Pre-Analysis Plan: Same Data, Same Hypotheses, Different Teams: Measuring and Understanding Dispersion in Results

Albert J. Menkveld Anna Dreber Felix Holzmeister
Jürgen Huber Magnus Johannesson Michael Kirchler
Sebastian Neusüss Michael Razen Utz Weitzel

January 11, 2021

Abstract

This document presents a strategy to analyze the dispersion of empirical results across research teams who all test the same set of hypotheses on the same data. We rely on a standard variance analysis approach, adapted to a setting where initial dispersion might change as teams learn from peer feedback. More specifically: How large is the initial dispersion? What explains it? And, does it decline during various stages of peer feedback? Finally, do researchers themselves accurately predict the size of such dispersion?

1 Objective

Academic research recognizes randomness in data samples by reporting confidence intervals around parameter estimates. It often does *not* recognize the "randomness" that is in the research process itself. It is the latter that is the focus of this study.

If multiple researchers analyze the same type of questions empirically, they might report different results. This dispersion can have a range of causes. It might be the result of straightforward implementation errors such as a glitch in the computer code. It might also be due to different routes traveled through the "garden of forking paths." Gelman and Loken (2014) use the metaphor to describe the choices a researcher has to make when analyzing data (e.g., which dataset to use, what hypotheses to test, how to process outliers, which statistical model to use, how to design statistical tests, what software to pick, etc.).

The presence of forks potentially corrupts the scientific 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 of there being no effect). If there are multiple paths, then researchers might be tempted to pick the path that produces the lowest *p*-value and write

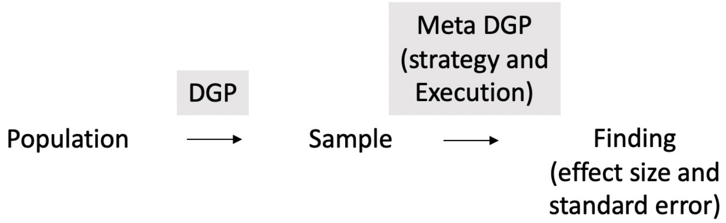
Meta DGP (strategy and Execution)
DGP
Population
Sample
Finding (effect size and standard error)

Figure 1: Two data-generating processes (DGPs).

the paper accordingly (i.e., they are hacking the p-value). Such worry is not only "academic," it seems to have corrupted published academic work. Studies that tried to replicate high-profile published papers typically find lower effect sizes and less statistical strength (Open Science Collaboration, 2015; Camerer et al., 2016, 2018). Munafò et al. (2017) survey the various threats to credible empirical science and propose several fixes.

Our objective is to measure and analyze the dispersion in results when multiple researchers *independently* test the same set of hypotheses on the same data. We are therefore interested in the dispersion of results inherent in the empirical strategy and execution.[1] We deliberately stop short of studying p-hacking as we are interested in genuine dispersion (similar to Botvinik-Nezer et al. (2020)).

Figure 1 illustrates what we have in mind. One could say that there are two data-generating processes (DGPs) underlying empirical science. There is one DGP that generated the realization (i.e., the data sample) that a research team uses to conduct their analysis. The team is mindful of the randomness in their realization. This is why they report point estimates of effect sizes along with confidence interval (based on standard errors).

There is another DGP that captures the empirical strategy picked by the research team and the execution. This adds randomness in the outcome that is not explicitly accounted for. We will refer to this as "meta randomness" due to the "meta DGP" to distinguish it from sample randomness due to the first (conventional) DGP. The presence of this additional meta DGP implies that *true* confidence intervals are wider than reported one.

We study the dispersion generated by the meta DGP for a given realization of the first DGP. More specifically, we will study the following questions:

1. How large is the dispersion inherent in the research process?
2. Can the level of dispersion be explained? For example, is it larger for less experienced researchers?
3. Part of the research process is getting evaluated by peers. Does such feedback from peer evaluators (PEs) remove some of this dispersion? In

[1] This is of interest given that modern technology has made sharing data relatively straightforward (e.g., by posting a sample on the internet). Concurrently, the academic journals and funding agencies increasingly incentivize authors to do so.

other words, to what extent can the scientific community reach consensus through peer feedback?

4. Researchers themselves are undoubtedly aware of the dispersion inherent in the research process, but are their beliefs about the size of it accurate?

Addressing all questions requires a setting where many teams test the same hypotheses on the same data. A dispersion estimate can then be based on many sets of results, and the size of dispersion be related to, for example, observable research-team or work-flow characteristics. The first two questions can be addressed with such results (for our statistical approach see Section 3).

Addressing the third question requires that teams receive feedback from PEs and are allowed to update their results based on it. This stage mimics 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). The dynamics at play in a refereeing process at a scientific journal are out of scope.[2]

Finally, the fourth question can be addressed by asking all teams about their beliefs on the within-community dispersion. All of the above motivated the #fincap project and the way it is structured. Collectively, these four questions serve as the project's (meta) objectives.

2 Project design and hypotheses

This section describes the project design in detail and, based on it, translates our objectives into testable hypotheses.

2.1 Project design

The #fincap project is about multiple research teams (RTs) testing the same set of six hypotheses on the same data. We henceforth will refer to these hypotheses as RT-hypotheses to avoid any confusion with our (meta) hypotheses presented in Section 2.2. For each RT-hypothesis, RTs report an estimate for the effect size along with its standard error.[3] They further report what they believe the

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

[3] Of course, both reported effect sizes and reported standard errors are estimates, as the true values are unknown. For ease of distinction, we will use the term "(point) estimate" when referring to reported effect size, "standard error" when referring to reported standard error, and "result" as hypernym for reported effect sizes, reported standard errors, and resulting t-statistics.

dispersion (in terms of standard deviation) will be across RTs in estimates and their associated t-statistics.[4]

The RTs are asked to write a short academic paper in which they present and discuss their findings. These papers get evaluated by PEs who were recruited outside the set of researchers who registered as research teams. The PEs will score the submissions of the research teams, both at the level of the RT-hypothesis and at the paper level. They will motivate their scores in the paper. The RTs will receive this feedback unabridged and are allowed to update their results based on the feedback. In all analysis in the remainder of the project, we remove a PE fixed effect by demeaning the PE scores. The project consists of four stages:

Stage 1 RTs report the results after doing their own analysis.

Stage 2 RTs report the results after re-considering their own analysis after receiving feedback from PEs.

Stage 3 RTs report the results after observing what PEs consider the best five papers. These papers are the ones that rank highest based on the average overall[5] PE score (at Stage 1). If there are ties, then the overall Cascad[6] reproducibility score will serve as tie-breaker. If the reproducibility score cannot fully break the ties, then we will randomly pick among the tied candidates to arrive at five papers in total.

Stage 4 RTs report their final results (i.e., their estimates of the effect sizes as well as the corresponding standard errors with all evidence available to them. Final results are thus not necessarily equal to the output of their code).

We will use variance as our dispersion measure because it is additive and the *de facto* standard dispersion measure in econometrics/statistics. Let us therefore define dispersion at stage t for RT-hypothesis j as:

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

where

- $i \in \{1, \ldots, n\}$ indexes RTs,

[4] The t-statistic is defined as the estimate divided by its standard error.

[5] With "overall" we mean the score on the PE score sheet that pertains to the overall paper.

[6] The Certification Agency for Scientific Code and Data, or Cascad for short, 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*). Cascad rates reproducibility on a five-category scale: RRR. Perfectly reproducible. RR. Practically perfect. R. Minor discrepancies. D. Potentially serious discrepancies. DD. Serious discrepancies. Cascad considered converting this categorical rating to an equal-distance numeric one reasonable: RRR, RR, R, D, and DD become 100, 75, 50, 25, 0, respectively.

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

To explain dispersion requires allowing for heterogeneity in the variance of idiosyncratic "errors" across teams, where errors are defined as:[7]

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

2.2 Hypotheses

With all the groundwork done, we can now translate our overall objectives into three sets of hypotheses. These hypotheses are tested for two types of results reported by the teams:

- The effect size (i.e., the point estimate).
- The t-statistic (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 outcomes on the variable of interest. The t-statistic captures the statistical strength of the reported effect in the sense of how likely it is to be different from no-effect (i.e., a zero effect). Both these statistical objects, the point estimate and its t-statistic, are key objects in empirical projects and both are therefore in focus here.[8] All hypotheses are stated as null hypotheses. The hypotheses tests will be two-sided tests.

The first set of three hypotheses focuses on whether error variance can be explained:

H1 Team quality, proxied by the first principal component by the following pre-defined variables, does not explain the size of errors in the first submission (more specifically, it does not explain error variance in Stage 1). These variables, defined at the level of RTs, are:[9]

(a) Prior top publication: Top-5 publication in economics or Top-3 publication in finance (0/1).[10]

[7] The term "error" is used in a statistical sense, where it defines a measure of how observed data differ from a population average. It should *not* be confused with errors necessarily being mistakes. Strictly speaking, $\hat{u}_{ijt}$ in (2) denotes the OLS residual, but we use the term "error" as a more tangible expression. Further, note that (2) results from the application of OLS on Equation (2) in Harvey (1976), with our model being a special case with only intercepts pertaining to the six dummies. Moreover, throughout our paper, we use the same notation as Harvey (1976) to enhance readability.

[8] We will also test the first seven hypotheses (laid out below) on the standard errors that RTs report. The only exception is the eighth hypothesis since we do not measure beliefs about the dispersion in standard errors. These test results will be made available in an online appendix.

[9] In further exploratory analysis we will, instead of the first principal component, add all these variables as explanatory variables in a multivariate regression.

[10] 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*.

(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: Having worked with datasets of similar size or larger than the Deutsche Börse dataset analyzed in #fincap (0/1).

(d) Academic seniority: One for holding an associate or a full professorship, zero otherwise (0/1).

(e) Team composition: Team comprising of two researchers (0/1).

H2 Work flow quality as proxied by the Cascad reproducibility score (hypothesis level) does not explain the size of errors in Stage 1 (scale from 0 to 100) which corresponds to the certification rating of Cascad (used by, for example, the *American Economic Review*).

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 (as stated earlier in the document).

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.[11]

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

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

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

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

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

H8 The belief of RTs about dispersion (as measured by the standard deviation) in results across RTs is correct.

3 Statistical approach

Heterogeneity in error variance is known as heteroskedasticity in econometrics. Greene (2007, Ch. 9.7), a standard econometrics textbook, 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:

$$\begin{aligned}\sigma^2_{ijt} &= \text{var}\left(u_{ijt}\right) \\ &= \exp\left(z'_{ijt}\alpha\right),\end{aligned} \tag{3}$$

[11] The decline across consecutive stages might be small but their sum across all might be sizeable. It is in this sense that the first-to-final stage might be statistically most powerful.

where u_{ijt} is the unobserved disturbance term, and z_{ijt} contains the set of multiplicative explanatory variables which includes six dummies corresponding to the six RT-hypotheses to account for possible heteroskedasticity across hypotheses. The parameter α measures the marginal *relative* change in variance of a unit change in the associated explanatory variable. For example, a coefficient of 0.1 for variable X means that the team's error variance increases by ten percent when variable X increases by one unit. This corresponds to a five percent increase in its standard deviation.[12]

Estimation. The parameter α is estimated as follows. First, the following model is estimated with ordinary least squares (OLS):[13]

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

Let $\tilde{\alpha}$ denote the resulting parameter estimate. Note that $\tilde{\alpha}$ estimates α consistently, except for the parameters associated with the six intercepts, which need to be adjusted. A consistent estimator for the α vector therefore is (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 Equation (6) in Harvey (1976). Then the predicted variance for team i's error for RT-hypothesis j in stage t is:

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

Statistical inference. The relevant hypotheses can then be tested by using the Wald statistic where the residuals w_{ijt} are clustered on RTs.[14] The R-squared of (4) in this case is a useful measure of how much of the dispersion can

[12] 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 $f(x) \approx \sqrt{\mu} + \frac{1}{2\sqrt{\mu}} 0.10\mu = 1.05\mu$.

[13] It is tempting to drop the natural logarithm on the dependent variable and simply explain variance OLS. Harvey (1976, p.6) states the following about his multiplicative model compared to such simple additive model:

> From the point of view of estimation, the multiplicative heteroscedasticity model considered here appears to be rather more attractive than the "additive" model in which either the variance or standard deviation of the ith disturbance term is assumed to be related to a linear combination of known variables [...]. There are three reasons for this. 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.

[14] The clustering is needed to account for possible non-zero correlation in residuals w_{ijt}. RT clustering is needed to account for the fact that if an RT shows a large error in testing

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.

4 Implementation

In this section we discuss in detail how we implement the proposed methodology to test our hypotheses.

4.1 Preliminary remarks

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 comparing the magnitude of errors across the stages as pure as possible as we follow the same set of teams through time.

Significance threshold. All hypotheses tests will be two-sided and we will refer to results with a p-value smaller than 0.005 as "statistically significant evidence" and results with a p-value smaller than 0.05 as "suggestive evidence" following the recommendation of Benjamin et al. (2018).

4.2 Summary statistics

In addition to tests on the various hypotheses, we will present descriptive results about the dispersion in results across RTs. For ease of interpretation we will report the standard deviation in results across RTs and its 95% confidence interval (as a measure of the precision in the estimated dispersion in results). In addition we will report the mean and median result across RTs and the minimum and maximum result. The above descriptive results will be reported for each hypothesis in each of the four stages, and separately for effect sizes and t-statistics.

4.3 Hypotheses tests

The tests of the various hypotheses are all implemented in the balanced sample (i.e., include only RTs that completed all stages). They all use the methodology presented in Section 3 except for the final hypothesis.

- The first three hypotheses (H1 through H3) are based on the sub-sample of initial submissions. Each hypothesis is tested by regressing log squared errors on the set of explanatory variables in focus. The RTs report results for the six RT-hypotheses they study. We allow for heterogeneity

one RT-hypothesis, they likely exhibit large errors for all RT-hypotheses. We decide not to cluster on RT-hypotheses as the model in (4) includes fixed effects for RT-hypotheses. We do not cluster on PEs as a possible PE fixed effect was removed by demeaning PE scores in all of our analysis.

across RT-hypotheses by including dummies for the second through sixth hypothesis.

- The hypotheses on convergence (H4 through H7) are based on the full sample (which is a balanced sample because throughout the manuscript we only include RTs that submitted in all four stages). The hypotheses can all be tested by regressing log squared errors on an intercept and dummies for the second, third, and fourth stage. The results can be presented graphically with confidence bounds by translating confidence intervals for (the sum of) parameter estimates to the variance of reported effects using (3) (or its standard deviation). Similar to the tests of H1 through H3, we include dummies to account for heterogeneity across RT-hypotheses.

- The eighth hypothesis requires a test on the equality of means: the mean belief about standard deviation in results across teams and the standard deviation of these results in the population. The results referred to here include

  - the reported effect size and
  - the t-statistic implied by the reported effect size and the reported standard error (i.e., the ratio of these two).

  We define the following test statistic which is based on the relative distance between the belief and the realization (only for Stage 1 as beliefs only pertain to that stage):

$$D = \frac{1}{6n} \sum_{i,j} \left(\frac{\mathit{BeliefOnStDev}_{ij} - \mathit{RealizationOfStDev}_j}{\mathit{RealizationOfStDev}_j} \right), \quad (7)$$

  where $\mathit{StDevBelief}_{ij}$ is the belief of team i on the standard deviation across teams for hypothesis j and $\mathit{StDevRealization}_j$ is the realized standard deviation for hypothesis j. The benefit of the relative measure as opposed to an absolute measure is that (i) it is easy to interpret as it allows for statements as teams over-estimate dispersion by 10% and (ii) it accounts for level differences across hypotheses (e.g., under the null of equal means of beliefs and realized dispersion, a uniform distribution of beliefs on the support 0.09 to 0.11 will exhibit the same dispersion as a uniform distribution of beliefs on 900 to 1100).
  The 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 stan-

dard deviation, from the beliefs:

$$\begin{aligned} \mathit{AdjBeliefOnStDev}_{ij} = \mathit{BeliefOnStDev}_{ij} - \Bigg[& \left(\frac{1}{n} \sum_i \mathit{BeliefOnStDev}_{ij} \right) \\ & - \mathit{RealizationOfStDev}_j \Bigg] \end{aligned} \tag{8}$$

In this new sample with adjusted beliefs from which we will draw in the bootstrap procedure, 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 of (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:[15]

$$\begin{aligned} & 2 \min \left(\mathit{EmpiricalQuantileFincapStatistic}, \right. \\ & \quad \left. 1 - \mathit{EmpiricalQuantileFincapStatistic} \right), \end{aligned} \tag{9}$$

where, e.g., the empirical quantile of the observed #fincap statistic is 0.25 if the observed #fincap statistic is closest in value to the value of the 2500th ordered simulated statistic.

4.4 Supplementary analyses

We will conduct the following robustness tests:

[15] Note that the procedure accounts for within-RT correlations (i.e., including non-zero correlations in the results an RT reports across RT-hypotheses and the beliefs it reports across hypotheses, and non-zero correlations across an RT's contribution to results and its beliefs). 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 experiment design.

1. We will test H1 with, instead of the first principal component, all RT characteristics added in a multivariate model, and each characteristic individually in univariate models.
2. We will redo all analysis (where we can) with the full dataset (i.e., the unbalanced panel).

We might do further exploratory analysis. If any such analyses enters our paper we will clarify that it was not part of this pre-analysis plan.

Appendices

A Explanatory variables for error variance

A.1 Team quality

Team quality measures are derived from the survey that participants filled out upon registration. The survey comprises additional items which we included to provide a better description of the sample but will not enter the regression analysis outlined in Section 2.2, either because (i) we do not expect them to affect error variance (country of residence/workplace, gender) or (ii) the corresponding quality is already captured by another variable (Google Scholar citations [captured by prior top publication], years since PhD [captured by academic seniority]), or (iii) the corresponding variable was used to assess eligibility as a research team (field of study, highest degree).[16]

To keep our 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 (i.e., academic seniority [holds associate/full professorship] and experience with big data, respectively). As the #fincap dataset comprises of 652 million observations, we set the latter variable to 1 if a researcher indicates that the largest dataset they have worked with contained at least between 100 million and 1 billion observations.

Using binary variables instead has the benefit that it forces a focus on what we deem to be a critical threshold value (i.e., holding a professorship for substantial academic achievement or having worked with an equally large dataset for prior dataset experience). Consequentially, we also treat these variables as binary on team level (which is formally obtained by taking the maximum of the team members' individual scores).

As for self-assessed experience, we asked for both empirical finance and market liquidity, which we deem equally relevant for answering the RT-hypotheses. Thus, and because of the anticipated high correlation, we use the average of these two self-reported measures to obtain the individual score. And, in the

[16] It is true that adding two highly correlated variables is unlikely to affect the first principal component of the set of regressors. We nevertheless prefer a parsimonious set of regressors to avoid multicollinearity when expanding on our first-principal-component analysis by performing a multivariate regression (see Section 4.4).

interest of consistency, we again use the maximum of the individual scores to obtain the team level score.

A.2 Workflow quality

We proxy workflow quality with an objectively obtained score of code quality. The latter will be obtained from Cascad (see footnote 6) using a scale from 0 (serious discrepancies) to 100 (perfect reproducibility).

A.3 Paper quality

To obtain quantifiable measures of paper quality, PEs will not only provide verbal feedback on the papers, but also rate the analyses individually and the paper in its entirety. They will do so using a scale from 0 (very weak) to 10 (excellent). The template that PEs will fill out is in Appendix B. As each paper will be evaluated by two PEs, the paper score will be the average overall score of the two PEs.[17]

B Instructions/templates for peer evaluators and research teams

B.1 Template for peer evaluators

[17] Note that in order to remove a possible PE fixed effect we use demeaned PE scores in all of our analysis, as stated earlier in the document.

Evaluator's report

For research team: ______________________________

For each hypothesis and for the full paper, please rate the quality of the analysis, briefly clarify and provide suggestions for possible improvement (a paragraph is sufficient). For your feedback please follow the guidelines of the Journal of Finance: "(P)lease focus on what you see as central weaknesses in terms of (...) the chosen econometric strategy (...). (Y)ou should provide clear and concise reasons why you see your proposed revision as material (...)."

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

Please rate the quality of the analysis:

0 very weak	1	2	3	4	5	6	7	8	9	10 excel- lent

Comments for possible improvement:

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

Please rate the quality of the analysis:

0 very weak	1	2	3	4	5	6	7	8	9	10 excel- lent

Comments for possible improvement:

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

Please rate the quality of the analysis:

0 very weak	1	2	3	4	5	6	7	8	9	10 excel- lent

Comments for possible improvement:

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

Please rate the quality of the analysis:

0 very weak	1	2	3	4	5	6	7	8	9	10 excel- lent

Comments for possible improvement:

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

Please rate the quality of the analysis:

0 very weak	1	2	3	4	5	6	7	8	9	10 excel- lent

Comments for possible improvement:

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

Please rate the quality of the analysis:

0 very weak	1	2	3	4	5	6	7	8	9	10 excel- lent

Comments for possible improvement:

Overall quality of the paper. Please rate the quality of the analysis:

0 very weak	1	2	3	4	5	6	7	8	9	10 excel- lent

Comments for possible improvement:

B.2 Instructions for the research teams

The instruction sheet that will be sent to the research teams is included in the remainder of the document. These pages will be cut out of this document and sent, as is, to the research teams.

After they followed these instructions and uploaded their results including the final PDF paper, we will ask them about how large they believe the dispersion in results to be across teams (i.e., only for Stage 1). We plan to ask something in the following spirit:

> Imagine you receive the short papers of all teams who completed all stages in the project (be aware that 221 teams registered and were given access to the data on January 11, but we expect some to drop out). What do you predict the dispersion in results to be across teams? More specifically, for each hypothesis, we ask you to assess
>
> - the *standard deviation* of reported estimates across teams (i.e., imagine that you collect the estimates y_i of all teams in a single data series, then what is the standard deviation $\frac{1}{n-1}\sum_i (y_i - \bar{y})^2$ of this series?)
> - the *standard deviation* of t-statistics across teams (the t-statistic is defined as the reported estimate divided by the reported standard error).

Instruction sheet for research teams

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

A Context

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

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

B Assignment

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

1. Propose a statistical measure, briefly motivate it, and present the formula to calculate it.

2. For this measure, estimate the average per-year change in percentage terms, based on the full sample (or at least the longest possible period because some series are not available yet at the beginning of the sample). Test it against the null of no change.

3. Report this estimate along with its standard error in four decimals (e.g., "measure Z declined by 1.251% with a standard error 0.241%")

4. Briefly discuss your result.

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

> "We propose measure Z to test hypothesis X because [...]. It is calculated as Z = f(DATA). Implementing it leads to the following result: We reject the null of no change. We find that Y declined as our measure Z declined by 1.251% on average per year where the standard error of this change is 0.421% and the resulting *t*-statistic is 2.971. This result shows [...]"

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

C Hypotheses

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

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

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

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

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

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

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

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

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

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

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

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

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

References

Benjamin, Daniel J., Magnus Johannesson, Valen E. Johnson et al. 2018. "Redefine Statistical Significance." *Nature Human Behavior* :6–10.

Botvinik-Nezer, Rotem et al. 2020. "Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams." *Nature* 582:84–88.

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.

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

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

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

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

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.