boxes depict the first and third quartile. The horizontal line in the box corresponds to the median. The whiskers depict the 2.5% and 97.5% quantile. All estimates are also plotted individually as gray dots. The red dots show the realized SD in estimates across RTs. The orange dots do the same, but are based on a 2.5%-97.5% trimmed sample.

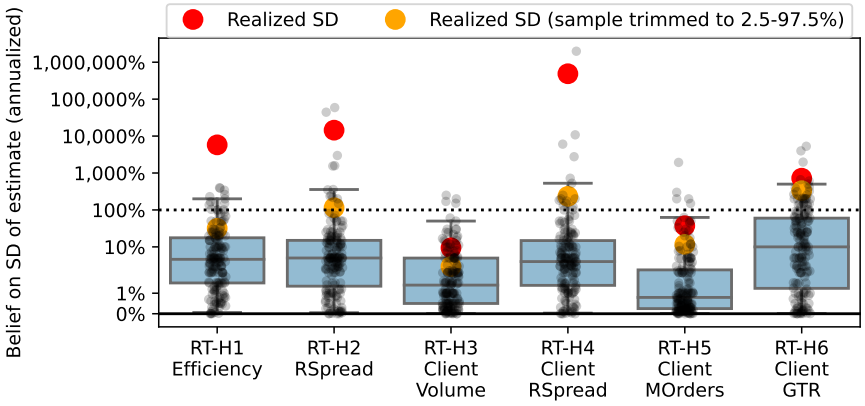


Figure 5: Dispersion in stage-1 estimates of multiverse analysis

This plot illustrates the dispersion in stage-1 estimates obtained from the multiverse analysis. The dispersion in *reported* estimates appears in gray and corresponds to panel (a) in Figure 1. The boxes depict the first and third quartile. The horizontal line in the box corresponds to the median. The whiskers depict the 2.5% and 97.5% quantile.

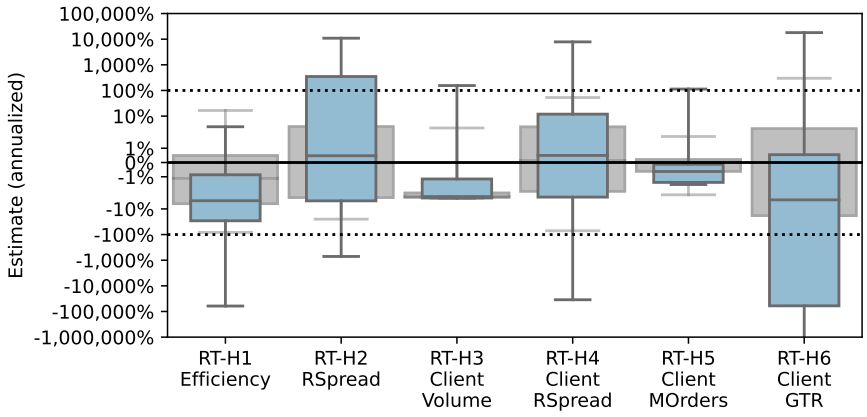


Figure 6: Fork sensitivity of estimates in multiverse analysis

This figure plots how sensitive the distribution in estimates is to the alternatives available at a fork in the multiverse analysis. The sensitivity is measured by the standardized Anderson-Darling test statistic. Higher values of the statistic imply that distributions become more dissimilar across alternatives at the fork.

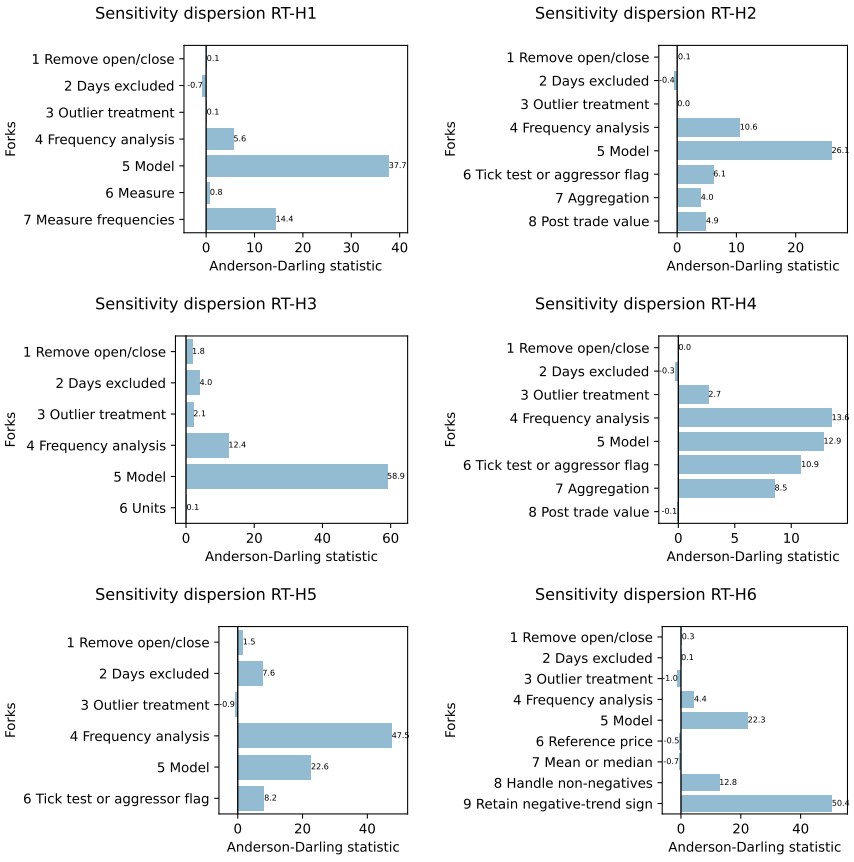


Figure 7: Sensitivity of estimates in multiverse analysis of RT-H1

This plot illustrates how the distribution of RT-H1 estimates depends on two influential forks in the multiverse analysis: (i) the model and (ii) the frequency of the analysis. Distributions are obtained by bootstrapping 1000 times from the original sample for each analysis path. To avoid clutter, the weekly frequency is dropped since it is used by only one team (out of 164).