UNIVERSITY OF CAMBRIDGE
Faculty of Economics

JANEWAY INSTITUTE CAMBRIDGE

CAMBRIDGE WORKING PAPERS IN ECONOMICS
JANEWAY INSTITUTE WORKING PAPERS

Non-Standard Errors

Albert J. Menkveld
Vrije Universiteit Amsterdam

Felix Holzmeister
University of Innsbruck

Magnus Johannesson
Stockholm School of Economics

Sebastian Neusüss
Aalto University

Utz Weitzel
Vrije Universiteit Amsterdam

Anna Dreber
Stockholm School of Economics

Juergen Huber
University of Innsbruck

Michael Kirchler
University of Innsbruck

Michael Razen
University of Innsbruck

Oliver Linton
University of Cambridge

Abstract

In statistics, samples are drawn from a population in a data generating process (DGP). Standard errors measure the uncertainty in sample 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. To study them, we let 164 teams test six hypotheses on the same sample. We find that non-standard errors are sizeable, on par with standard errors. Their size (i) co-varies only weakly with team merits, reproducibility, or peer rating, (ii) declines significantly after peer-feedback, and (iii) is underestimated by participants.

Online appendix available at https://bit.ly/3DIQKrB

Please note a full list of authors is available in the working paper

Reference Details

2182 Cambridge Working Papers in Economics
2021/12 Janeway Institute Working Paper Series

Published 25 November 2021

Key Words Market Efficiency, P-hacking, Publication bias
JEL Codes A14, C10, C12, C59, C90, G14, G40

Websites www.econ.cam.ac.uk/cwpe
www.janeway.econ.cam.ac.uk/working-papers

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], 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 23, 2021

*The first nine authors in italics are the project coordinators. They designed the project, managed it, 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. This manuscript represents the views of the authors and does not reflect the views of any of the institutions authors are affiliated with or receive financing from. The coordinators thank Lucas Saru for valuable comments. They further thank Adam Gill, Eugénie de Jong, 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, and (Menkveld) NWO-Vici.

Non-Standard Errors

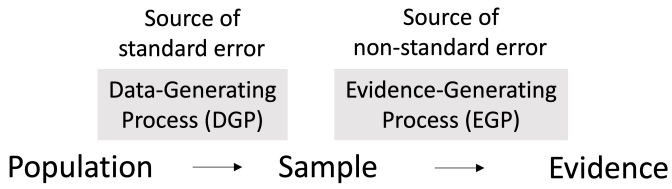
Abstract

In statistics, samples are drawn from a population in a data-generating process (DGP). Standard errors measure the uncertainty in sample 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*. To study them, we let 164 teams test six hypotheses on the same sample. We find that non-standard errors are sizeable, on par with standard errors. Their size (i) co-varies only weakly with team merits, reproducibility, or peer rating, (ii) declines significantly after peer-feedback, and (iii) is underestimated by participants.

Online appendix available at https://bit.ly/3DIQKrB.

1 Introduction

Academic research recognizes randomness in data samples by computing standard errors (SEs) for parameter estimates. It, however, does *not* recognize the randomness that is in the research process itself. We believe that such randomness is the cause of, what we will call, non-standard errors (NSEs).



The above schema depicts 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. These samples are used to estimate population parameters, with error: standard error. Using the same language, one could say that scientists collectively engage in an evidence-generating process (EGP). This EGP exposes variation across researchers, adding further error, which we refer to as *non-standard error*. Can this additional uncertainty be safely ignored, or is it sizeable? Does it decline with peer feedback? This has been our core motivation when designing the *Finance Crowd Analysis Project* (#fincap).[1]

Our working definition of non-standard errors is the standard deviation across researchers for the results they report when independently testing the same hypotheses on the same sample. For example, in #fincap we provide researchers with identical data and then ask them to independently propose a measure for market efficiency, estimate its average per-year change, and report it. The NSE for this reported change is simply the standard deviation across researchers of their reported changes.

Non-standard errors exist for a variety of reasons. They might, of course, be due to glitches in computer code or due to machine

[1] To increase the credibility of the findings and address concerns of specification searching, we wrote a pre-analysis plan (PAP) and filed it with the Open Science Foundation at https://osf.io/h82aj/. This was done *before* distributing the sample with instructions to the #fincap participants. We follow the PAP throughout unless stated otherwise.

precision. A more intriguing source of NSEs is the different routes that researchers travel through the "garden of forking paths." Gelman and Loken (2014) use this metaphor to describe the non-trivial set of choices that researchers have to make when generating evidence. For example, they have to specify an appropriate econometric model, they have to pre-process the sample to ready it for estimation (e.g., purge outliers), they have to pick a programming language, et cetera. Error is therefore to be understood in the sense of erratic rather than erroneous.

Our objective is to measure and explain the size of non-standard errors. Are they large? And, how large are they compared to standard errors?[2] The four questions that we focus on are the following:

1. How large are non-standard errors for research 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 more easily reproducible results?
   (c) for papers that score higher in peer evaluations?
3. Does peer feedback reduce these non-standard errors?
4. Are researchers accurately aware of the size of non-standard errors?

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

We believe that #fincap is close to this "ideal experiment" for three main reasons. First, Deutsche Börse supported the project by offering exclusive access to 720 million trade records spanning 17

[2]Note that the relative size of NSEs in total uncertainty is likely to grow in view of the trend of ever larger datasets. NSEs are invariant to sample size whereas SEs decline with sample size.

years of trading in Europe's most actively traded instrument: the EuroStoxx 50 index futures. It enables researchers to test important hypotheses on how the market changes when migrating to superhuman speeds. Second, #fincap being the first crowd-sourced empirical paper in finance might have pushed the hesitant few over the line (in addition to us arguing in the invitation that this is an opportunity to forego future regret).[3] 164 research teams and 34 peer evaluators participated in #fincap. A back-of-the-envelope calculation shows that this effort alone cost 27 full-time equivalent person years (164×2 months + 34×2 days $\approx$ 27 years). Third, we believe that #fincap participants did, in fact, exert serious effort for a variety of reasons, such as, strong on-time delivery statistics and a high average rating of papers (more in Section 3.3).

Summary of our findings. First off, we show that the group of #fincap participants is representative of the academic community that studies empirical finance/liquidity. The hypotheses to be tested on the Deutsche Börse sample are tailored to this group. To distinguish them from the hypotheses we test in our (meta) study, we henceforth refer to them as RT-hypotheses. About a third of the 164 research teams (RTs) have at least one member with publications in the top-three finance or the top-five economics journals.[4] For the group of peer evaluators (PEs), 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 research is 8.1 for RTs and 8.4 for PEs. For experience with market-liquidity research, it is 6.9 for RTs and 7.8 for PEs.

The dispersion in results across research teams is sizeable. All six 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 across-RT dispersion in estimates is enormous, but is mostly due to

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

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

the presence of extreme values. One RT reported a 74,491% increase, which explains why the simple across-RT mean and standard deviation (SD) are 446.3% and 5,817.5%, respectively. If, however, following a standard practice in finance,[5] outliers are treated by winsorizing the sample using 2.5 and 97.5 percentiles, the mean and SD become -7.5% and 20.6%, respectively.

An average 7.5% per-year decline in efficiency from 2002 to 2018 for a leading exchange, is remarkable. The really remarkable result given *our* purposes, however, is the across-RT dispersion. An SD of 20.6% for a *winsorized* sample is striking. But, how does this non-standard error compare to the standard error? The across-RT mean SE for the winsorized sample is 13.2%. The NSE-SE ratio therefore is 1.6. The NSE, therefore, is non-trivial in a relative sense as well. The NSE-SE ratio ranges between 0.6 and 2.1 across RT-hypotheses, including some that were deliberately non-fuzzy (e.g., one on the share of client volume in total volume). Not surprisingly, the non-fuzzy ones tend to show lower NSE-SE ratios.

We further find that the size of non-standard errors is hard to explain in the cross-section of researchers. We formally test three hypotheses that relate team quality, work-flow quality, and paper quality to dispersion, or more specifically, the size of "errors." These errors are defined as the difference between a particular RT result, and the across-RT mean result. A standard approach to model heteroskedasticity of errors regresses log squared error on explanatory factors (Harvey, 1976). Although almost all coefficients are negative consistent with a high-quality small-error association, none of them are statistically significant.[6] The only exceptions are team quality and work-flow quality in the 2.5%-97.5% winsorized sample. The latter is proxied by verifying how easy it is to reproduce the results of a particular RT with the code that the RT provided. For work-flow quality, we find a statistically significant negative sign. For team quality we find

[5] Adams et al. (2019), for example, review how outliers are dealt with for all ordinary least-squares (OLS) regressions in articles that appeared in premier finance journals in the period from 2008 through 2017: *Journal of Finance*, *Journal of Financial Economics*, *Review of Financial Studies*, and *Journal of Financial and Quantitative Analysis*. We like to note that outlier treatment was not pre-registered.

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

only a suggestive (weakly significant) negative sign. A one standard-deviation increase in work-flow quality reduces non-standard error by 12%. For team quality, a one SD increase reduces it by 8%.

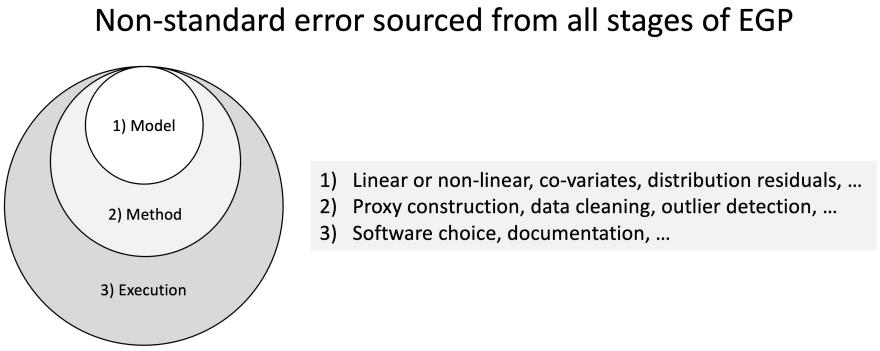
Peer feedback significantly reduces non-standard errors. The peer-feedback process involves multiple stages. By and large, we find that each stage reduces NSEs. The overall reduction across all four stages is 8.5% for the raw sample, and 53.5% for the 2.5%-97.5% winsorized sample. This stark difference is driven by extreme-result teams largely staying put in the feedback process.

RTs mostly underestimate non-standard errors as tested in an incentivized belief survey. The extent to which the average team underestimates the realized across-RT standard deviation ranges between 9.0% and 99.5% across the six hypotheses. It is not merely due to the outliers that they did not foresee as the vast majority also underestimated the NSE for the 2.5%-97.5% winsorized sample. Such underestimation might well be the reason why non-standard errors never attracted much attention, until recently.

These findings on dispersion in percentage-change estimates carry over to the associated *t*-values. In a sense, *t*-values represent the statistical strength of a positive, or a negative percentage-change finding. The non-standard error of *t*-values therefore measures the extent of agreement across RTs on statistical strength. Our finding of non-trivial NSEs also for *t*-values therefore implies large dispersion in findings of statistical strength. Most telling, perhaps, is that for the first RT-hypothesis on efficiency, 23.8% of RTs find a significant decline, 8.5% find a significant increase, and 67.7% find no significance if their *t*-values are evaluated at a conventional significance level (in finance) of 5%. We find a similar pattern for all other RT-hypotheses.

Finally, we like to point out that the sizeable non-standard errors are *not* driven by the presence of poor-quality results. This could already be concluded from the very weak traction obtained from quality variables in the error-size regressions (discussed above). To further corroborate this point, we reduce the sample to research teams who score high on *all* quality variables. Doing so yields a sample of only nine out of 164 teams. Non-standard errors for this subsample remain large, including implied *t*-values that for some RTs are above 1.96 and for others are below -1.96 for the *same* RT-hypothesis.

Contribution to the literature. The issue of variability in the research process is not new.[7] The extant literature is best discussed by adding some structure. The following chart depicts the various ways in which the existing literature touches on non-standard errors.



1. *Model.* One source of variability across researchers is model specification. Is the empirical model linear or non-linear? Which co-variates should be included? Et cetera. Economists are cognizant of specification errors and they typically do *robustness analysis* to show that their main results do not critically depend on specification choices. They have even endogenized the awareness specification error in their theoretical models. Agents in these models are assumed to make decisions in full awareness of potential model-specification errors (e.g., Hansen and Sargent, 2001, for economics or Garlappi, Uppal, and Wang, 2006, for finance).

2. *Method.* An additional source of variability is in the empirical method. What are reasonable empirical proxies for the parameters of interest? What filters should be applied to the sample (e.g., how to deal with outliers)? What are appropriate test statistics? It has long been known that not only is this a source of variability, it can actually produce misleading inference. In finance, Lo and MacKinlay (1990) were the first to warn of "data snooping" in which properties of the data are used to construct test statistics. More recently, Mitton (2021) provides di-

[7] 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."

rect evidence on substantial methodological variation by analyzing 604 empirical corporate-finance studies published in the top-three finance journals (mentioned in footnote 4).

3. *Execution.* Variability remains, *even* if one fixes model and method. This is the domain of computational reproducibility studies that go far back in economics and finance. In economics, the *American Economic Review* (AER) was the first to introduce a *Data Availability Policy* in 2005 after the AER had published two studies that illustrated how hard it was to reproduce empirical studies (Dewald, Thursby, and Anderson, 1986; McCullough and Vinod, 2003). Reproducibility issues persisted as, *at most*, half of the empirical studies published in the top economics journals could be computationally reproduced (Glandon, 2011; Chang and Li, 2017; Gertler, Galiani, and Romero, 2018). Reproducibility has become an issue for finance journals as well as exemplified by the *Journal of Finance* retracting a published article for the first time in its history (Rampini, Viswanathan, and Vuillemey, 2021).

The presence of sizeable non-standard errors potentially corrupts the research process through a phenomenon known as "p-hacking." The p-value of a statistical test refers to the probability (p) that the reported effect is solely due to chance (under the null hypothesis of there being no effect). Large non-standard errors create opportunities for researchers to pick a model, a method, and an execution that produces a low p-value and write a report accordingly (i.e., they are hacking the p-value).

Replication studies typically find much weaker effects and less statistical strength,[8] suggestive of p-hacking. We caution that these results are not necessarily due to ill intent on the side of researchers, because they might be demand- rather than supply-driven. That is, journals might selectively publish papers with low p-values. Munafò

[8] See, for example, Open Science Collaboration (2015) and Camerer et al. (2016, 2018). Replication is an issue in finance as well (Harvey, 2017) and replication studies started to appear recently. Hou, Xue, and Zhang (2018), for example, study 452 asset-pricing anomalies, including "the bulk of published anomalies in finance and accounting" (p. 2024). They conclude that most anomalies fail to replicate. Black et al. (2021) try to replicate the results of four top finance/accounting/economics publications. They were not able to reproduce any of them.

et al. (2017) survey the various threats to credible empirical science and propose several fixes.

Our study on the magnitude of non-standard errors, therefore, not only measures the additional uncertainty in empirical results, it also reveals the scope for *p*-hacking. #fincap was carefully designed to minimize any *p*-hacking dynamics so as to obtain unbiased estimates of NSEs. For example, researchers signed contracts that clarified that their results would remain anonymous in all public communication about #fincap.

We are the first to study how large these errors are in finance, but not the first in science. Silberzahn et al. (2018) pioneered the multi-analyst study by letting multiple research teams test whether soccer referees are more likely to draw red cards for players with 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. Our contribution relative to these studies is that we not only document the size of NSEs,[9] we also try to explain them in the cross-section, and to study convergence in a peer-feedback process. Furthermore, we ascertain whether research teams themselves are accurately aware of the size of NSEs. A further strength of our study is the large sample size: N=164. It is at least twice and up to 20 times larger than 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 also presents the hypotheses that we will test and how we test them. Section 3 presents our results and Section 4 concludes.

2 Project design and hypotheses

This section starts with presenting the details of #fincap. With this in mind, we then translate our overall objectives into testable hypotheses.

[9] We like to note that documenting the size of NSEs has been done before, but, to the best of our knowledge, only in Huntington-Klein et al. (2021). Their sample size is seven RTs, ours is 164 RTs.

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

2.1 Project design

The core of #fincap is letting multiple research teams (RTs) independently test the same six hypotheses on the same Deutsche Börse sample. We refer to these hypotheses as RT-hypotheses and to this sample as RT-sample. This is to avoid potential confusion with the hypotheses that *we* will test based on sample generated by RTs and peer evaluators (PEs).[11]

The RT-sample is a plain-vanilla trade dataset for the EuroStoxx 50 index futures with, added to it, an agency/principal flag.[12] For each buy and sell, we therefore know whether the exchange member traded for its 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).

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

RT-H1 market efficiency,

RT-H2 the realized bid-ask spread,

RT-H3 the share of client volume in total volume,

RT-H4 the realized spread on client orders,

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

RT-H6 the gross trading revenue of clients.

Appendix A discusses these RT-hypotheses in detail. For the purpose of our analysis, it suffices to know that these market

[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. More details on the sample are in Figure OA.1 of the Online Appendix.

characteristics should not surprise a researcher familiar with empirical-finance/liquidity. There is, purposefully, considerable variation across RT-hypotheses in the level of abstraction. Or, in other words, the garden of forking paths is much larger for some than for others. RT-H1, for example, is on the relatively abstract notion of market efficiency. RT-H3, on the other hand, is on the market share of client volume. Such share should be a relatively straightforward to calculate because, in the RT-sample, each buy and sell trade is flagged client or proprietary.

RTs are asked to test these RT-hypotheses by estimating an average yearly change for a self-proposed measure.[13] In the interest of readability, we refer to this estimate as the "effect size." 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. Collectively, we refer to these results as RT results.

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 was evaluated twice, and each PE evaluated nine or ten papers. PEs scored the papers in single-blinded process: PEs saw the names of RTs but not vice versa. This was made clear to all ex-ante.[14] PEs scored the papers both at the RT-hypothesis level, and at the aggregate, paper level. They motivated their scores in a feedback form and were encouraged to add constructive feedback. RTs received this feedback unabridged and were allowed to update their results based on it. Importantly, the design of #fincap was common knowledge to all as it was communicated ex-ante via a dedicated website (see footnote 3). More specifically, #fincap consisted of the following four stages:

Stage 1 (Jan 11 - Mar 23, 2021.) RTs receive the detailed instructions along with access to the RT-sample. They conduct their

[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 effect size estimates (and the corresponding standard errors); research teams are not required, however, to consider only annualized data."

[14] We picked a single-blinded process instead of double-blinded process in order to incentivize RTs to exercise maximum effort.

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., effect sizes 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.

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 (e.g., feedback from colleagues over lunch or at the water cooler, during seminars, or in the coffee breaks after conferences, but also hearing about and seeing similar papers endorsed by others). The dynamics at play in a refereeing process at a scientific journal are out of scope.[16]

[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 the Cascad. For more information on Cascad, see the statement of H2 in Section 2.2.

[16] Testing the dynamics in a refereeing process requires a different experiment that involves "publishing" papers, *including* the names of the authors. Note that

2.2 Hypotheses

The overall objectives outlined in the introduction were translated into a set of pre-registered hypotheses (see footnote 1 for pre-registration details). The hypotheses all center on across-RT dispersion. We use variance as our dispersion measure because it is additive and the *de facto* standard dispersion measure in statistics. Let us therefore define dispersion at stage t for RT-hypothesis j as:

$$\text{var}\left(y_{jt}\right) = \frac{1}{n-1}\sum\left(y_{ijt} - \bar{y}_j\right)^2, \tag{1}$$

where

- $i \in \{1, \ldots, n\}$ indexes RTs,
- $j \in \{1, \ldots, 6\}$ indexes RT-hypotheses,
- $t \in \{1, \ldots, 4\}$ indexes group stages, and

where $\bar{y}_j$ is the overall average for RT-hypothesis j. To study the extent of dispersion requires modeling variance, which is known in econometrics as modeling heteroskedasticity. The idea is that the size of "errors" are related to co-variates. In our application these errors are defined as:[17]

$$\hat{u}_{ijt} = y_{ijt} - \bar{y}_j. \tag{2}$$

With all the groundwork done, we can now translate our overall objectives into three sets of hypotheses. These hypotheses are tested for the dispersion in two types of RT-results:

---

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." In other words, a refereeing system has two components. The first is anonymous feedback from peers/experts and the second is an incentive to revise according to this feedback in order for one's paper to become published. We capture the first component by asking PEs to rate the paper, but not the second one.

[17] The term error is used in a statistical sense of how observed data differ from a population mean. It should *not* be confused with errors necessarily being mistakes. Jumping ahead a little, we will model heteroskedasticity following Harvey (1976) (Section 2.3). $\hat{u}_{ijt}$ in (2) denotes the OLS residual that results from estimating Harvey (1976, Equation (2)) with only intercepts pertaining to the six dummies for the RT-hypotheses.

- The effect size (i.e., the point estimate).
- The t-value (i.e., the point estimate divided by the standard error).

We believe that both are of interest. The effect size is of intrinsic interest as it is about the outcome in the variable of interest. The t-value captures the statistical strength of the reported effect in the sense larger magnitudes are less likely to occur under the null of there being no effect. All hypotheses on the dispersion in effect sizes and t-values are stated as null hypotheses. The hypotheses tests will be two-sided.

The first set of three hypotheses focuses on whether error variance relates to various quality measures:

H1 Team quality does not explain the size of errors in Stage 1 results. 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 standardized series that include:[18]

(a) *Top publications*: The team has at least one top-five publication in economics or top-three publication in finance (0/1) (see footnote 4).

(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 datasets that are at least as big as the Deutsche Börse dataset analyzed 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).

[18] 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 PCA results will be discussed in Section 3.2.1.

H2 Work flow quality does not co-vary with the size of errors in Stage 1 results. We proxy for work-flow quality with a reproducibility score per RT × RT-hypothesis. The score measures to what extent an RT result is reproducible from the 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 institutions. The goal of this agency is to provide researchers with an innovative tool allowing them to signal the reproducibility of their research (used by, for example, the *American Economic Review*).[19]

H3 Paper quality as judged by the average score of PEs (hypothesis level) does not explain the size of errors in the first submission (scale from 0 to 10). To remove a possible PE fixed effect we use demeaned PE scores in all of our analysis.

The second set of hypotheses are about convergence of RT results across the four stages, and about convergence from the first to the final stage to have a test with maximum power.[20]

H4 The error variance does not change from first to second stage.

H5 The error variance does not change from second to third stage.

H6 The error variance does not change from third to fourth stage.

H7 The error variance does not change from first to fourth stage.

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

H8 The average belief of RTs on the across-RT dispersion in results, is correct. The dispersion beliefs are solicited in terms of the standard deviation measure.

[19]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.

[20]The decline across consecutive stages might be small, but after summing them the total decline might be sizeable. It is in this sense that the first-to-final stage is expected to be statistically most powerful.

2.3 Statistical approach

Heterogeneity in error variance is known as heteroskedasticity in econometrics. A standard textbook in econometrics, Greene (2007, Ch. 9.7), refers to Harvey (1976) as a common approach to modeling heteroskedasticity. We adapt this approach to our setting and follow Harvey's notation for ease of comparison. Error variance is modeled as:

$$\sigma_{ijt}^2 = \text{var}\left(u_{ijt}\right) = \exp\left(z_{ijt}'\alpha\right), \tag{3}$$

where u_{ijt} is the unobserved disturbance term, and z_{ijt} contains the set co-variates which, instead of an intercept, includes six dummies corresponding to the six RT-hypotheses to account for possible heteroskedasticity across hypotheses. The parameter vector α measures the marginal *relative* change in variance when the associated co-variate is increased by one unit. For example, if the coefficient for co-variate X is 0.1, then increasing X by one unit raises variance by ten percent. It raises standard deviation by approximately half that amount: five percent.[21]

Estimation. An attractive feature of the Harvey model is that it can be estimated with ordinary least squares (OLS). More specifically, the OLS model is:[22]

$$\log \hat{u}_{ijt}^2 = z_{ijt}'\alpha + w_{ijt}. \tag{4}$$

Let $\tilde{\alpha}$ denote the resulting parameter estimate. Harvey shows that $\tilde{\alpha}$ estimates α consistently, except for the intercept coefficients. If,

[21] This follows directly from a first-order Taylor approximation of $f(x) = \sqrt{x}$ around μ: $f(x) \approx \sqrt{\mu} + \frac{1}{2\sqrt{\mu}}(x - \mu)$. If the predicted change in variance is ten percent, then it follows that $f(x) \approx \sqrt{\mu} + \frac{1}{2\sqrt{\mu}} 0.10\mu = 1.05\sqrt{\mu}$.

[22] It is tempting to drop the natural logarithm on the dependent variable and simply use squared errors as the dependent variable in an OLS regression. Harvey (1976, p.6) mentions three reasons for why his model is more attractive than this alternative one: "Firstly, the likelihood function is bounded and no problems arise due to estimated variances being negative or zero. Secondly, the error terms in the two-step equation [...] are (asymptotically) homoscedastic and so the estimated covariance matrix of the two-step estimator, $\tilde{\alpha}$, is consistent. Finally, the likelihood ratio test has a much simpler form in the multiplicative model."

however, w_{ijt} is assumed to be Gaussian, then a consistent estimator for the full α vector becomes (where the subscript k refers the kth element of the vector):

$$\hat{\alpha}_k = \begin{cases} \tilde{\alpha}_k & \text{for } k > 6, \\ \tilde{\alpha}_k - \psi\left(\frac{1}{2}\right) + \log\left(\frac{1}{2}\right) & \text{for } k \leq 6, \end{cases} \tag{5}$$

where $\psi(\nu)$ is the psi (digamma) function defined as $d \log \Gamma(\nu) / d\nu$ where $\Gamma(\nu)$ is the Gamma function with ν degrees of freedom. This follows from Harvey (1976, Equation (6)). The predicted variance for RT i's error for RT-hypothesis j in stage t becomes:

$$\hat{\sigma}^2_{ijt} = \exp\left(z'_{ijt}\hat{\alpha}\right), \tag{6}$$

Importantly, the non-intercept elements of the parameter vector α are estimated consistently, even if w_{ijt} is non-Gaussian. The tests of our hypotheses all pertain to non-intercept elements, and therefore remain valid in the absence of normality.

Statistical inference. All hypotheses are tested using standard errors that cluster residuals w_{ijt} by RTs.[23] The R-squared of the heteroskedasticity regression in (4) is a useful measure of how much of the total dispersion can be explained. More precisely, the R-squared captures how much of the variance of log squared errors is explained by the set of explanatory variables.

Balanced sample. We will implement our tests on a balanced sample. We therefore include only observations for RTs that participated in *all* of the four stages. This makes the results in all four stages comparable as they pertain to the same set of RTs. The balanced sample, however, is not that different from the unbalanced one as only four of the 168 RTs did not complete all rounds. The balanced sample therefore consists of 164 RTs.

[23] The clustering is needed to account for possible non-zero correlation in residuals w_{ijt} per RT, across RT-hypotheses and stages. For example, if the error is large for a particular RT on a particular RT-hypothesis, then it is likely to be large for the other RT-hypotheses as well. We do not cluster on RT-hypotheses because the model in (4) includes fixed effects for RT-hypotheses, which, we believe, removes most of the commonality. We do not cluster on PEs as a possible PE fixed effect was removed by demeaning PE scores in all of our analysis.

Significance levels. The hypothesis tests are two-sided and tested at significance levels of 0.005 and 0.05. Following the recommendations of Benjamin et al. (2018), we refer to results with a p-value smaller than 0.005 as statistically significant. Results with a p-value smaller than 0.05 are referred to as suggestive (weakly significant) evidence.

3 Results

This section summarizes all our findings. The first subsection presents various summary statistics to familiarize with the sample. The second subsection presents and discusses the test results for our hypotheses.

3.1 Summary statistics

(Insert Table 1 about here.)

Table 1 summarizes our sample by means of three sets of statistics, organized in three panels. Panel (a) summarizes the qualities of the #fincap community. It consists of 164 research teams (RTs) and 34 peer evaluators (PEs). The 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 4 for the list of journals). For PEs this is 85%. Those who provide feedback are therefore better published, which probably mirrors reality as the feedback flow is most likely to come from more senior or successful scholars. This finding is echoed in the percentage of RTs that have at least one member who is tenured at the associate or full professor level. This fraction is 52% for RTs and 88% for PEs.

(Insert Figure 1 about here.)

Figure 1 illustrates how RT members and PEs cover the global academic-finance community reasonably well. RT members reside in 34 countries with most (51 out of 293) residing in the United States. PEs reside in 13 countries with, again, most (13 out of 34) residing in the United States. The stronger skew towards the Unites States (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 in market-liquidity research these average scores are 6.9 for RTs and 7.8 for PEs. There is considerable variation around these averages as the across-RT standard deviation for these scores range from 1.7 to 2.4. When it comes to working with the large RT-sample (720 million trade records), again, most RTs and PEs seem up to it. 52% of RTs have worked with samples of similar size or larger. For PEs, this percentage is 88%.

Panel (b) of Table 1 shows that the average quality of the RT analysis is solid, and the dispersion is large. The average Cascad reproducibility score is 64.5 on a scale from 0 (low) to 100 (see footnote 19). This is high when benchmarked against other studies on reproducibility surveyed in Colliard, Hurlin, and Pérignon (2021). The across-RT standard deviation is 43.7, which implies extreme variation across RTs (most code either reproduces perfectly or not at all). The paper-quality score of PEs show a similar pattern, albeit with considerably lower dispersion. The average score across RTs is 6.2 on a scale from 0 (low) to 10. The across-RT standard deviation is 2.0.

Panel (c) provides descriptive statistics on non-standard errors: the dispersion in results across RTs. It does so, by hypothesis, and by type of result: Estimate of the effect size, standard error, and t-value. Our focus is on dispersion across RTs, which is why we relegate a discussion of RT means to Appendix A. 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 mean instead of the across-RT standard deviation.

(Insert Figure 2 about here.)

Perhaps the most salient feature of the extensive Panel (c) is that there is substantial across-RT variation in all hypotheses and for all results. For RT-H1 on efficiency, for example, the across-RT mean annual change is 446.3% with an across-RT standard deviation (NSE) of 5,817.5%. These extraordinary numbers are intimately linked to an extremely large value of 74,491.1% for a particular RT. If, as is common in finance (Adams et al., 2019), the RT-sample is winsorized

at a 2.5%-97.5% level (which is the default level for the remainder of this text), then the mean becomes -7.4% with standard deviation of 20.6%. Note that such NSE of 20.6% is similar in magnitude to the mean reported SE of 13.2%. The NSE-SE ratio therefore is 1.6. The dispersion is not much lower for RT-H3, where the corresponding numbers are a mean change is -2.6% with an NSE of 1.4% and a mean SE of 1.3%, leading to an NSE-SE ratio of 1.3. This pattern emerges for all RT-hypotheses with NSE-SE ratios ranging from 0.6 to 2.1. Figure 2 illustrates the substantial across-RT dispersion with boxplots.

The panel further shows that *t*-values also exhibit sizeable dispersion across RTs. For RT-H1, for example, the mean and standard devation for *t*-value across RTs is -3.6 and 28.4, respectively. For the winsorized sample, these values are -1.4 and 5.2, respectively. At conventional threshold levels for 95% significance, -1.96 and 1.96, almost a third of RTs find statistically significant results. Interestingly, there is no agreement in this group of RTs as the 32.3% decomposes into 23.8% who find a significant decline and 8.5% who find a significant increase. This pattern is common across RT-hypotheses. Even for the relatively straightforward RT-H3, 49.4% report an (implied) *t*-value that is larger than 1.96 in absolute value. There is more agreement on the sign of the significance, as 45.7% finds significant decline, whereas 3.7% finds a significant increase. The 3.7% is not a single RT, it corresponds to *six* RTs who find a significant increase whereas 75 RTs (45.7%) find a significant decline.

Overall, the summary statistics show that there is substantial dispersion across RTs at various levels. There is substantial dispersion in team quality (panel (a)), the quality of their analysis (panel (b)), and in the results they report (panel (c)). These findings are promising for the hypotheses tests that we will turn to next. The sizeable dispersion in results implies that NSEs are large and therefore worthy of study. The strong variation in both team quality and analysis quality creates the statistical power needed to test whether they are significant covariates for the heteroskedasticity in results (H1-3). Are errors larger in magnitude for lower quality teams or for teams with lower quality analysis? This is what will be tested in the next subsection.

3.2 Hypotheses tests

The results on the three sets of hypotheses are discussed in the next three subsections.

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

(Insert Table 2 about here.)

The first set of hypotheses aims to measure whether various quality variables significantly co-vary with the size of errors. The first hypothesis centers on RT quality, which we measure by picking the first principal component of five standardized proxies. Table 2 summarizes the principal component analysis (PCA). Panel (a) presents the simple correlation matrix that enters the PCA. Note that this matrix is the covariance matrix since the five proxies are standardized. It is reassuring that the lion share of these correlations are positive. The only exception is that big-data experience is negatively related to experience. Since both are positively correlated with the rest, this finding suggests that the sample could be cut in two subgroups: those who are relatively more experienced in the field and those who are relatively more experienced in big-data analysis.

Panel (b) and (c) document that the largest principal component explains 38.3% of all variance and loads positively on all variables. It loads strongest on publications and weakest on big-data but, importantly, it loads positively on all of them. This is the principal component that serves as the team-quality measure in hypothesis test. Noteworthy is that the second principal component picks up 23.6/(100-38.3)=38.2% of the remaining variance. The strongest loadings are 0.79 on big-data and -0.55 on experience. This component neatly picks up the experience/big-data distinction in the group of RTs that we commented on earlier.

(Insert Table 3 about here.)

Table 3 summarizes our findings when regressing log-squared errors on the various quality co-variates. These regressions were done for the raw sample as pre-registered in the PAP. We, however, had not foreseen the types of extreme values that some RTs reported (as evidenced, for example, by one RT reporting an average per-year increase in efficiency of 74,491.1%). Following standard practice in

finance, we treat these "outliers" by winsoring and trimming the sample with cut-off levels at the 1% and 99% quantile and at the 2.5% and 97.5% quantile (Adams et al., 2019). Winsorizing replaces values beyond these levels by the quantile level, whereas trimming simply removes them.

For effect-size estimates, the regression results show that log-squared errors seem unrelated to the various quality co-variates, for the most part. Most coefficients are negative, and thus consistent with higher quality being associated with smaller errors. Loosely speaking, higher quality is in the center of the distribution of RT results and lower quality in the periphery. There is, however, no statistical significance with two exceptions for the 2.5%-97.5% winsorized sample.

First, we find suggestive evidence that higher team quality coincides with smaller errors. A one standard-deviation (SD) increase in team quality coincides with a 16% drop in error-variance, which, by approximation,[24] implies an SD drop of 0.5×16=8%.

Second, we find significant evidence that better reproducibility is associated with smaller errors. A one-SD increase in reproducibility coincides with an approximate 12% drop in SD. Phrased differently, it coincides with a 12% drop in NSE. The findings for dispersion in t-values are similar with, again, statistical significance only for reproducibility, with higher reproducibility being associated with smaller errors.

In summary, the evidence is such that the first three hypotheses cannot be firmly rejected for both effect-size estimates and t-values. In other words, team quality (H1), reproducibility (H2), and paper quality (H3) seem unrelated to the size of errors. The only exception is that for the 2.5%-97.5% winsorized sample we do reject the null of no effect for H1 and H2.

(Insert Figure 3 about here.)

The almost absent statistical significance is unlikely to be due to low statistical power. The sample contains almost one thousand observations: 164 RTs times six hypotheses yields 984 observations. The result suggests that NSEs are sizeable, *also* for high-quality teams and/or high-quality results. The large sample allows us to make this

[24] This approximation is derived in footnote 21.

point in a more straightforward way: Pull out a sub-sample of RTs that score high on all quality measures. We note that this additional analysis was not pre-registered.

Figure 3 illustrates the result. Nine RTs score highest on all measures: all RT-quality proxies, reproducibility, and the average peer-evaluator rating of their paper.[25] The figure shows that NSEs remain large. For the effect-size estimates depicted in the top plot, the NSE for RT-H1 is 10.1% (reported on top of the plot) and for RT-H3 it is 0.4%. For the winsorized full sample, these values were 20.6% and 1.8%, respectively (see Table 1). The bottom plot shows that also for t-values the across-RT dispersion is large. It is so large, in fact, that for RT-H4, two RTs report values larger than 1.96 in magnitude, but with *opposite* signs. If interpreted conventionally, the one RT would conclude that the realized spread on client order increased significantly in the course of the sample, whereas the other RT would conclude it significantly decreased.

3.2.2 Convergence across stages? (H4-7)

(Insert Table 4 about here.)

Whereas initial dispersion is large and seemingly only weakly related to quality, the dispersion does decline with peer feedback. Table 4 shows that error variance declines significantly when comparing the first to the last stage.

(Insert Figure 4 about here.)

Panel (a) presents the results for the effect-size estimates. In the raw sample, error variance declines by 17% and the standard deviation, therefore, by approximately half that amount: 8.5%. Figure 4 illustrates this finding by showing box plots for all four stages. One salient pattern in the figure is that extreme values seem to stay put or move only marginally. Most of the convergence seems to be in the boxes that depict the interquartile range. This observation explains why for

[25] For the binary variables, the conditioning is trivial. For non-binary variables, we picked 7.5 for the scales from zero to 10, and 75 for the scales from 0 to 100.

the winsorized and trimmed samples the convergence across stages is much stronger.[26]

For the winsorized sample, the standard deviation across RT estimates decline by approximately 53.5%. The decline is distributed rather evenly across the three stages. After receiving written feedback on their paper by two PEs in the second stage, it declines by 14.5%. It declines by another 20% in the third stage after RTs see the best five papers. There is a final 19% reduction in the final stage when RTs hand in their final estimates, this time not constrained by having to hand in the code to support it. All these changes are statistically significant, except for first change where the significance is weak and the result is therefore only suggestive.

The pattern for t-values is somewhat different for this winsorized sample. The SD also declines significantly across all stages. The decline, however, is smaller in magnitude: 20.5% instead of 53.5%. The most salient difference is that the SD *increases* significantly by 14% from Stage 1 to Stage 2. The subsequent decline of 4% is insignificant. The real convergence is in the final stage where it drops by 30.5%. This maybe due to the nature of this final stage where RTs might simply be Bayesian updating with less than full weight on their own paper. This necessarily leads to convergence across RTs.

This evidence makes us reject most of the null hypotheses on convergence. H7 is most firmly rejected. For both effect-size estimates and t-values do we find rejection in all stages in favor of the alternative that error variance declines. The other three hypotheses that pertain to changes across the three consecutive stages, the evidence is weaker but largely rejecting the null of no changes.

3.2.3 Are RT-beliefs about dispersion accurate? (H8)

The eighth and final hypothesis is on whether RTs are accurately aware of the size of non-standard errors: the standard deviation of results across RTs. Beliefs were solicited in an incentivized way. If their beliefs turn out to be within 50% of the realized standard devi-

[26] The winsorization and trimming was done per hypothesis, per stage. To retain a balanced panel, the trimming procedure removes an RT for all stages in case it was removed in a single stage. The balanced-panel condition is particularly important when studying convergence across stages, because one wants to compare a fixed set of RTs across stages.

ation, they can earn a monetary reward of $300. The details of the reward scheme are in the instruction sheet they obtained prior to reporting their beliefs (see Figure OA.10 in the Online Appendix). The hypothesis pertains to Stage 1 results because they were solicited only for this stage.

As H8 is stated in terms of the average belief being correct, testing it requires a test on the equality of means: the mean belief about standard deviation in results across RTs and the standard deviation of these results in the population. Let us define the test statistic D that measures relative distance between beliefs and realizations:

$$D = \frac{1}{6n} \sum_{i,j} \left(\frac{BeliefOnStDev_{ij} - RealizationOfStDev_j}{RealizationOfStDev_j} \right), \qquad (7)$$

where $BeliefOnStDev_{ij}$ is the belief of team i on the standard deviation across RTs for hypothesis j and $RealizationOfStDev_j$ is the realized standard deviation for this 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 C.

(Insert Table 5 about here.)

Table 5 presents the test results which show that non-standard errors are severely underestimated. Effect-size estimates are significantly underestimated by 71.7% and t-values by a significant 70.6%. Redoing the test by RT-hypothesis shows that there is strong heterogeneity but, where significant, there is underestimation. For effect-size estimates, it seems that for the RT-hypotheses that are relatively straightforward to test, the null of accurate beliefs cannot be rejected. An example is RT-H3, which is on the market share of client volume. The average RT-belief for this RT-hypothesis is not significantly different from the realization. For RT-hypotheses that require more creative effort to test, NSEs are significantly underestimated. For RT-H1 on market efficiency, for example, the underestimation is *99.5%*. In

[27] The benefit of a relative measure as opposed to an absolute measure 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 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).

summary, the vast majority of tests show significant underestimation and we therefore firmly reject the null of an accurate average belief.

(Insert Figure 5 about here.)

Figure 5 plots the entire distribution of reported beliefs along with the realized values that are depicted by red dots. It illustrates that the vast majority of RTs underestimated dispersion. The interquartile range denoted by the boxes is consistently below the red dot. One might think that RTs simply overlooked the extreme values that prop up SDs. This, however, does not seem to be true since even if one trims the RT-results (which is stronger than winsorizing in this case), the boxes stay below the realized value which, for the trimmed sample, are depicted by orange dots. The only exception is RT-H3 and effect size, for which the orange dot is just within the top of the box. 75% of the RTs therefore underestimate the NSEs for 11 out of the 12 cases.

3.3 Alternative explanations

After having presented all our results, it is useful to discuss alternative explanations. Might the sizeable 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 research teams and peer evaluators who 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 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 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.2 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 the 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.6 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 explanation for sizeable 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 are sizeable for relatively precise hypotheses as well. 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 informing them about the following before signing 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;[28] and, only those who completed all stages would 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 did not 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.

[28]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).

4 Conclusion

Testing an hypothesis on a data sample involves many decisions on the side of the researcher. He needs to find an appropriate econometric model, clean the data, choose suitable software for estimation, et cetera. These choices inherently generate some dispersion in results across researchers. There is no right or wrong here, there simply are many roads that lead to Rome. We propose the across-researcher standard deviation in results as a measure of such dispersion and refer to it as the non-standard error.

Studying NSEs involves letting a representative group of researchers *independently* test the same set of hypotheses on the same sample. #fincap did exactly this. 164 research teams tested relatively standard types of hypotheses on novel and proprietary trade data.

We compute NSEs and show that they are sizeable, similar in magnitude to SEs. They, for the most part, do not co-vary with various quality variables: research-team quality, work-flow quality, and paper quality. The last one is assessed by 34 external peer evaluators. PEs also provided written feedback, so that each research team received anonymized feedback from two PEs. After receiving this feedback, teams were allowed to update their analysis and post new results. NSEs declined weakly signficantly in this stage. They declined significantly further in two subsequent stages, in which teams received new information. The total decline across all stages is substantial, up to 50%. Finally, we find that teams grossly underestimate NSEs, and more so for more abstract hypotheses.

In sum, we believe that our study shows that NSEs add non-negligible uncertainty to the outcome of hypotheses tests in finance. We document that NSE uncertainty is similar in magnitude as SE uncertainty. This is particularly worrisome as it creates substantial space for *p*-hacking (which had been scoped out here). An encouraging result, however, is that peer feedback reduces NSEs. We therefore believe that the profession should further encourage researchers to make data available, to post a pre-registration plan, to consider a multi-analyst approach, and to stimulate interaction among researchers. This, hopefully, does not leave us lost in the larger metropolitan area, but leads us right to the Forum Romanum.

Appendices

A RT-sample, RT-hypotheses, and results

This appendix presents the RT-hypotheses in detail and the test results of #fincap RTs as a group. We start by providing a motivating context.

A.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 the dealers' quoted ask prices when buying, and accepting their bid prices when selling. The wedge between dealer bid and ask prices, the bid-ask spread, was a useful measure of trading cost, and often still is.

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

A.2 RT-hypotheses and test results

The RT-hypotheses and results are discussed based on RT-results in the final stage of the project (Table OA.3 in the Online Appendix). We therefore base our discussion on the results that RTs settled on after receiving feedback. What do RTs find after having shown some convergence across the stages? And, consistent with the main text, the presence of extreme values makes us prefer to analyze the 2.5%-97.5% winsorized sample.

(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. RTs predominantly reject the hypothesis in favor of a decline of market efficiency. In the final stage, 61.6% of RTs report an implied t-value that is larger than 1.96 in absolute value, and therefore significant at a conventional 5% significance level. Of these, 51.8/61.6 = 84.1% report a significant decline. The decline

seems modest as the across-RT mean[29] is -1.7% per year. The small changes add up, though, to a total decline in the 2002-2018 sample of $(0.983^{17} - 1) = 25.3\%$. 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. The regulation constrained the supply of liquidity by sell-side banks (e.g. Bao, O'Hara, and Zhou, 2018; Jovanovic 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 majority of RTs (53.7%) find statistical significance, mostly (90.9%) in the direction of a decline in the realized spread. The decline is 2.1% per year, which implies a 30.3% 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. Almost all RTs (86.6%) find statistical significance, almost exclusively (97.9%) pointing towards a decline in the share of client volume. The average decline is 2.7% per year, which implies a total decline of 37.2% 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.[30]

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

[30] 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, their volume changes status from agency to principal.

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. 32.3% of RTs would reject the null hypothesis, with the majority (71.8%) in favor of a decline in realized spread. The average decline is 0.7% per year and 11.3% for the full sample. The decline in client realized spread is therefore only about a third of the total realized spread decline, which suggests that market orders of intermediaries benefited most from a 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. 31.1% of RTs would reject the null, but this time, interestingly, about half find a significant decline, whereas the other half finds a significant increase. The average per-year change is -0.2% which adds up to -3.3% for the full sample. The results 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 limit orders[31] as opposed to market orders. The benefit of execution via limit order is that one earns half the bid-ask spread rather than pays for 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. Only 12.8% of RTs find significance with about half in favor of GTR decline and the other half in favor of an increase. The average change is a 1.5% increase, which for the full sample implies a total change of 28.8%. Since the

[31] By executing via a limit order we mean, submitting a limit order that cannot be executed immediately and, therefore, enters the book and eventually gets matched with an incoming market order.

average client GTR is most likely negative, this result implies that it became more negative thus suggesting worse overall execution quality. We caution that this result is very weak in terms of significance and in terms of agreement across RTs on the sign. If, however, overall execution quality did indeed decline, then we would come full circle with the RT-H1 finding of a market efficiency decline, potentially due to a decline in market depth.

B Explanatory variables for error variance

B.1 Team quality

The quality measures for research teams are based on the survey that participants filled out upon registration (see Figure OA.2 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.

B.2 Workflow quality

We proxy for workflow quality with an objectively obtained score of code quality provided by Cascad (see footnote 19). The scale ranges from 0 (serious discrepancies) to 100 (perfect reproducibility).

B.3 Paper quality

Papers are rated by an external group of peer evaluators (PEs). They rate the analyses associated with each RT-hypothesis individually, but also the paper in its entirety (see Figure OA.11 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).

C 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:

$$AdjBeliefOnStDev_{ij} = BeliefOnStDev_{ij} - \left[\left(\frac{1}{n}\sum_{i} BeliefOnStDev_{ij}\right) - RealizationOfStDev_{j}\right] \quad (8)$$

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 (7). 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:[32]

$$2\min(EmpiricalQuantileFincapStatistic, 1 - EmpiricalQuantileFincapStatistic). \quad (9)$$

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,

[32]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.

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 (forthcoming)* .

Adams, John, Darren Hayunga, Sattar Mansi, David Reeb, and Vincenzo Verardi. 2019. "Identifying and Treating Outliers in Finance." *Financial Management* 48:345–384.

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

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.

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, Sebastian 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.

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.

Chang, Andrew C. and Phillip Li. 2017. "A Preanalysis Plan to Replicate Sixty Economics Research Papers That Worked Half of the Time." *American Economic Review: Papers & Proceedings* 107:60–64.

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

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *American Economic Review* 76:587–603.

Garlappi, Lorenzo, Raman Uppal, and Tan Wang. 2006. "Portfolio Selection with Parameter and Model Uncertainty: A Multi-Prior Approach." *Review of Financial Studies* 20:41–81.

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

Gertler, Paul, Sebastian Galiani, and Mauricio Romero. 2018. "How to Make Replication the Norm." *Nature* 554:417–419.

Glandon, Philip J. 2011. "Appendix to the Report of the Editor: Report on the American Economic Review Data Availability Compliance Project." *American Economic Review: Papers & Proceedings* 101:695–699.

Greene, W.H. 2007. *Econometric Analysis*. London: Prentice Hall.

Hansen, Lars Peter and Thomas J. Sargent. 2001. "Robust Control and Model Uncertainty." *AEA Papers and Proceedings* 91:60–66.

Harvey, Andrew C. 1976. "Estimating Regression Models with Multiplicative Heteroscedasticity." *Econometrica* 44:461–465.

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

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.

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)* .

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

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.

Lo, Andrew W. and A. Craig MacKinlay. 1990. "Data-Snooping Biases in Tests of Financial Asset Pricing Models." *Review of Financial Studies* 3:431–467.

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

McCullough, Bruce D. and Hrishikesh D. Vinod. 2003. "Verifying the Solution from a Nonlinear Solver: A Case Study." *American Economic Review* 93:873–892.

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.

Rampini, Adriano A., S. Viswanathan, and Guillaume Vuillemey. 2021. "Retracted: Risk Management in Financial Institutions." *Journal of Finance* 76:2709–2709.

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, Alexandra 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.

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. Nijstad, 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.

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

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 econ/finance publications (see footnote 4)	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 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 effect size						
Mean	446.3	-1,093.4	-3.5	-38,276.1	-3.5	-87.1
Mean (wins.)[a]	-7.4	9.0	-2.6	-2.1	-0.3	-27.6
Mean (trim)[b]	-6.2	5.4	-2.7	0.6	-0.2	-22.1
SD	5,817.5	14,537.2	9.4	490,024.2	37.6	728.5
SD (wins.)[a]	20.6	49.5	1.8	46.8	1.1	187.8
SD (trim)[b]	16.6	29.0	1.5	24.9	0.9	128.7
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
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
Mean (wins.)[a]	13.2	23.5	1.4	26.9	1.7	104.8
Mean (trim)[b]	10.8	16.4	1.3	18.8	1.4	80.7
SD	5,810.6	14,711.9	29.5	489,929.5	40.1	526.0
SD (wins.)[a]	27.0	60.8	1.6	75.3	2.6	237.5
SD (trim)[b]	21.2	38.8	1.3	53.1	1.7	174.1
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
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
Mean (wins.)[a]	-1.4	-1.3	-13.3	-0.4	0.2	-0.4
Mean (trim)[b]	-1.1	-1.0	-11.6	-0.3	0.0	-0.2
SD	28.4	541.2	269.9	406.0	60.1	21.2
SD (wins.)[a]	5.2	4.1	25.4	3.0	4.2	2.0
SD (trim)[b]	3.4	3.3	22.0	2.4	2.6	1.4
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
Max	51.6	6,880.5	29.5	5,119.5	89.6	100.6
More t-value statistics						
t <-1.96	23.8%	22.6%	45.7%	17.7%	11.0%	9.8%
t >1.96	8.5%	6.1%	3.7%	9.1%	9.8%	3.0%
$\|t\|$ >1.96	32.3%	28.7%	49.4%	26.8%	20.7%	12.8%
Relative size NSE (wins.)[c]						
NSE/SE ratio	1.6	2.1	1.3	1.7	0.6	1.8

[a]: Winsorized at 2.5%-97.5%. [b]: Trimmed at 2.5%-97.5%. [c]: The non-standard error of effect size is compared to the mean standard error of effect size for the winsorized sample.

Table 2: Principal component analysis team quality

This table presents the results of a principal component analysis of the standardized team quality variables.

Panel (a): Correlation team quality measures

	Publications	Experience	Big Data	Position	#Members
Publications		0.34	0.10	0.54	0.30
Experience			-0.18	0.25	0.12
Big Data				0.14	0.14
Position					0.16

Panel (b): Fraction of variance explained

	PC1	PC2	PC3	PC4	PC5
Variance explained	38.3%	23.6%	17.1%	12.4%	8.6%

Panel (c): Loading of principal components on variables

	Publications	Experience	Big Data	Position	#Members
PC1	0.61	0.40	0.13	0.55	0.37
PC2	-0.01	-0.55	0.79	0.05	0.26
PC3	-0.10	0.06	-0.21	-0.46	0.86
PC4	-0.20	0.71	0.56	-0.35	-0.12
PC5	-0.76	0.14	-0.02	0.60	0.22

Table 3: Stage-1 error-variance regressions

This table presents the results of stage-1 error-variance regressions. The standard errors in parentheses are based on clustering on research teams. The incremental R^2 measure how much more is explained beyond a model with simple dummies for RT-hypotheses. $^{*}/^{**}$ correspond to significance at the 5/0.5% level, respectively.

Panel (a): Estimates

	Raw	Winsorized		Trimmed	
		1% to 99%	2.5% to 97.5%	1% to 99%	2.5% to 97.5%
Team quality (standardized)	0.01	-0.13	-0.16*	-0.12	-0.12
	(0.06)	(0.07)	(0.08)	(0.06)	(0.08)
Reproducibility score (standardized)	-0.13	-0.15	-0.24**	-0.13	-0.08
	(0.07)	(0.09)	(0.08)	(0.09)	(0.08)
Average rating (standardized)	-0.12	-0.14	-0.18	-0.17	-0.17
	(0.07)	(0.10)	(0.09)	(0.11)	(0.09)
RT-hypotheses dummies	Yes	Yes	Yes	Yes	Yes
R^2	0.95	0.73	0.65	0.72	0.63
#Observations	984	984	984	972	936

Panel (b): *t*-values

	Raw	Winsorized		Trimmed	
		1% to 99%	2.5% to 97.5%	1% to 99%	2.5% to 97.5%
Team quality (standardized)	-0.00	-0.01	0.00	-0.05	0.05
	(0.08)	(0.09)	(0.11)	(0.09)	(0.10)
Reproducibility score (standardized)	-0.12	-0.18	-0.23*	-0.21*	-0.09
	(0.09)	(0.09)	(0.11)	(0.08)	(0.10)
Average rating (standardized)	-0.15	-0.21	-0.09	-0.01	0.02
	(0.12)	(0.13)	(0.12)	(0.10)	(0.11)
RT-hypotheses dummies	Yes	Yes	Yes	Yes	Yes
R^2	0.71	0.49	0.37	0.49	0.32
#Observations	984	984	984	972	936

Table 4: All-stages error-variance regressions

This table presents the results of error-variance regressions including all stages. The standard errors in parentheses are based on clustering on research teams. $^{*}/^{**}$ correspond to significance at the 5/0.5% level, respectively.

Panel (a): Estimates

	Raw	Winsorized 1% to 99%	Winsorized 2.5% to 97.5%	Trimmed 1% to 99%	Trimmed 2.5% to 97.5%
Dummy Stage 2 - Dummy Stage 1	-0.11*	-0.24*	-0.29*	-0.30**	-0.21*
	(0.05)	(0.09)	(0.11)	(0.10)	(0.11)
Dummy Stage 3 - Dummy Stage 2	0.04	-0.26**	-0.40**	-0.21*	-0.40**
	(0.03)	(0.07)	(0.09)	(0.09)	(0.09)
Dummy Stage 4 - Dummy Stage 3	-0.10**	-0.26**	-0.38**	-0.25**	-0.39**
	(0.03)	(0.05)	(0.06)	(0.05)	(0.06)
Dummy Stage 4 - Dummy Stage 1	-0.17*	-0.75**	-1.07**	-0.76**	-1.00**
	(0.06)	(0.10)	(0.12)	(0.10)	(0.12)
RT-hypotheses dummies	Yes	Yes	Yes	Yes	Yes
R^2	0.96	0.58	0.51	0.54	0.50
#Observations	3,936	3,936	3,936	3,888	3,744

Panel (b): *t*-values

	Raw	Winsorized 1% to 99%	Winsorized 2.5% to 97.5%	Trimmed 1% to 99%	Trimmed 2.5% to 97.5%
Dummy Stage 2 - Dummy Stage 1	-0.01	0.04	0.28**	0.10	0.24*
	(0.07)	(0.09)	(0.09)	(0.08)	(0.08)
Dummy Stage 3 - Dummy Stage 2	-0.20**	-0.23*	-0.08	-0.14	0.14
	(0.06)	(0.09)	(0.08)	(0.09)	(0.07)
Dummy Stage 4 - Dummy Stage 3	-0.06	-0.32**	-0.61**	-0.43**	-0.73**
	(0.03)	(0.05)	(0.08)	(0.07)	(0.09)
Dummy Stage 4 - Dummy Stage 1	-0.26**	-0.51**	-0.41**	-0.46**	-0.36**
	(0.09)	(0.11)	(0.12)	(0.11)	(0.12)
RT-hypotheses dummies	Yes	Yes	Yes	Yes	Yes
R^2	0.76	0.57	0.42	0.47	0.39
#Observations	3,936	3,936	3,936	3,888	3,744

Table 5: Dispersion on research team beliefs

This presents test statistics on whether the beliefs of research teams about dispersion across research teams matches realized dispersion. More precisely, the test statistic is defined as the difference between the average belief on the standard deviation across teams minus the realized standard deviation, divided by the latter. The p-values in parentheses are obtained through bootstrapping. This analysis is the only one that uses the unwinsorized sample as beliefs were solicited for the raw data. $^{*}/^{**}$ correspond to significance at the 5/0.5% level, respectively.

	RT-H1 Efficiency	RT-H2 RSpread	RT-H3 Client Volume	RT-H4 Client RSpread	RT-H5 Client MOrders	RT-H6 Client GTR	All
Estimate	-99.5%** (0.00)	-95.4%** (0.00)	-9.0% (0.64)	-97.5%** (0.00)	-45.3% (0.50)	-83.3%** (0.00)	-71.7%** (0.00)
t-value	15.9% (0.19)	-96.0%** (0.00)	-92.4%** (0.00)	-97.1%** (0.00)	-86.0%** (0.00)	-68.2%** (0.00)	-70.6%** (0.00)

Figure 1: Countries of #fincap community

This plot illustrates how the #fincap community is dispersed around the globe. The top plot depicts how the members of the research teams are dispersed across countries. The bottom plot does the same for the peer evaluators.

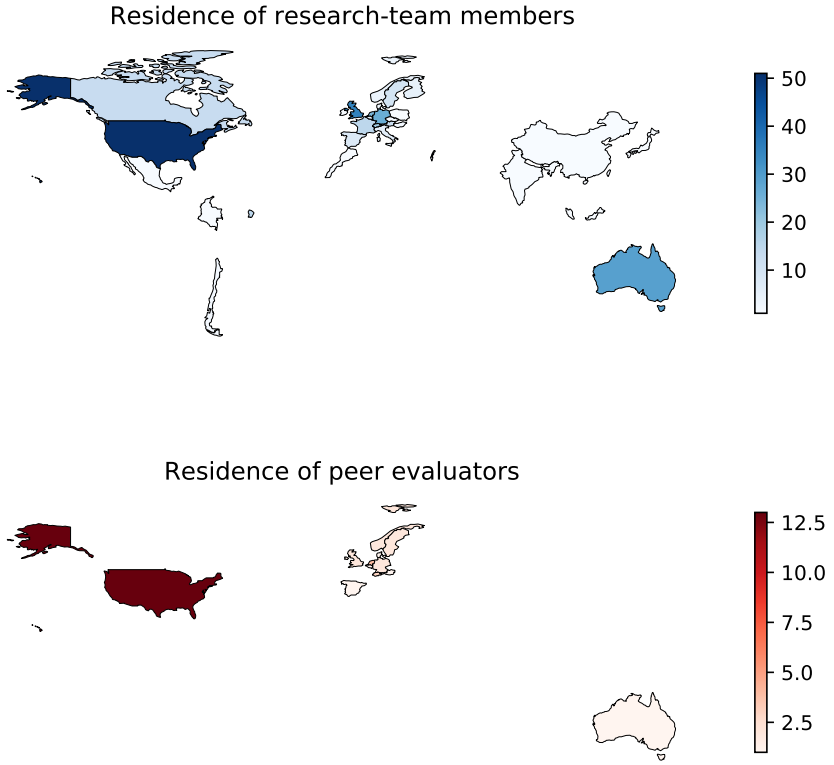


Figure 2: Dispersion in research-team estimates

This plot illustrates the dispersion in the stage-1 estimates reported by the research teams for all six hypotheses. It is based on the raw sample. The boxes span the the first to the third quartile with the median as the interior horizontal line. The whiskers span 95% of the the observations, starting from the 2.5% quantile to the 97.5% quantile.

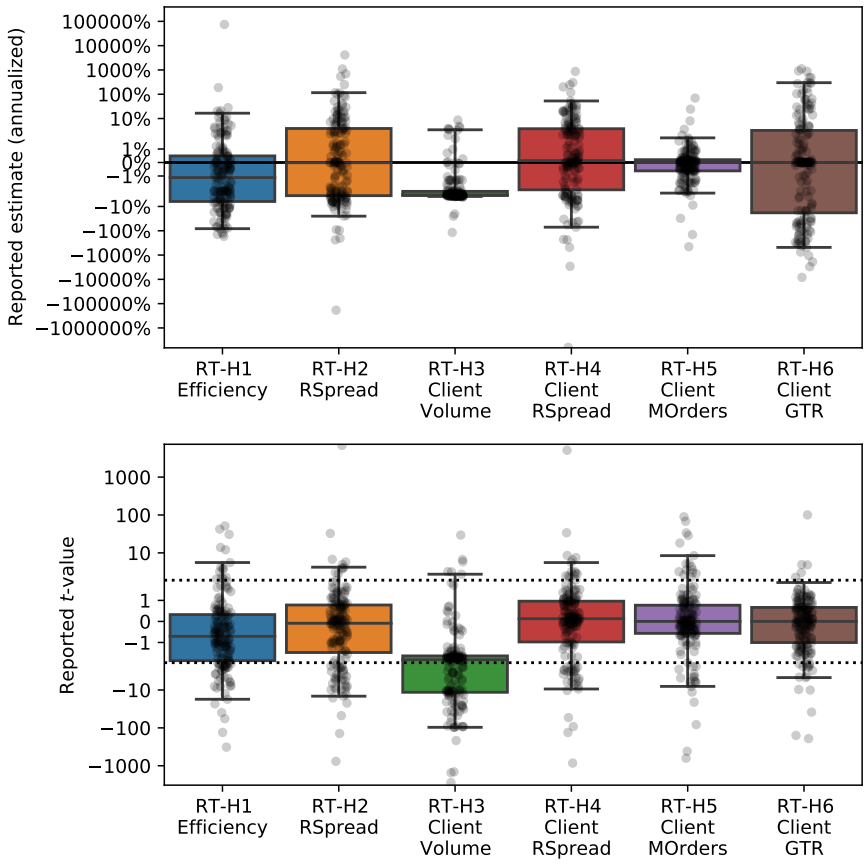


Figure 3: Dispersion research-team estimates: A high-quality subset

This plot mirrors Figure 2 but reports dispersion only for the subset of the highest-quality research teams with highest-quality results (N=9). The teams score one on *all* of the binary team-quality measures used in Table 3 and at least 7.5 in the experience field (on a 1-10 scale). The reproducibility score of their analysis is at least 75 (out of 100) and their average peer evaluator rating is at least 7.5 (out of 10). The boxes span the the first to the third quartile with the median as the interior horizontal line. The whiskers span 95% of the the observations, starting from the 2.5% quantile to the 97.5% quantile.

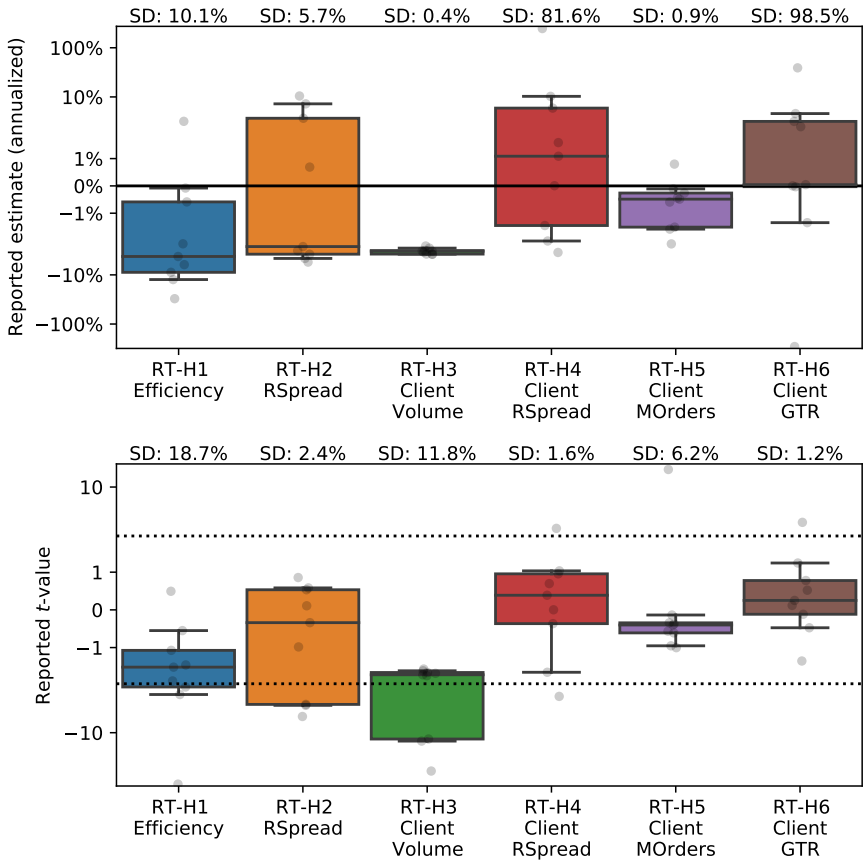


Figure 4: Dispersion research team estimates across all stages

This plot illustrates the dispersion in the estimates reported by the research teams for all six hypotheses. The boxes span the the first to the third quartile with the median as the interior horizontal line. The whiskers span 95% of the the observations, starting from the 2.5% quantile to the 97.5% quantile.

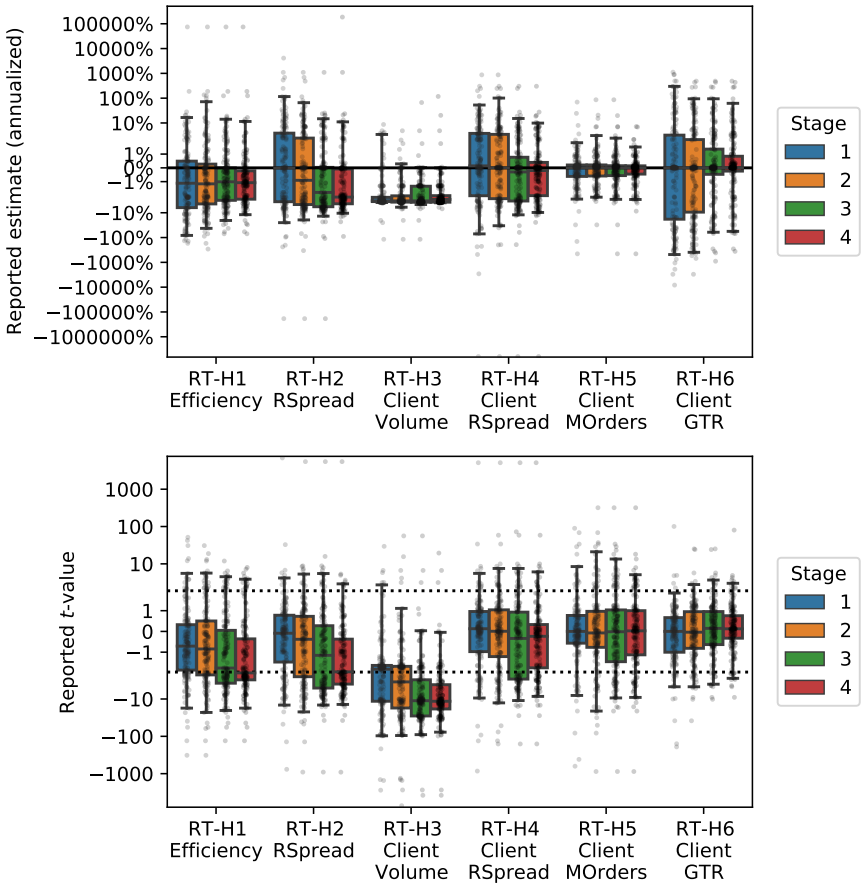


Figure 5: Dispersion research team beliefs

This plot illustrates the dispersion in beliefs across research teams for all six hypotheses. The teams were asked to report their belief on how large the standard deviation would be across all teams of the reported estimates and *t*-values. The whiskers span 95% of the reported beliefs, starting from the 2.5% quantile to the 97.5% quantile. The red dots represent the "truth," the observed standard deviation across research teams. This figure is the only one that uses the unwinsorized sample as beliefs were solicited for the raw data.