Huber, Christoph et al.

**Working Paper**

# Do Experimental Asset Market Results Replicate? High-Powered Preregistered Replications of 17 Claims

I4R Discussion Paper Series, No. 190

**Provided in Cooperation with:**
The Institute for Replication (I4R)

*Suggested Citation:* Huber, Christoph et al. (2024) : Do Experimental Asset Market Results Replicate? High-Powered Preregistered Replications of 17 Claims, I4R Discussion Paper Series, No. 190, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/307930

INSTITUTE for REPLICATION

No. 190
I4R DISCUSSION PAPER SERIES

# Do Experimental Asset Market Results Replicate? High-Powered Preregistered Replications of 17 Claims

Christoph Huber
Magnus Johannesson
Anna Dreber
Michael Kirchler
Felix Holzmeister
Christian König-Kersting
Jürgen Huber

December 2024

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 190

# Do Experimental Asset Market Results Replicate? High-Powered Preregistered Replications of 17 Claims

**Christoph Huber[1], Felix Holzmeister[2], Magnus Johannesson[3], Christian König-Kersting[2], Anna Dreber[3], Jürgen Huber[2], Michael Kirchler[2]**

*[1]Aalto University School of Business, Espoo/Finland*
*[2]University of Innsbruck/Austria*
*[3]Stockholm School of Economics, Stockholm/Sweden*

DECEMBER 2024



E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

# Do experimental asset market results replicate? High-powered preregistered replications of 17 claims

*By* Christoph Huber[†], Felix Holzmeister[†], Magnus Johannesson[†], Christian König-Kersting[†], Anna Dreber, Jürgen Huber, and Michael Kirchler*

*December 9, 2024*

*Experimental asset markets provide a controlled approach to studying financial markets. We attempt to replicate 17 key results from four prominent studies, collecting new data from 166 markets with 1,544 participants. Only 3 of the 14 original results reported as statistically significant were successfully replicated, with an average replication effect size of 2.9% of the original estimates. We fail to replicate findings on emotions, self-control, and gender differences in bubble formation but confirm that experience reduces bubbles and cognitive skills explain trading success. Our study demonstrates the importance of replications in enhancing the credibility of scientific claims in this field.* (*JEL* G12, G41, C91, C92)

[†] The first four authors contributed equally to this work. *C. Huber: Aalto University School of Business (email: christoph.huber@aalto.fi); Holzmeister: University of Innsbruck (email: felix.holzmeister@uibk.ac.at); Johannesson: Stockholm School of Economics (email: magnus.johannesson@hhs.se); König-Kersting: University of Innsbruck (email: christian.koenig@uibk.ac.at); Dreber: Stockholm School of Economics (email: anna.dreber@hhs.se); J. Huber: University of Innsbruck (email: juergen.huber@uibk.ac.at); Kirchler: University of Innsbruck (email: michael.kirchler@uibk.ac.at). We thank Aurélien Baillon, Colin Camerer, and Séverine Toussaert for their helpful comments. We also thank the following original authors who generously provided us with information about their studies and helpful comments: Brice Corgnet, Sascha Füllbrunn, Terrance Odean, and David Schindler. For the opportunity to conduct our experiment in their laboratories, we thank the Innsbruck EconLab, University of Innsbruck; Lab$^2$ and the WZB-TU lab, WZB Berlin Social Science Center; the Vienna Center of Experimental Economics (VCEE), University of Vienna; and WULABS, WU Vienna University of Economics and Business. We furthermore thank Nina Bonge, Nilay Buhlan, Geoffrey Castillo, Levent Neyse, Sebastian Peters, Julian Quandt, Lukas Seewitz, Teresa Steinbacher, and Yaoyao Xu for supporting our research at the labs. This research was funded in part by the Austrian Science Fund (FWF) 10.55776/P29362. We also thank the Jan Wallander and Tom Hedelius Foundation (grants P21-0091 and P23-0098 to A.D.), the Knut and Alice Wallenberg Foundation (grants KAW 2018.0134 and KAW 2023.0363 to A.D.), the Marianne and Marcus Wallenberg Foundation (grant KAW 2019.0434 to A.D.), and Riksbankens Jubileumsfond (grant P21-0168 to M.J.) for financial support. This study was approved by the Institutional Review Board at the University of Innsbruck (no. 118/2023) and the Ethics Board at WU Vienna University of Economics and Business (ref. WU-RP-2023-064). The online appendix accompanying the manuscript is available at osf.io/uxrgk.

Asset price bubbles and crashes lie at the heart of financial markets and come with tremendous costs for individuals, households, and society (Brunnermeier and Schnabel 2016; Miao and Wang 2018; Guerron-Quintana, Hirano, and Jinnai 2023; Hori and Im 2023). Moreover, they are recurring phenomena, tend to follow common patterns, and occur across a wide range of financial and non-financial asset classes such as stocks, bonds, foreign exchange, derivatives, real estate, and commodities (e.g., Galbraith 1994; Kindleberger and Aliber 2011; Brunnermeier and Oehmke 2013). Asset price bubbles are generally defined as positive price deviations from an asset's fundamental value—as such, they represent periods of inefficient pricing. In empirical data, however, attempts to gauge mispricing—and hence, also the identification of causes and potential policy measures—suffer from a joint hypothesis problem (Fama 1970). Typically, an asset's fundamental value is unobservable, implying that estimating deviations hinges on auxiliary assumptions.

Experimental asset markets, by contrast, allow researchers to directly induce a well-defined and, thus, observable fundamental value (see, e.g., Bloomfield and Anderson 2010) so that the existence and determinants of price bubbles can be studied systematically. In experimental asset markets, the literature following and building upon the pioneering work of Smith, Suchanek, and Williams (1988; SSW henceforth) has indeed demonstrated the emergence of bubble and crash patterns in the laboratory and has identified a multitude of factors contributing to their formation. Palan (2013) offers a comprehensive review of over 60 experimental studies employing the SSW paradigm, and Powell and Shestakova (2016) add to that by surveying more recent developments in this literature.

While the high level of control exercised through experimental environments implies important upsides, findings obtained in lab settings also come with limitations beyond concerns regarding validity and generalizability. Since typical experimental asset market designs require six to ten traders to form a single independent observation, data collection is inherently resource-intensive. As a result, the literature is plagued by various methodological, statistical, and practical restraints, potentially jeopardizing the credibility and reliability of empirical claims. Many findings are grounded on a single study examining a particular hypothesis. While several "stylized facts" have emerged through conceptual replications (see, e.g., Palan 2013; Powell and Shestakova 2016), there are barely any direct replications of causal effects, which is hardly surprising given the high cost involved and the

incentives faced by researchers. A potentially even more severe issue is that experimental asset market studies usually rely on very few independent observations: most studies comprise, at most, ten market-level observations per treatment, with many influential contributions relying on six or fewer independent observations.[1] Sample sizes that small curtail the hypothesis tests' statistical power to low levels, making it difficult to detect genuinely true effects, increasing the likelihood of false positive results, and inflating the effect size of statistically significant findings (Ioannidis 2008; Zhang and Ortmann 2013; Maniadis, Tufano, and List 2014). This, in turn, opens the door to questionable research practices, such as selective reporting and p-hacking (Simmons, Nelson, and Simonsohn 2011; A. Gelman and Loken 2014; Brodeur, Cook, and Heyes 2020). Moreover, many experimental setups lack proper randomization of treatments, with data collection occurring sequentially.[2] The lack of proper randomization of treatments weakens causal inference, further increasing the false positive risk and limiting confidence in the validity of reported findings. It is, therefore, important to assess the credibility of reported findings.

An effective method for assessing the credibility of published claims is through the process of replication, which involves testing original hypotheses against new data. Over the past decade, various social science disciplines have started to scrutinize published findings through extensive systematic replication projects. Beginning with the seminal *Reproducibility Project: Psychology* (RPP; Open Science Collaboration 2015), which put 100 original studies published in three leading psychology journals to a replicability test, several other notable large-scale replication projects emerged, including the *Experimental Economics Replication Project* (EERP; Camerer et al. 2016), the *Social Sciences Replication Project* (SSRP; Camerer et al. 2018), the *Management Science Replication Project* (MSRP; Davis et al. 2023), and the *Mechanical Turk Replication Project* (MTRP; Holzmeister et al.

[1] Some examples of prominent studies with six or fewer independent observations (markets) per treatment are: James and Isaac (2000), Lei, Noussair, and Plott (2001), Dufwenberg, Lindqvist, and Moore (2005), Haruvy and Noussair (2006), Haruvy, Lahav, and Noussair (2007), Hussam, Porter, and Smith (2008), Huber and Kirchler (2012), Kirchler, Huber, and Stöckl (2012), Haruvy, Noussair, and Powell (2014), Eckel and Füllbrunn (2015), and Kirchler et al. (2015).

[2] Data for different treatments are typically collected in different experimental sessions without randomization of participants to different sessions. This for instance appears to be the case in the two treatment comparison studies included in our replication project. Andrade, Odean, and Lin (2016) explicitly mention that the treatments were conducted sequentially, implying no randomization. The study by Kocher, Lucks, and Schindler (2019) does not explicitly mention whether or not participants were randomized to treatments but from the data it becomes clear that only one treatment was carried out per session, ruling out randomization to treatments within sessions.

2024). The findings from these projects reveal varying levels of replicability across disciplines. In the RPP, only 36% of the original claims reported as statistically significant were successfully replicated ($n$ = 97). In contrast, the EERP found that 61% of experimental studies published in the *American Economic Review* and *Quarterly Journal of Economics* could be replicated ($n$ = 18). Similarly, the SSRP reported a replication rate of 62% for social science experiments published in *Science* and *Nature* ($n$ = 21). The MSRP demonstrated a somewhat higher success rate, with 70% of the operations management experiments being replicable ($n$ = 10),[3] whereas the MTRP reports a replication rate of 54% for online experiments published in the *Proceedings of the National Academy of Sciences*.

Replications can be categorized into two types: direct and conceptual replications (Dreber and Johannesson 2024). Direct replications involve testing the same hypothesis as the original study against new data, utilizing the same research design and analysis. In contrast, conceptual replications test the same hypothesis in new data but do so using a different research design or analytical approach. Within experimental asset market research, there has been a notable prevalence of conceptual replications, as many settings build upon established research designs such as the paradigms put forth by Smith, Suchanek, and Williams (1988), Plott and Sunder (1988), or Smith et al. (2014). Direct replications are rare in this literature. A notable exception is a recent study by Corgnet et al. (2023), which attempts—and fails—to replicate Plott and Sunder's (1988) seminal study on information aggregation in asset markets. Corgnet et al. (2023) effectively showcase some of the challenges within this area of research, particularly the issue of low statistical power in original studies. The model put forth by Maniadis, Tufano, and List (2014) highlights that studies yielding strong and surprising results—often those that are most likely to be featured in prestigious journals—are typically less likely to reflect genuine associations. Against this backdrop, it is striking that there have been relatively few attempts to directly replicate findings in the experimental asset markets literature, particularly since even a single replication attempt can significantly enhance the likelihood of achieving accurate inference (Dreber et al. 2015; Maniadis, Tufano, and List 2017).

[3] The MSRP (Davis et al. 2023) reports a replication rate of 70%, but the project involves replications of multiple hypotheses and sites for some of the papers and weighting each paper equally, the results reported in Table 3 in their paper implies a replication rate of about 60% rather than 70%.

This study reports the results of high-powered, preregistered replications of 17 claims from the experimental asset markets literature, with replication sample sizes varying between 1.6 and 9 times the original sample sizes (7.2 times the original, on average). The 17 findings put to a replicability test were taken from four prominent articles published in leading journals in economics and finance: the *American Economic Review*, the *Journal of Finance*, the *Review of Financial Studies*, and the *Review of Finance*. In particular, we attempt to replicate published findings on the association between asset market pricing and emotions, self-control, experience, and gender and re-examine empirical claims regarding traders' characteristics—cognitive reflection, fluid intelligence, and Theory of Mind—explaining individual trading success.

In selecting the original studies for replication, we focused specifically on studies examining *behavioral* motives for mispricing—a literature in which many influential results so far rely on only a single study.[4] In particular, we aim to replicate focal findings from two prominent studies focusing on the causal impact of self-control and emotions on mispricing: Andrade, Odean, and Lin (2016; AOL henceforth) and Kocher, Lucks, and Schindler (2019; KLS henceforth).[5] We gathered data for two treatments from AOL to test their two key hypotheses about excitement increasing overpricing, and collected data for two treatments from KLS to evaluate their two main findings concerning the impact of low self-control increasing bubble formation. Overall, our study involved collecting data for 166 markets with a total of 1,544 participants. Each session included either the two treatments from AOL or the two treatments from KLS to ensure proper randomization of treatments within sessions. The replication sample sizes were determined based on a priori power

[4] The literature on experimental asset markets can roughly be divided into two strands. The first investigates the implications of institutional factors such as constraints on short-selling and a market's cash-to-asset ratio (e.g., Caginalp, Porter, and Smith 1998; 2001; Haruvy and Noussair 2006; Haruvy, Noussair, and Powell 2014; Noussair and Tucker 2016; Kirchler et al. 2015; Razen, Huber, and Kirchler 2017; Weitzel et al. 2020). The second strand, which we focus on in this study, examines the impact of behavioral factors and individual traits such as inexperience, confusion about fundamentals, risk preferences, cognitive abilities, emotional states, self-control, and gender (e.g., Dufwenberg, Lindqvist, and Moore 2005; Kirchler, Huber, and Stöckl 2012; Cueva and Rustichini 2015; Eckel and Füllbrunn 2015; Andrade, Odean, and Lin 2016; Breaban and Noussair 2018; Bosch-Rosa, Meissner, and Bosch-Domènech 2018; Kocher, Lucks, and Schindler 2019).

[5] AOL uses an emotion manipulation to induce excitement and there is a substantial body of work testing if emotions affect economic behavior, particularly risk-taking, but the jury is still out on whether there are important effects of emotion manipulations on economic behavior (e.g., Wake, Wormwood, and Satpute 2020; Marini 2023). The study by KLS uses a so-called ego-depletion paradigm to manipulate self-control, but it is controversial whether such a paradigm successfully manages to manipulate outcomes on subsequent tasks (e.g., Hagger et al. 2016; Friese et al. 2019; Dang et al. 2021). Our replications help shed light on these controversial issues.

calculations to have at least 90% statistical power to detect two-thirds of the original effect sizes of the treatment comparisons in AOL and KLS at the 5% significance level (as in Holzmeister et al. 2024). Furthermore, our extensive dataset enables us to replicate four key hypotheses from Eckel and Füllbrunn (2015; EF henceforth) about the association between gender composition and bubble indicators in their meta-analysis.[6] By gathering data on cognitive reflection, fluid intelligence, and Theory of Mind from experimental participants, we can also conceptually replicate nine key hypotheses in Corgnet, Desantis, and Porter (2018; CDP henceforth), exploring the effect of individual-level characteristics on trading performance. Moreover, we revisit the impact of market experience on overpricing based on paired comparisons of bubble measures between two repetitions of market trading.[7]

To evaluate replicability, we use two replication indicators: the statistical significance indicator, defining a successful replication as a statistically significant ($p < 0.05$) effect in the same direction as the original study, and the relative effect-size indicator, defined as the ratio between the effect size estimate in the replication and the original estimate (Dreber and Johannesson 2024).

We could neither replicate AOL's finding that excitement causally contributes to overpricing nor KLS's result regarding the causal impact of low self-control on overpricing. The point estimates in all four replication tests point in the opposite direction of the original claims. Likewise, we failed to replicate the four meta-analytic results from EF suggesting a negative correlation between the fraction of female traders in a market and the extent of bubbles, with all point estimates pointing in the opposite direction. Regarding the conceptual replication of CDP, we find support for three out of six results that were originally reported as statistically significant: the association between trader earnings and cognitive reflection, fluid intelligence, and Theory of Mind. However, the estimated effect sizes for fluid intelligence and Theory of Mind are substantially lower than those in the

[6] We attempt to replicate the meta-analytical results reported in EF, which are based on previous experimental data. Note that EF also conducted original market experiments with single- and mixed-gender compositions, which we did not attempt to replicate.

[7] We refer to the replication of the CDP as a conceptual replication, since the original findings were established based on Plott and Sunder's (1988)—rather than SSW's—market paradigm. The replication test of the "experience effect" is not tied to a specific original result but rather re-examines a stylized fact in literature, with many landmark contributions such as Dufwenberg, Lindqvist, and Moore (2005), Hussam, Porter, and Smith (2008), and Kopanyi-Peuker and Weber (2021) providing evidence for a moderating effect of experience.

original study (about 60% and 40%, respectively). The three results, which were originally reported as statistically significant and which we could not replicate, have in common that they involve interaction effects among cognitive variables. Three more interaction effects in CDP that were reported as statistically insignificant in the original study also turned out not to be statistically significant in our replication attempt, confirming the original null findings. Consequently, we find no evidence of moderation effects between cognitive reflection skills, fluid intelligence, and Theory of Mind. Overall, three (21.4%) out of the 14 results originally reported as statistically significant were successfully replicated according to the statistical significance indicator, with an average relative effect size of only 2.9%. Apart from the replication results tied to specific original claims, our study finds support for the stylized fact that market experience curbs the extent of mispricing, with bubble indicators cut in about half in the second repetition of market trading.

Our study centers on behavioral factors in experimental asset markets and highlights the important role of replication in bolstering the credibility of scientific claims within this field by presenting new evidence on 17 claims in the literature. Through high-powered replications, we were unable to confirm 11 hypotheses while confirming only three that had previously been reported as statistically significant. Additionally, we validate three negative findings concerning the moderating effects of cognitive skills on trading performance and provide strong support for the experience hypothesis, suggesting that market experience can effectively reduce mispricing. Our study makes a vital contribution to the literature on experimental asset markets, identifying several likely false positive findings and offering novel and well-powered insights into the behavioral determinants of market behavior and trading performance. It also underscores the essential nature of replication in enhancing the process of knowledge generation and accumulation of evidence in economic research.

# I. Replication Protocol

We preregistered the study design, analyses, and statistical tests in a detailed pre-analysis plan prior to the start of the data collection, available at osf.io/aepxt.[8] Unless explicitly noted, the study design and all analyses and tests reported follow the preregistration exactly. We explicate any deviations from the pre-analysis plan in section H of the Online Appendix.

## *A. Treatments and Market Settings*

Our study comprises four conditions, each of which involves a treatment manipulation before participants take part in an asset market experiment: *(i)* the *Excitement* condition and *(ii)* the *Calm* condition from AOL,[9, 10] and *(iii)* the *Low Self-Control* condition (*LowSC*) and *(iv)* the *High Self-Control* condition (*HighSC*) from KLS. In the *Excitement* and the *Calm* conditions, the treatment manipulation involved watching a movie clip from either an action movie or a placid movie intended to manipulate participants' emotions toward excitement or calmness.[11] In the *LowSC* and the *HighSC* conditions, the treatment manipulation involved

[8] Before filing the preregistration, we contacted the original authors of AOL, KLS, and CDP, and informed them about our intent to replicate key findings of their articles, and asked them to share materials (e.g., instructions, stimuli, software, etc.) used in their study. However, we did not ask them to approve of our replication designs.

[9] AOL's published article involves treatment comparisons across three conditions: *Calm*, *Excitement*, and *Fear*. Footnote 3 in AOL states that they collected data for another treatment condition (*Sad*, $n = 8$) not reported in the article "for simplicity." Based on the data kindly shared by the original authors, it appears that the *Calm* treatment was initially split into two conditions: *Neutral* ($n = 8$) and *Low Arousal* ($n = 7$). In their analysis code, the *Low Arousal* indicator is replaced with *Neutral* and the merged condition is referred to as *Calm* in the published study. AOL's article does not mention the *Low Arousal* condition and it is unclear which stimuli were used in the two initial conditions that ended up as a single treatment in the published manuscript. In our replication, we used the stimuli attributed to the *Calm* condition as per AOL's article.

[10] Footnote 3 in AOL clarifies that they ran 24 markets in the *Excitement* condition before completing the data collection for the *Calm* and *Fear* conditions. The sequential, treatment-wise collection of data undermines statistical inference, ruling out interpreting treatment differences as causal effects. Moreover, AOL noted that they decided to only run 16 markets for *Calm* and *Fear* "[b]ecause of the high cost of the experiments."

[11] To avoid deceiving participants, our instructions concerning the treatment manipulation differed from those in AOL, who motivated the inclusion of a video clip as follows (p. 461): "When the Practice Session is over, it will take some time to re-initialize and configure the trading program. The preparation could take around 5–8 min. Because the waiting is a bit long, we will play a video clip. We intend to use the video in another experiment and want to get some feedback from you. After you've finished watching the clip, please answer a few questions about it. Note that the video is not related to your earnings today. So thank you in advance for helping out." To circumvent deception in our replication, we instructed participants that "In this part of the experiment, you will watch a short movie clip and answer a few questions about it." AOL also included a question to assess whether participants could guess the purpose of the study, which was not included in the replication.

either completing a hard (*LowSC*) or an easy version (*HighSC*) of the Stroop task (Stroop 1935), with the hard version intended to deplete participants' self-control.

In the subsequent asset market experiment, participants in the same treatment condition were grouped into markets resembling the design introduced by Smith et al. (1988), in which shares of a long-lived asset with risky dividend payments were traded in a continuous double auction market over ten periods of 120 seconds each. The parameterizations applied in our replication attempt are summarized in Table 1, alongside those used by AOL and KLS. Our parameterization differed slightly from AOL but aligned with KLS, with the exception that we permitted markets of eight traders instead of ten in cases where not enough lab participants showed up for a session.[12] Dividend payments per share were either 0 or 10 points at the end of every period, each occurring with equal probability, which implies a linearly declining fundamental value typical for the SSW setting.[13] At the beginning of the first trading period, one-half of the participants were randomly assigned to receive an initial endowment of 20 shares and 3,000 points; the other half received an endowment of 60 shares and 1,000 points. Shares and points carried over from one period to the next.

AOL and KLS only carried out one repetition of the market trading, meaning each participant was involved in only one instance of the asset market experiment. We carried out two repetitions of each market, keeping the traders and market parameters constant. Each participant was thus involved in two subsequent instances (repetitions) of the asset market experiment. The two possible endowments were independently assigned at random in each repetition. Our replication tests of AOL and KLS are solely based on the first repetition.[14] However, repeating the market trading twice allows us to assess whether potential treatment effects persist among more experienced traders in the second run.

[12] In his review of the literature on bubbles in continuous double auction markets, Palan (2013) noted that the number of traders in SSW-like markets typically varies between 6 and 15, and argued that there is no evidence of a systematic effect of the number of traders on pricing.

[13] While AOL and KLS used a single sequence of realized dividends for all markets, we applied the same dividend structure as KLS with five low and five dividends in each market, but randomized the order of dividend realizations across markets and repetitions. The pre-analysis plan did not specify that we deviated from KLS in this regard as we misinterpreted KLS's description of dividend sequences in their manuscript and their experimental instructions and thought to mirror their implementation.

[14] Similarly, the replication hypothesis tests of EF and CDP are solely based on the first repetition, while the market outcomes of the second repetition (and/or the average outcomes across both repetitions) enter robustness tests.

However, the primary intent of having two repetitions was to test for systematic differences in mispricing between repetitions, in line with extensive literature documenting experience effects in experimental asset markets (e.g., Dufwenberg, Lindqvist, and Moore 2005; Hussam, Porter, and Smith 2008; Kopányi-Peuker and Weber 2021).

**Table 1. Treatment overview and market parameterization.** The table summarizes the market parameterizations used in the original studies by Andrade, Odean, and Lin (2016) and Kocher, Lucks, and Schindler (2019) as well as the parameterization implemented in the replication experiments, which mirrors the parameterization in Kocher, Lucks, and Schindler (2019).

| | **Original studies** | | **Replication** | |
|---|---|---|---|---|
| | *Andrade et al.* | *Kocher et al.* | *Andrade et al.* | *Kocher et al.* |
| *No. of markets per condition* | 24 (*Excitement*)<br>15 (*Calm*) | 8 (*LowSC*)<br>8 (*HighSC*) | 31 (*Excitement*)<br>31 (*Calm*) | 52 (*LowSC*)<br>52 (*HighSC*) |
| *Traders per market* | 9 | 10 | 8–10† | 8–10† |
| *Periods* | 15 | 10 | 10 | 10 |
| *Period length* | 210 s | 120 s | 120 s | 120 s |
| *Dividend* | 0, 8, 28, 60 | 0, 10 | 0, 10 | 0, 10 |
| *Exp. dividend* | 24 | 5 | 5 | 5 |
| *FV (t = 0)* | 360 | 50 | 50 | 50 |
| *FV (t = T)* | 24 | 5 | 5 | 5 |
| *Endowments (shares, cash)* | (1, 1800),<br>(2, 1440),<br>(3, 1080) | (60, 1000),<br>(20, 3000) | (60, 1000),<br>(20, 3000) | (60, 1000),<br>(20, 3000) |
| *C/A ratio (t = 0)* | 2 | 1 | 1 | 1 |
| *C/A ratio (t = T)* | 44 | 19 | 19 | 19 |
| *Market mechanism* | continuous double auction | continuous double auction | continuous double auction | continuous double auction |
| *Repetitions* | 1 | 1 | 2 | 2 |

**Notes.** † We were targeting ten traders per market but allowed for markets with only eight traders if not enough participants showed up for a particular lab session. All markets conducted in the same session had the same number of traders. *FV* = fundamental value, *C/A ratio* = cash-to-asset ratio.

## *B. Data Collection*

We collected data at several experimental economics laboratories across Austria and Germany; all experimental sessions were conducted in German. The instructions were based on those used by AOL and KLS, and the experimental software (programmed in oTree; Chen,

Schonger, and Wickens 2016) resembled their original implementations;[15] both are available at osf.io/bm2dx. We only invited participants who had not previously participated in an asset market experiment, matching one of the inclusion criteria in KLS.[16,17]

Based on our preregistered a priori power calculation, we collected data for 31 markets per condition in the replication of AOL and 52 markets per condition in the replication of KLS for a total of 166 markets. This implies that our replication sample sizes (in terms of the number of independent market-level observations) are 1.6 times larger than the original sample size for AOL and 6.5 times larger for KLS. These sample sizes provide us with 90% statistical power to detect at least two-thirds of the original effect size at the 5% significance level in a two-tailed test for the two primary replication tests conducted for AOL and KLS.

To ensure effective randomization of traders to treatment conditions, each experimental session consisted of two markets. In each session, we included either the *Excitement* and the *Calm* conditions from AOL or the *Low Self-Control* (*LowSC*) and the *High Self-Control* (*HighSC*) conditions from KLS. Participants in each session were randomly assigned to either of the two conditions (markets), with an equal number of eight or ten traders

[15] Our experimental instructions closely follow the original German-language instructions from KLS. Our software implementation differed from KLS only in technical aspects of the trading. In contrast to KLS's implementation, traders could not choose which offer to trade against but always traded against the best available offer. Moreover, traders could not post limit offers that would be automatically cleared against existing offers by the exchange. We cannot rule out that these differences may affect the prevalence or the extent of mispricing; however, we are not aware of any evidence supporting this conjecture. Additionally, in our replication, the stakes were increased to align with the current rates used in the involved laboratories; research by Kocher, Martinsson, and Schindler (2017), however, suggests that such adjustments are unlikely to systematically impact pricing.

[16] KLS also excluded participants *"potentially familiar with the cognitive reflection test or the Stroop task"* (pp. 2158–2159). We did not apply these exclusion criteria as the information about participants' previous involvement in experiments utilizing the cognitive reflection test or the Stroop task was not consistently available across all laboratories where the replication took place. Screening participants ex ante for their familiarity with the tasks is challenging because it would require revealing details about the tasks, which could inadvertently lead to familiarity with them. Furthermore, it is important to note that we are replicating the main treatment effects in KLS that do not rely on the cognitive reflection test data. The original study gathered this data to examine whether cognitive reflection served as a mechanism for potential treatment effects; however, they found no statistically significant moderating effects.

[17] As color perception is crucial in the Stroop task, used to manipulate self-control in KLS, we emphasized in the invitations to the sessions for the *Low Self-Control* and the *High Self-Control* conditions that participants who suffer from color blindness (achromatopsia) should refrain from participating in the experiment. It is unclear whether KLS included this information in their invitations to the experiment; however, they did inquire if participants suffered from any deficiency in perceiving colors, and reported that five out of 400 participants responded affirmatively. We posed the same question as part of our replication protocol, and 17 out of 904 participants (1.9%) indicated they had some form of achromatopsia; the question provided options for two forms of achromatopsia, an "other" category, and a "no impairment" option, with the 1.9% reflecting those who did not select "no impairment").

(participants) in each of them. We aimed to have ten traders per market, but we still conducted the session even if there were only eight traders per market due to no-shows. Markets in the same session consistently had the same number of traders. For instance, if 19 participants attended a session, we operated two markets with eight participants each. Note that we only required effective randomization between the two conditions in the replication of AOL and the two conditions in the replication of KLS to causally identify the focal effects. Since we do not compare treatment effects across the two primary studies, randomization between the AOL and KLS conditions was not a requirement. Also note that our replications of the meta-analysis in EF and of CDP pertain to correlational rather than causal effects and, thus, do not rely on randomization.[18]

For the conceptual replication of CDP, we complemented the experimental protocol to elicit participants' cognitive reflection (Cognitive Reflection Test; Frederick 2005; Toplak, West, and Stanovich 2014), fluid intelligence (Advanced Progressive Matrices; Raven, Raven, and Court 1998; Raven and Raven 2008), and Theory of Mind ("Reading the Mind in the Eyes" test; Baron-Cohen et al. 1997; 2001) toward the end of each session. Note that KLS included Frederick's (2005) original inventory and a risk preference elicitation task (Dohmen et al. 2011) between the Stroop task and the markets to test if cognitive reflection and risk preferences were potential mechanisms for an eventual treatment effect. To keep the experimental protocol of the replication as close as possible to the original, we included the three-item cognitive reflection test and the same risk elicitation task despite not using the data in any analysis.[19] Since the seven-item cognitive reflection test administered at the end of the experiment for the replication of CDP comprises Frederick's (2005) inventory, we replaced the three items used by KLS with items 2–4 of Thomson and Oppenheimer's (2016) CRT-2 inventory to avoid using the same items twice within the same experiment.

In the original study by KLS, participants earned experimental points converted to Euros using an exchange rate of 500 points = €1.00. Participants received €4 as a show-up fee, €3

[18] EF conducted a meta-analysis of several published studies to test whether the fraction of female traders in a market is associated with four different "bubble measures." Our data collection results in a dataset comprising about five times the number of markets considered in EF's meta-analysis, which we use to put their meta-analytic claim to a replicability test.

[19] The cognitive reflection test and the risk preference elicitation task in KLS were incentivized; we adopted the same incentive schemes for the two tasks to maintain consistency with their design. However, we chose not to incentivize the seven-item cognitive reflection test, the fluid intelligence task, or the Theory of Mind task toward the end of each experimental session, aligning with CDP's setting, which did not use incentives for these tasks.

for completing the Stroop task (irrespective of their performance), €0.50 per correct answer on the three-item cognitive reflection test, between €0.20 and €4.20 in the risk preference elicitation task, and earned, on average, about €8 in the market, resulting in average earnings of €18.27. In the original study by AOL, participants received a $5 show-up fee and earned, on average, an additional $21.68. For the replication of KLS, we implemented the same incentives for the side tasks but adjusted the exchange rate to 160 points = €1.00 and the show-up fee to €5 to match the rates used in the involved labs. We used the same market payments and show-up fee in the two conditions replicating AOL. In all conditions, participants in our replication experiment were paid based on either repetition 1 or 2 determined at random to avoid portfolio-building and cross-task contamination effects (e.g., Cubitt, Starmer, and Sugden 1998; Azrieli, Chambers, and Healy 2018). On average, participants in our study earned €30.28 (*sd* = 8.79; min = 5.00, max = 68.00) in the two conditions replicating AOL and €37.07 (*sd* = 11.02; min = 9.20, max = 94.40) in the two conditions replicating KLS; the higher payments in those two conditions were due to the inclusion of payments for the Stroop task, cognitive reflection test, and risk preference task. On average, the experiment took 1 hour and 50 minutes; thus, the average earnings exceeded the targeted hourly rate of €15.

### *C. Key Variables*

The key outcome variables used in the original studies by AOL and KLS are relative deviation (*RD*), peak overpricing ($RD_{\text{MAX}}$), and relative absolute deviation (*RAD*), measured on the market level and separately for each of the two repetitions (see, e.g., Stöckl, Huber, and Kirchler 2010). As a measure of overpricing, *RD* is defined as the average difference between the market price and its fundamental value across trading periods; $RD_{\text{MAX}}$ captures peak overpricing, defined as the maximum discrepancy between the market prices and the fundamental value. *RAD* is determined as the average absolute difference between market prices and fundamental values across trading periods relative to the fundamental value, serving as a measure of market mispricing.

In the replication of EF, we used the four outcome measures considered in the original article: average bias (*AB*), positive deviation (*PD*), boom duration, and bust duration. *AB* is

defined as the average discrepancy between median market prices and fundamental values across trading periods, whereas *PD* is given as the average absolute difference between median prices and fundamentals. Thus, similar to *RD* and *RAD*, *AB* and *PD* are measures of overpricing and mispricing, respectively. Boom duration (bust duration) is defined as the greatest number of consecutive periods for which the median price exceeds the fundamental value. Section B.1 in the Online Appendix provides a comprehensive overview of all outcome measures entering the replication tests, including formal definitions of the variables.

While the replications of AOL and KLS test for causal effects on outcomes induced through exogenous treatment variations, the replications of EF and CDP test for correlational effects between outcome variables and independent variables. In the replication of EF, the focal independent variable is the share of female traders in a market. In the replication of CDP, the key independent variables are cognitive reflection, fluid intelligence, and Theory of Mind. Following CDP, we used the extended version of the cognitive reflection test, which adds four questions developed by Toplak, West, and Stanovich (2014) to the three original items devised by Frederick (2005). Regarding fluid intelligence, we slightly deviate from CDP's original protocol and implement the test used by Farago et al. (2022). While CDP used Raven's (1941) Standard Progressive Matrices test as a measure of fluid intelligence, we used the second set of Raven's Advanced Progressive Matrices (Raven, Raven, and Court 1998; Raven and Raven 2008). As in CDP, we used the odd-numbered items of the 36-item inventory, resulting in a set of 18 matrices (see also Jaeggi et al. 2010) and limited the duration of the test to 10 minutes. Finally, we followed CDP's protocol with respect to the implementation of the "Reading the Mind in the Eyes" test to elicit participants' Theory of Mind skills (Baron-Cohen et al. 1997). Each item involved an image of the eyes of an individual, and the participant had to choose one of four feelings that best describes the mental state of the person whose eyes are shown. The test was limited to 10 minutes and comprised 36 items (Baron-Cohen et al. 2001). All three participant-level measures were coded as the number of correct answers (0–7, 0–18, and 0–36, respectively) and treated as continuous predictor variables in the regression analyses.

### *D. Replication Indicators*

We conducted replication tests of two key results in AOL, two key results in KLS, four key results in EF, and nine key results in CDP. Below, in describing each hypothesis and test, we refer to these 17 replication tests as "replication hypothesis tests" to distinguish them from additional tests that do not replicate a specific original result.

We report the results for two replication indicators: the statistical significance indicator and the relative effect size indicator (Dreber and Johannesson 2024). For original results reported as statistically significant, the statistical significance indicator is binary and is defined as a statistically significant effect ($p < 0.05$; two-tailed test) in the same direction as in the original study. The relative effect size indicator is defined as the ratio of the effect size estimate in the replication study to that of the original study. For CDP, we only report the results for the relative effect size indicator for the six replication results that were reported as statistically significant in the original study ($\alpha = 0.10$ was used as the statistical significance threshold in the original study, with $p < 0.05$ for four of the six positive results).

We estimate the two replication indicators for each replication hypothesis test but also pool the indicators for each replication study. This includes calculating the fraction of results that replicate according to the statistical significance indicator and the average relative effect size for each of the four articles selected for replication. Additionally, we pool the two replication indicators across the four replication studies based on the study-level averages per article. The pooled results are estimated separately for original results reported as statistically significant and for original null results. The latter only applies to the replication of CDP, which involves replicating three original null results.[20]

To estimate relative effect sizes, we converted both the original effect sizes and the replication effect sizes in AOL and KLS to Cohen's $d$ units (i.e., standardized mean differences). Following Szucs and Ioannidis (2017), the conversion of test statistics obtained from unpaired $t$-tests to Cohen's $d$ units is given by $d = 2 \cdot t \cdot n^{-0.5}$, where $n$ denotes the sample size, and $t$ is the $t$-statistic. Note that the primary claims in the original study by KLS were

[20] For the three original null results in CDP, a successful replication according to the statistical significance indicator is defined as a two-tailed $p$-value > 0.05, whereas a two-tailed $p$-value < 0.05, irrespective of the direction of the effect, is considered a replication failure. Since "absence of evidence is not evidence of absence" (Altman and Bland 1995), evaluating null effects in terms of replicability is inherently challenging (see, e.g., Patil, Peng, and Leek 2016; Pawel et al. 2024).

obtained using Mann-Whitney *U*-tests rather than two-sample *t*-tests. To estimate the original effect sizes of KLS, we re-evaluated the two hypotheses using unpaired *t*-tests based on the original data (we refer to Section D.1 in the Online Appendix for details). For the replication hypothesis tests of EF, the relative replication effect size was determined as the ratio of the Spearman correlation coefficients obtained in the replication test and the original study. For the replication hypothesis tests of CDP, we apply the same conversion to Cohen's *d* units as for AOL and KLS to approximate the original effect sizes and the replication effect sizes based on the *t*-values of the particular regression coefficients, defining the sample size *n* as the number of participants (*n* = 167 in CDP and *n* = 1,542 in our replication study).[21]

### *E. Hypotheses, Tests, and Statistical Power*

As noted above, we examined the replicability of 17 key results from four papers relating to the role of emotions, self-control, and gender differences in bubble formation, as well as the relationship between cognitive skills and trading success. Related to these replication tests, we also conducted a number of preregistered secondary hypothesis tests for consistency or additional insights, as well as various preregistered robustness tests. Additionally, we revisit the experience hypothesis without relating our tests to a specific original result. All hypotheses and tests are described in detail in the Results section.

In evaluating hypotheses, we adhere to our preregistration (osf.io/aepxt) and interpret two-sided *p*-values below 0.05 as "suggestive evidence" and two-sided *p*-values below 0.005 as "statistically significant evidence" (Benjamin et al. 2018). However, as noted above, a 5%-threshold is applied to determine whether or not a replication is successful according to the statistical significance indicator of replication.

The sample size in our study was based on having at least 90% power to detect two-thirds of the original effect size at the 5% level (in a two-tailed test) for the two replication hypothesis tests of AOL and KLS. We powered the study to have at least 90% power to detect two-thirds of the original effect size to account for the empirical

[21] Two out of 1,544 participants that started the experimental sessions dropped out during the experiment due to illness; one without making any trades and one during the Theory of Mind task. Consequently, the replication tests of CDP involve *n* = 1,542 observations; see section H in the Online Appendix for details.

observation that even true positive original findings tend to have inflated effect sizes, which is an immediate consequence of insufficient statistical power in the original studies (Ioannidis 2008; Zhang and Ortmann 2013; Maniadis, Tufano, and List 2014).[22] The a priori power calculations resulting in sample sizes of 62 markets for the replication of AOL and 104 markets for the replication of KLS are available at osf.io/rf8mc. We refer to section A of the Online Appendix for details about the a priori power calculations.

For the replication tests of the meta-analytical results in EF, we have an approximately five times larger sample size than the original study. This gives us a statistical power of 90% power to detect a correlation coefficient of 0.25 at the 5% significance level. Consequently, we have 90% power to detect 52.1%, 70.8%, 63.7%, and 47.0% of the original correlation coefficients in the four replication hypothesis tests. This implies that the minimum detectable effect size for one of the four tests is slightly larger than our target of two-thirds of the original effect size.

For the conceptual replication tests of CDP, the replication sample size is approximately nine times larger than in the original study. We estimated the fraction of the original effect size we would have 90% statistical power to detect at the 5% significance level for each of the six replication hypothesis tests deemed "statistically significant" based on an $\alpha = 10\%$ threshold in the original study. The statistical power to detect two-thirds of the original effect size exceeds 90% for the six replication tests of CDP for which the original study reports a statistically significant effect, which implies that the minimum detectable effect size is even lower than our target of two-thirds of the original effect size.[23]

The statistical power of the 14 replication hypothesis tests pertaining to original results reported as statistically significant, expressed as the percentage of the original effect size we could detect with 90% power at the two-tailed 5% significance level ($\delta_{\text{pre}}$), is reported in the supplementary tables in the Online Appendix tabulating the replication hypothesis test

[22] In both the *SSRP* (Camerer et al. 2018) and the *MTRP* (Holzmeister et al. 2024) replication projects, the relative replication effect size of studies that were successfully replicated was about 70%; the *MTRP* also used the same target power of having 90% statistical power to detect two-thirds of the original effect size in their replications.

[23] Note that the eventual sample size for the replication tests of CDP were unknown ex-ante, since we targeted markets of ten traders each but allowed for markets of eight in the case of no-shows. The power calculations were carried out ex ante as part of the pre-analysis plan and were based on an average of nine traders per market. The actual average number of traders turned out to be 9.3, implying that the eventual minimum detectable effect sizes in the replication hypothesis tests are even smaller than expected in the preregistration.

results.[24] On average, for the 14 replication hypothesis tests of original results reported as statistically significant, we have 90% power to detect 54.0% of the original effect size; the average replication sample size for the 17 replication hypothesis tests is 7.2 times larger than that of the original studies.[25]

## II. Results

Following the typology proposed by Dreber and Johannesson (2024), we refer to the replications of AOL, KLS, and EF as direct replications with new data from a similar population.[26] In contrast, the replication of CDP is considered a conceptual replication with new data from a similar population. The reason for classifying the replication of CDP as a conceptual rather than a direct replication is that CDP's design is based on the market setting introduced by Plott and Sunder (1988) rather than the market paradigm initiated by Smith, Suchanek, and Williams (1988) employed in this study.[27] All data and code used to generate the results presented below are available at osf.io/sr4nv.

[24] We also report the post-hoc minimum detectable effect size ($\delta_{post}$) that we had 90% statistical power to detect for all hypothesis tests (but not the robustness tests), expressed in the tests' units of measurement. The $\delta_{post}$ estimates are reported in the supplementary tables in the Online Appendix Tables summarizing the replication results. See section A in the Online Appendix for details.

[25] Instead of averaging minimum detectable effect sizes and sample sizes across the 17 replication hypothesis tests, they can be averaged for each of the four studies first and then aggregated across studies. Weighting each study equally, the average minimum detectable effect size is 55.5% and the replication sample exceeds the original sample size by a factor of 5.5.

[26] As noted above, our research design follows the SSW paradigm with a long-lived asset traded across multiple periods, and features its key characteristics. Over time, the cash-to-asset ratio increases, while the asset's fundamental value declines—a pattern found to usually generate considerable overpricing (e.g., Caginalp, Porter, and Smith 2001; Dufwenberg, Lindqvist, and Moore 2005; Kirchler, Huber, and Stöckl 2012; Kocher, Lucks, and Schindler 2019). While our parameters slightly differ from the particular implementation of AOL, we consider the design sufficiently close to the original study to be defined as a direct rather than a conceptual replication. We acknowledge, however, that this classification is not obvious and that there will always be borderline cases. This also applies to the replication of the meta-analysis in EF, where the parameterization in the 35 studies included in the meta-analysis slightly differs from the parameterization in our replication. We cannot dismiss the possibility that differences in the market parameterization have affected our replication results. However, we are not aware of any empirical evidence pointing at systematic differences attributable to the parameterization of dividend sequences, the period length, or the like, and the findings in the original articles are not qualified in light of the particular market settings. Consequently, we deem our replication tests adequate and diagnostic of the original studies' focal claims.

[27] The market environment introduced by Plott and Sunder (1988) differs in various aspects from the bubble environment put forth by Smith, Suchanek, and Williams (1988). Particularly, the setting in Plott and Sunder (1988) is mainly concerned with information aggregation, focusing on private rather than public information about the asset value, and consists of a sequence of independent one-period markets instead of multi-period markets with long-lived assets and endowments carrying over from one period to the next.

### *A. Direct Replication of Andrade, Odean, and Lin (2016)*

*Manipulation Check.*—Following AOL's protocol, our replication attempt commenced with an ex-ante manipulation check of the movie clips used in the Excitement ("Knight & Day") and Calm ("Peace in the Water") conditions.[28] The manipulation check was conducted on Prolific. Participants were randomized to the two conditions (with $n = 95$ in the Excitement condition and $n = 103$ in the Calm condition). We used two-sample $z$-tests to test for differences in the proportions of participants choosing the "excited/eager/enthusiastic" and the "calm/relaxed/peaceful" options to describe their emotions while watching the movie clip between conditions. The observed fraction of participants choosing "excited/eager/enthusiastic" was 67.4% in the Excitement condition and 2.9% in the Calm condition ($z = 9.576$, $p < 0.001$). The observed fraction choosing "calm/relaxed/peaceful" was 4.2% in the Excitement condition and 68.0% in the Calm condition ($z = 9.263$, $p < 0.001$). Consequently, the treatment manipulation was deemed successful according to our preregistered criteria; we refer to section C.4 in the Online Appendix for details.

*Replication Hypothesis Tests.*—The replication of AOL involves two replication hypotheses, conjecturing that *(i)* overpricing ($RD$) and *(ii)* peak overpricing ($RD_{\text{MAX}}$) are higher in the Excitement condition than in the Calm condition in the first repetition of the experiment.[29] Figure 1 plots the average period-by-period market price for the two treatment conditions in the replication of AOL, separated for the two repetitions. The mean price developments in repetition 1 follow the pattern typically observed in SSW markets, with prices substantially exceeding the fundamental value during the intermediate trading phase and collapsing toward the end of market trading. Yet, eyeballing the figure already indicates that the extent of mispricing hardly differs between the two treatment conditions.

[28] We filed a separate preregistration for the manipulation check, which is available at osf.io/eqy42. For details about the design and implementation, we refer to section C.4 in the Online Appendix and the preregistration; all data and code pertaining to the manipulation check is available at osf.io/z4cty.

[29] Based on a sample of 39 markets, the original study by AOL reports sizable and statistically significant treatment effects for both outcome measures. Prices, on average, overshot fundamental values by 152.1% in the *Excitement* condition ($n = 24$) and by 85.1% in the *Calm* condition ($n = 15$), implying a treatment difference of 67.0 percentage points for $RD$ ($t(37) = 5.380$, $p < 0.001$). The maximum relative deviations, $RD_{\text{MAX}}$, were 266.8% and 173.1% in the *Excitement* and the *Calm* conditions, respectively, implying a treatment effect of 93.7 percentage points ($t(37) = 4.010$, $p < 0.001$).

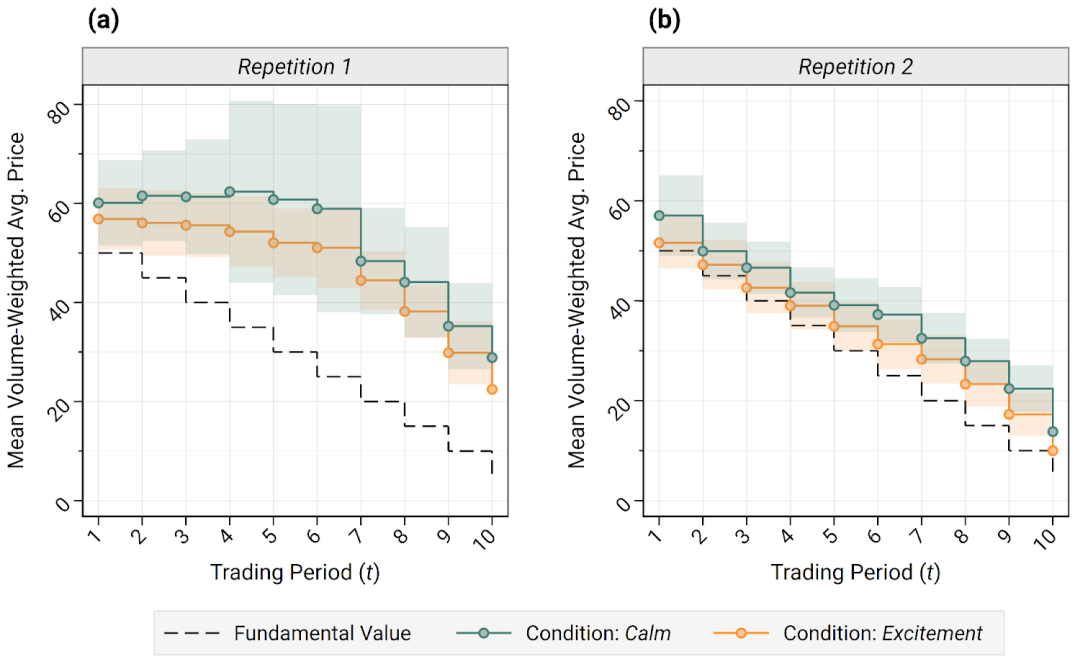


**Figure 1. Average prices by trading period, separated by treatment conditions, in the replication of Andrade, Odean, and Lin (2016).** The figure shows the period-by-period average of the volume-weighted average price (*VWAP*) across markets in the *Calm* condition ($n = 31$) and the *Excitement* condition ($n = 31$) for **(a)** the first repetition and **(b)** the second repetition of the experiment; the shaded areas depict the 95% confidence intervals around the mean. The dashed line indicates the asset's linearly declining fundamental value.

Figure 2 illustrates the results of the two replication hypothesis tests, evaluated using two-sided unpaired *t*-tests. Neither for *RD* ($t(60) = -0.905$, $p = 0.369$; $n = 62$) nor for $RD_{MAX}$ ($t(60) = -0.705$, $p = 0.483$; $n = 62$) do we find evidence for a difference between treatment conditions. Thus, we fail to reject the null for both hypotheses, and the replication rate, according to the statistical significance indicator, is 0% for AOL. The relative effect sizes are −13.3% and −13.9% for hypotheses 1 and 2, respectively, implying an average relative effect size of −13.6%.[30] The negative point estimates indicate that the replication estimates point in the opposite direction of the original effects. Detailed test results are reported in Table C1 in the Online Appendix.

[30] As indicated in Figure 1, the *Calm* condition comprises one outlying market (*RD* = 6.71, $RD_{MAX}$ = 11.94). Importantly, the results are not driven by the outlying observation. Excluding the outlier leaves the conclusions unaltered: *(a) RD*: $t(59) = -0.178$, $p = 0.859$, $n = 61$; *(b)* $RD_{MAX}$: $t(59) = 0.318$, $p = 0.752$, $n = 61$. The relative effect sizes excluding the outlying market are −2.6% and 4.7%, respectively. These robustness tests were not preregistered.

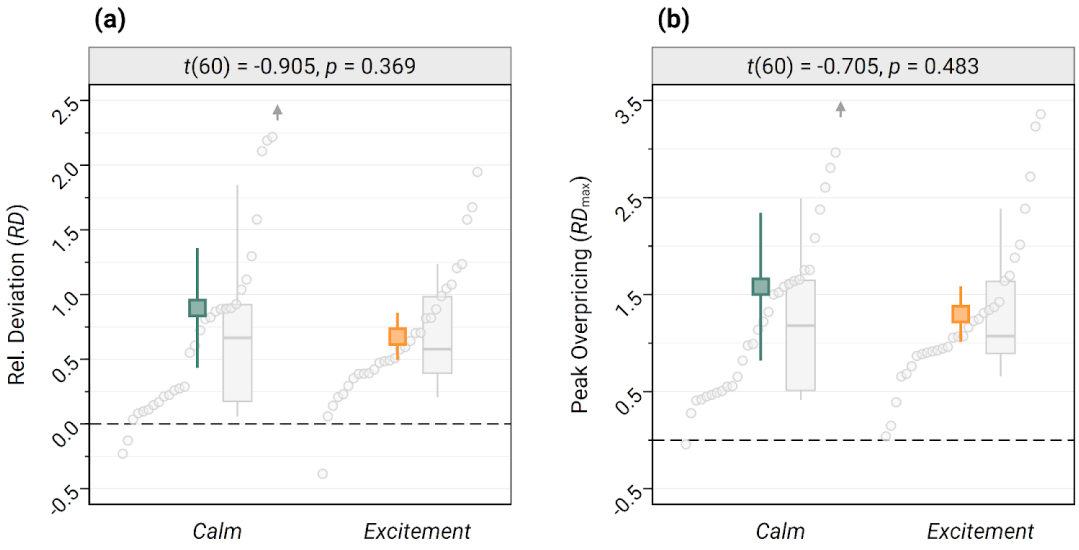


**Figure 2. Direct replication of Andrade, Odean, and Lin (2016).** The figure shows **(a)** the relative deviation (*RD*) and **(b)** peak overpricing ($RD_{MAX}$) for the *Calm* condition ($n = 31$) and the *Excitement* condition ($n = 31$), respectively, for the first repetition of the experiment. The square markers and the associated whiskers indicate the means and 95% confidence intervals of the mean, respectively; boxplots indicate the $p_{10}$, $p_{25}$, $p_{50}$, $p_{75}$, and $p_{90}$ percentiles. The results of two-sided two-sample *t*-tests (assuming equal variances), corresponding to replication hypothesis tests 1 and 2, are reported in the panel headers. One market is omitted from the figure in both panels (indicated by the arrow markers) as values exceed the y-axis scaling ($RD = 6.71$, $RD_{MAX} = 11.94$); the observations are not omitted in determining the means, confidence intervals, and test results, though. As preregistered robustness tests, we report Wilcoxon rank-sum tests for *RD* and $RD_{MAX}$ between treatment conditions: *(a)* $z = 0.049$, $p = 0.961$; and *(b)* $z = 0.092$, $p = 0.927$.

*Secondary Hypothesis Tests.*—In addition to the replication hypothesis tests, we conducted a series of preregistered secondary analyses, which are not replication tests. First, we test for a treatment effect on mispricing (*RAD*) in repetition 1, a measure not included in AOL but included in KLS and frequently employed in the experimental asset market literature. We find no evidence of a treatment effect on mispricing (unpaired *t*-test; $t(60) = -0.975$, $p = 0.333$; $n = 62$). Second, we test for treatment effects on relative deviation (*RD*), peak overpricing ($RD_{MAX}$), and mispricing (*RAD*) in repetition 2. Consistent with the results for repetition 1, we find no evidence of treatment effects for any of the three measures (see Table C2 in the Online Appendix for details).

*Existence of Overpricing.*—Finally, we test for the existence of overpricing in repetitions 1 and 2, separately for each of the two conditions, by testing if the relative deviation (*RD*) differs from zero. Based on the previous literature, we hypothesized overpricing in repetition 1 but had no directional hypothesis in repetition 2. In repetition 1, we find statistically significant evidence of overpricing in both conditions; in repetition 2, we find statistically significant evidence of overpricing in the *Calm* treatment and suggestive evidence in the *Excitement* treatment (see Table C3 in the Online Appendix for details). The magnitude of overpricing and its development over the trading periods in repetition 1 is consistent with the literature (see, e.g., Palan 2013 for a review); however, we do not observe the frequently encountered pattern of average prices exceeding fundamental values already in the first period.[31] We will revisit the extent of overpricing in repetition 2 (relative to repetition 1) in section F, zeroing in on experience's moderating effects on mispricing.

*Non-parametric Robustness Tests.*—All primary and secondary hypothesis tests are robust (regarding statistical significance) in terms of using distribution-free tests instead of parametric tests. The results of the non-parametric robustness tests are reported alongside the primary analyses in Tables C1–C3 in the Online Appendix.

## *B. Direct Replication of Kocher, Lucks, and Schindler (2019)*

*Manipulation Check.*—We mirror the manipulation checks in KLS and test (using unpaired *t*-tests; $n = 920$ in all tests) whether participants' Stroop task performance in the *Low Self-Control* (*LowSC*) and the *High Self-Control* (*HighSC*) conditions differs in terms of *(i)* the number of attempted problems in the Stroop task ($t(918) = -15.601$, $p < 0.001$), *(ii)* the number of correctly solved problems in the Stroop task ($t(918) = -15.833$, $p < 0.001$), *(iii)* the number of mistakes in the Stroop task ($t(918) = -4.351$, $p < 0.001$), and *(iv)* how demanding participants perceived the Stroop task to be ($t(918) = -7.511$, $p < 0.001$). We find statistically significant differences in the hypothesized direction for the four manipulation checks, lending support to the effectiveness of the treatment manipulation.

[31] Various studies in the previous literature find that prices start well below the fundamental value in the first few periods. Miller (2002) and Porter and Smith (2008) argue that initial prices falling below fundamentals may reflect traders' risk aversion, which alleviates as they become more familiar with the trading environment in the course of the first few trading periods. Notably, mean prices in AOL do not start below the fundamental value either.

*Replication Hypothesis Tests.*—The replication of KLS involves two replication hypotheses, conjecturing that the *LowSC* condition inflates *(1)* overpricing (*RD*) and *(2)* mispricing (*RAD*) as compared to the *HighSC* condition in the first repetition of the market experiment.[32] The average market prices per period and treatment for each of the two repetitions are illustrated in Figure 3. Similar to the data pertaining to the replication of AOL, the market prices in repetition 1 invoke the typical bubble pattern but without any apparent differences between treatment conditions.

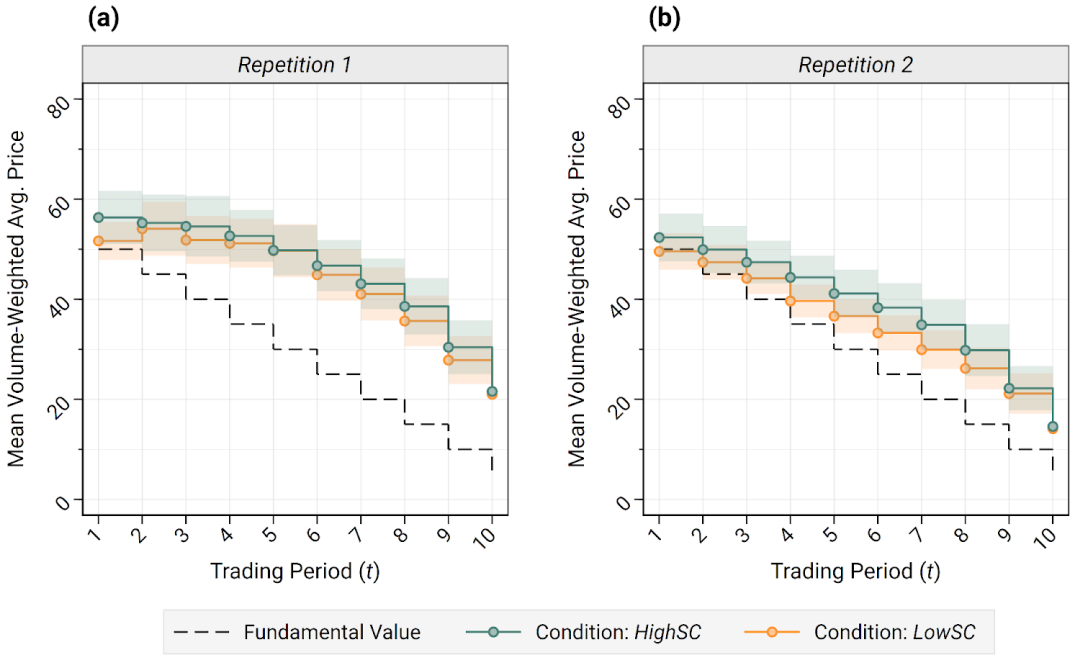


**Figure 3. Average prices by trading period, separated by treatment conditions, in the replication of Kocher, Lucks, and Schindler (2019).** The figure shows the period-by-period average of the volume-weighted average price (*VWAP*) across markets in the *High Self-Control* (*HighSC*) condition ($n = 52$) and the *Low Self-Control* (*LowSC*) condition ($n = 52$) for **(a)** the first repetition and **(b)** the second repetition of the experiment; the shaded areas depict the 95% confidence intervals around the mean. The dashed line indicates the asset's linearly declining fundamental value.

[32] Based on KLS's original data, we re-estimated the two focal hypothesis tests using unpaired *t*-tests yielding similar results as the original non-parametric tests for both *RD* ($t(14) = 2.065$, $p = 0.058$) and *RAD* ($t(14) = 2.386$, $p = 0.032$). In KLS, prices, on average, outvalue fundamentals by 18.8% in the *HighSC* and 49.9% in the *LowSC* condition, implying a treatment difference of 34.4 percentage points for *RD*; the relative absolute deviation (*RAD*) is 32.5% in *HighSC* and 58.9% in *LowSC*, entailing a treatment effect of 26.4 percentage points. KLS evaluated the two hypotheses using Mann-Whitney *U*-tests reported *p*-values of 0.074 for the treatment effect on *RD* and 0.046 for the treatment effect on *RAD*, respectively. Both results are qualified as "statistically significant" by the original authors based on an $\alpha = 10\%$ significance threshold.

Figure 4 illustrates the results of the two replication hypothesis tests, evaluated using two-sided unpaired *t*-tests. We cannot reject the null for either of the two conjectures: Neither *RD* ($t(102) = -0.659$, $p = 0.512$; $n = 104$) nor *RAD* ($t(102) = -0.755$, $p = 0.452$; $n = 104$) differs significantly between the *LowSC* and the *HighSC* condition in repetition 1. Hence, the replication rate, according to the statistical significance indicator, is 0% for the replication of KLS. The relative replication effect size is −12.5% for the treatment effect of *RD* and −12.4% for the effect of *RAD*, implying an average relative effect size of −12.5%. As for the replication of AOL, the point estimates of the replication effect sizes are in the opposite direction of those in the original study.

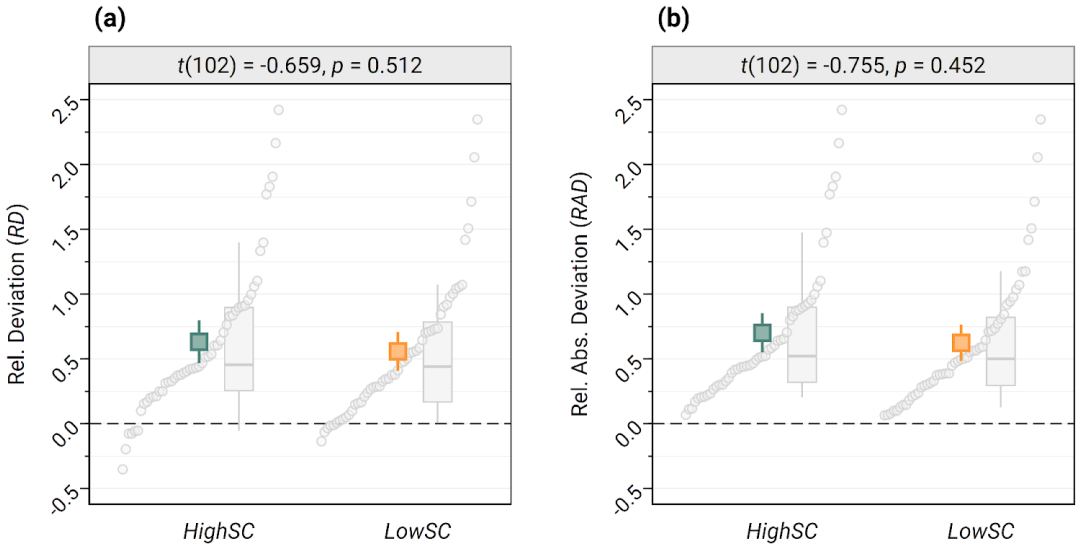


**Figure 4. Direct replication of Kocher, Lucks, and Schindler (2019).** The figure shows **(a)** the relative deviation (*RD*) and **(b)** the relative absolute deviation (*RAD*) for the *High Self-Control* (*HighSC*) condition ($n = 52$) and the *Low Self-Control* (*LowSC*) condition ($n = 52$), respectively, for the first repetition of the experiment. The square markers and the associated whiskers indicate the means and 95% confidence intervals of the mean, respectively; boxplots indicate the $p_{10}$, $p_{25}$, $p_{50}$, $p_{75}$, and $p_{90}$ percentiles. The results of two-sided two-sample *t*-tests (assuming equal variances), corresponding to replication hypothesis tests 1 and 2, are reported in the panel headers. As preregistered robustness tests, we report Wilcoxon rank-sum tests for *RD* and *RAD* between treatment conditions: *(a)* $z = -0.670$, $p = 0.503$; and *(b)* $z = -0.774$, $p = 0.439$.

*Secondary Hypothesis Tests.*—To complement the replication hypothesis tests, we perform various preregistered secondary tests that are not considered replication tests. Particularly, we test for a treatment effect on peak overpricing ($RD_{\text{MAX}}$) in repetition 1, a measure not included in KLS but included in AOL and commonly used to quantify overpricing in the context of asset market experiments. We find no evidence of a treatment effect on mispricing (unpaired $t$-test; $t(102) = -0.648$, $p = 0.519$, $n = 104$). In addition, we test for systematic differences between treatment conditions in terms of relative deviation ($RD$), peak overpricing ($RD_{\text{MAX}}$), and mispricing ($RAD$) in repetition 2. Consistent with the results for repetition 1, we find no evidence of treatment-induced effects in any of the three measures in repetition 2, and all estimates point in the opposite direction of KLS's claims (see Table D2 in the Online Appendix for details).

*Existence of Overpricing.*—To evaluate the existence of overpricing, mirroring the analysis in the replication of AOL, we test if the relative deviation ($RD$) differs from zero in each condition and repetition. Again, we hypothesized overpricing in repetition 1 but had no directional hypothesis about overpricing in repetition 2. We find statistically significant evidence of overpricing in both conditions and both repetitions (see Table D3 in the Online Appendix for details).

*Non-parametric Robustness Tests.*—To gauge the results' analytical robustness, we report the results of robustness tests, replacing the parametric tests with distribution-free alternatives for all tests reported above. The corresponding results are tabulated alongside the results of the primary analysis in Tables D1–D3. All primary and secondary tests turn out to be robust in terms of both statistical significance and sign.

### *C. Direct Replication of Eckel and Füllbrunn (2015)*

*Replication Hypothesis Tests.*—The replication of EF involves four replication hypothesis tests for the meta-analytic claims that the fraction of female traders in a market is associated with four different "bubble measures:" *(i)* average bias (*AB*), *(ii)* positive deviation (*PD*), *(iii)* boom duration, and *(iv)* bust duration.[33] Following EF, we test the four hypotheses using Spearman rank correlations.[34]

The associations between the fraction of female traders and the four outcome measures used in replication hypothesis 1–4 are illustrated in Figure 5. The Spearman correlations ($\rho_S$; $n = 166$ in all tests) are *(a)* $\rho_S = 0.139$ ($z = 1.774$, $p = 0.075$) for average bias, *(b)* $\rho_S = 0.153$ ($z = 1.962$, $p = 0.049$) for positive deviation, *(c)* $\rho_S = 0.093$ ($z = 1.186$, $p = 0.234$) for boom duration, and *(d)* $\rho_S = -0.023$ ($z = -0.299$, $p = 0.765$) for bust duration. A comprehensive summary of the four replication hypothesis tests is provided in Table E1 in the Online Appendix. Therefore, all four original claims fail to replicate according to the statistical significance indicator. All four tested associations even point in the opposite direction of the original results, with one replication hypothesis test yielding suggestive evidence of a correlation in the opposite direction of the original claim. Consequently, the fraction of successful replication tests is 0% according to the statistical significance indicator. The relative effect sizes in the four replication tests are −29.1%, −43.7%, −23.8%, and −4.4%, respectively, implying an average relative replication effect size of −25.2%.

[33] Our replication study includes EF's meta-analytic results of 35 SSW markets from the literature pooled in a meta-analysis and consolidated as Observation 4 in the original article. Of the four observations reported in EF, Observations 1–3 are based on data collected within their experiment, comparing all-female to all-male markets. As part of Observation 4, EF also conducts seven markets with both male and female participants that they do not include in the meta-analysis but compare to their all-male and all-female markets (this part of Observation 4 is not included in our replication study). In our replication, participants were not informed about the gender composition within markets, and could not reliably infer the fraction of female traders. While Eckel and Füllbrunn (2017) argued that the prevalence of gender effects might be linked to the observability of the gender composition, it is unclear whether the gender composition per market was also ambiguous in the 35 markets included in EF's meta-analysis.

[34] Based on their sample comprising 35 markets, EF reported correlation coefficients of *(i)* $\rho_S = -0.477$ ($p = 0.013$) for average bias, *(ii)* $\rho_S = -0.351$ ($p = 0.057$) for positive deviation, *(iii)* $\rho_S = -0.390$ ($p = 0.037$) for boom duration, and *(iv)* $\rho_S = 0.529$ ($p < 0.001$) for bust duration. Hypothesis tests in EF are evaluated by the original authors based on an $\alpha = 10\%$ significance threshold.

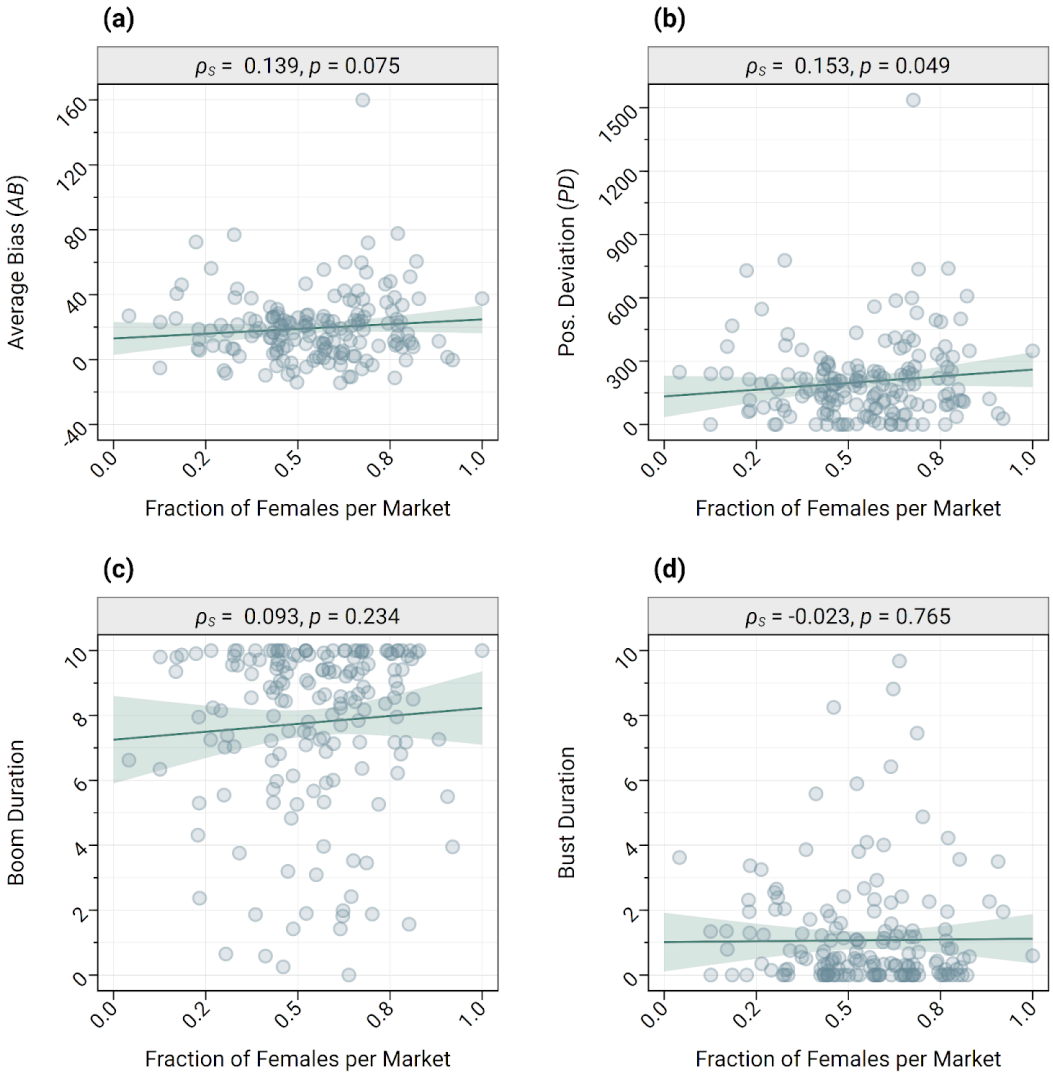


**Figure 5. Direct replication of Eckel and Füllbrunn (2015).** The figure plots the relationship between **(a)** average bias (*AB*), **(b)** positive deviation (*PD*), **(c)** boom duration, and **(d)** bust duration and the fraction of female participants per market in the first repetition of the experiment ($n$ = 166 in each panel); markers are randomly jittered to enhance exposition. Solid lines and shaded areas indicate linear trends and the associated 95% confidence intervals. Spearman correlation coefficients and the associated $p$-values, corresponding to replication hypothesis tests 1–4, are reported in the panel headers. As preregistered robustness tests, we report Pearson correlation coefficients ($\rho_P$): *(a)* $\rho_P$ = 0.102, $p$ = 0.191; *(b)* $\rho_P$ = 0.114, $p$ = 0.144; *(c)* $\rho_P$ = 0.064, $p$ = 0.413; and *(d)* $\rho_P$ = 0.010, $p$ = 0.896.

*Secondary Hypothesis Tests.*—To complement our analysis of the association between the fraction of female traders per market and the prevalence and extent of mispricing, we report the results of preregistered secondary tests (not considered direct replication tests). Particularly, we test for an association between the fraction of female traders and the bubble measures used in the replications of AOL and KLS: overpricing (*RD*), peak overpricing ($RD_{MAX}$), and mispricing (*RAD*). The Spearman correlations in these tests ($\rho_P$; $n = 166$ in all tests) are $\rho_P = 0.135$ ($z = 1.731$, $p = 0.082$) for overpricing, $\rho_P = 0.174$ ($z = 2.224$, $p = 0.025$) for peak overpricing, and $\rho_P = 0.186$ ($z = 2.381$, $p = 0.016$) for mispricing; see Table E2 in the Online Appendix for details. Hence, in two of the three tests, we find suggestive evidence of an association between the fraction of female traders and the mispricing measures in the opposite direction of EF's claim.

*Robustness Tests.*—As preregistered robustness tests, we examine the replication hypotheses and the secondary hypotheses based on the average across the two repetitions for each of the outcome measures (*AB*, *PD*, *Boom Duration*, *Bust Duration*, *RD*, $RD_{MAX}$, and *RAD*). These robustness tests, reported in Table E3 in the Online Appendix, are consistent with the above results, with the exception that there is no longer suggestive evidence of an association in the opposite direction for any of the four replication hypothesis tests and only for one of the secondary hypothesis tests. In addition, we preregistered using Pearson's product-moment correlation instead of Spearman's rank correlation as a robustness test for all correlations reported above. The corresponding results are tabulated alongside the main analyses in Tables E1–E3 in the Online Appendix. The conclusions regarding the replicability of EF's claims are robust to the parametric alternative, with all seven correlation coefficients still pointing in the opposite direction of the original hypotheses. However, there is no longer suggestive evidence of a correlation in the opposite direction of the original results for any of the hypotheses.

### *D. Conceptual Replication of Corgnet, Desantis, and Porter (2018)*

Unlike the replications of AOL and KLS, our conceptual replication attempt of CDP does not address all focal hypothesis tests in the original article since some of CDP's conjectures relate to aspects of Plott and Sunder's (1988) paradigm not applicable to the SSW setting used in our study. Notwithstanding, we replicate all tests related to the association of traders' performance and their cognitive reflection, fluid intelligence, and Theory of Mind that our study setting permits to address.[35] We attempted to replicate nine of CDP's hypothesis tests, six of which were reported as statistically significant in the original study (based on the $\alpha = 0.10$ significance threshold employed by CDP). Replication hypotheses 5, 7, and 8 below were reported as null results in the original study.

*Replication Hypothesis Tests.*—The conceptual replication of CDP involves nine replication hypotheses, which are tested using four ordinary least squares regressions of participants' final market earnings in repetition 1 of the experiment, with standard errors clustered on the market level ($n = 1{,}542$ in all regressions). Particularly, we estimate the following estimating equations:

$$\boldsymbol{\mu} = \boldsymbol{\alpha} + [CRT, APM, TOM]\boldsymbol{\beta} + \boldsymbol{\Omega}\boldsymbol{\omega} + \boldsymbol{\epsilon}, \tag{1}$$

$$\boldsymbol{\mu} = \boldsymbol{\alpha} + [CRT, APM, TOM]\boldsymbol{\beta} + [CRT \times TOM, APM \times TOM]\boldsymbol{\gamma} + \boldsymbol{\Omega}\boldsymbol{\omega} + \boldsymbol{\epsilon}, \tag{2}$$

$$\boldsymbol{\mu} = \boldsymbol{\alpha} + [SI]\boldsymbol{\beta}_1 + [HET]\boldsymbol{\beta}_2 + [SI \times HET]\boldsymbol{\eta} + \boldsymbol{\Omega}\boldsymbol{\omega} + \boldsymbol{\epsilon}, \tag{3}$$

$$\boldsymbol{\mu} = \boldsymbol{\alpha} + [CRT, APM, TOM]\boldsymbol{\beta}_1 + [HET]\boldsymbol{\beta}_2 + [CRT \times HET, APM \times HET, TOM \times HET]\boldsymbol{\zeta} + \boldsymbol{\Omega}\boldsymbol{\omega} + \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{\mu}$ indicates the traders' earnings vector; $\boldsymbol{\alpha}$ denotes a unity vector for the constant; $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$, $\boldsymbol{\zeta}$, and $\boldsymbol{\omega}$ are vectors of the coefficient estimates; and $\boldsymbol{\epsilon}$ indicates the residuals. $\boldsymbol{\Omega}$ denotes a matrix of control variables (covering indicator variables for women, participants that answer "prefer not to say" on the gender question, and treatment conditions).

Replication hypotheses 1–3, conjecturing a positive association between traders' earnings and cognitive reflection (*CRT*), fluid intelligence (*APM*), and Theory of Mind (*TOM*), are evaluated based on the coefficient estimates in $\boldsymbol{\beta}$ in estimating equation (1); hypotheses

[35] CDP encompasses five conjectures, each of which involves multiple hypotheses and tests. We conceptually replicate all of the hypothesis tests in three out of the five conjectures (conjectures 1, 2, and 4; the conjectures not specific to the Plott and Sunder (1988) design employed in CDP).

4 and 5 test for a positive moderation effect of Theory of Mind on cognitive reflection and fluid intelligence on earnings, captured by the interaction term estimates comprised by $\boldsymbol{\gamma}$ in equation (2); hypothesis 6, conjecturing that the effect of traders' skills (*SI*) on earnings is more pronounced when the traders' skills in a market are heterogeneous (*HET*),[36] is evaluated based on the interaction term estimate $\boldsymbol{\eta}$ in estimating equation (3); hypotheses 7–9 test for a positive moderation effect of heterogeneity in skills on cognitive reflection, fluid intelligence, and Theory of Mind, captured by the $\boldsymbol{\zeta}$ estimates in equation (4).

We summarize the results in Figure 6, illustrating the effect size estimates and their 95% confidence intervals for the nine replication hypothesis tests. Note that all independent variables are *z*-standardized so that the regression coefficients measure the change in the dependent variable for a one-standard-deviation change in the independent variable. Detailed test results are provided in Table F1 in the Online Appendix; the results for estimating equations (1) through (4) are tabulated in Tables F2–F5 in the Online Appendix.

For the six replication results reported as statistically significant in the original study, we find statistically significant evidence of an association in the same direction as the original study in two tests, suggestive evidence in one test, and no evidence of an association in three tests. This implies a replication rate of 50% according to the statistical significance indicator. Notably, the three original results that replicate successfully according to the statistical significance indicator are the three tested main effects; in contrast, the three tested interaction effects fail to replicate. The relative replication effect sizes in these six tests are 114.7%, 60.9%, 39.3%, −24.9%, 7.7%, and −4.0%, and the average relative effect size is 32.3%. For the three original results reported as statically insignificant, we find no evidence of an association in our replication tests either. According to the statistical significance indicator, the replication rate is thus 100% for these three tests.

[36] The *Skills Index* (SI) is defined as the average of a trader's *z*-standardized cognitive reflection, fluid intelligence, and Theory of Mind scores. To construct the dichotomous *Heterogeneity* variable (HET), we first estimate the interquartile range of SI for each market. HET takes the value 1 for markets at or above the median interquartile range in the skills index and 0 for markets below the median interquartile range. We refer to the Online Appendix for details.

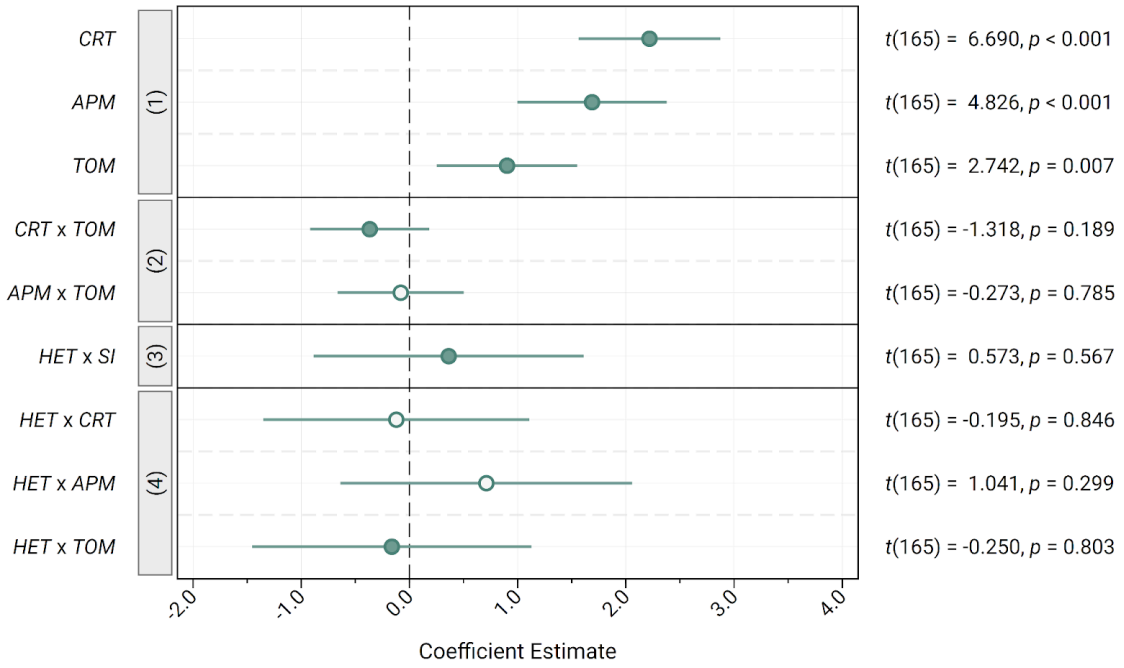


**Figure 6. Conceptual replication of Corgnet, Desantis, and Porter (2018).** The figure plots the coefficient estimates of *(i)* cognitive reflection, *CRT*; *(ii)* fluid intelligence, *APM*; *(iii)* Theory of Mind skills, *TOM*; *(iv)* the interaction of cognitive reflection and Theory of Mind, *CRT* × *TOM*; *(v)* the interaction of fluid intelligence and Theory of Mind, *APM* × *TOM*; *(vi)* the interaction of heterogeneity in skills (*HET*) and the skills index (*SI*), *HET* × *SI*; *(vii)* the interaction of heterogeneity in skills and cognitive reflection, *HET* × *CRT*; *(viii)* the interaction of heterogeneity in skills and fluid intelligence, *HET* × *APM*; and *(ix)* the interaction of heterogeneity in skills and Theory of Mind skills, *HET* × *TOM* on traders' earnings (in €; *m* = 25.00, *sd* = 10.04; min = 0.01, max = 81.79) in the first repetition of market trading. Estimates *(i)–(iv)* were obtained based on estimating equation (1), *(v)–(vi)* are based on equation (2), *(vi)* is based on equation (3), and *(vii)–(ix)* are based on equation (4); with all metric explanatory variables (i.e., *CRT*, *APM*, *TOM*, and *SI*) entering in *z*-standardized terms (i.e., estimates are in standard deviation units). The corresponding regression estimates (in unstandardized terms) are reported in Tables F2–F8 in the Online Appendix. Hollow markers indicate hypothesis tests for which the original study reports statistically insignificant effects. Whiskers indicate 95% confidence intervals based on cluster robust standard errors (*n* = 1,542 in 166 clusters for all models).

*Secondary Hypothesis Tests.*—Above and beyond the replication hypothesis tests detailed above, we complement our analysis with a set of preregistered secondary hypothesis tests. Other than the primary tests, the secondary analyses do not test whether individuals with higher cognitive reflection, fluid intelligence, and Theory of Mind outperform their peers in terms of earnings but whether markets populated with traders with higher mean cognitive reflection, fluid intelligence, and Theory of Mind perform better in terms of lower levels of

overpricing ($RD$), peak overpricing ($RD_{\text{MAX}}$), and mispricing ($RAD$). These analyses are inspired by, for instance, Hefti, Heinke, and Schneider (2016) and Bosch-Rosa, Meissner, and Bosch-Domenech (2018), showing that markets populated with traders with higher cognitive, analyzing, and mentalizing skills are less prone to mispricing. We also test if $RD$, $RD_{\text{MAX}}$, and $RAD$ are associated with the fraction of female traders per market, which can be thought of as robustness tests of the claims in EF controlling for the mean cognitive reflection, fluid intelligence, and Theory of Mind among traders in a market.[37] We test these hypotheses through the following estimating equations:

$$\boldsymbol{RD} = \boldsymbol{\alpha} + [\overline{CRT}, \overline{APM}, \overline{TOM}, \overline{Fem}]\,\boldsymbol{\beta} + \boldsymbol{\Omega}\,\boldsymbol{\omega} + \boldsymbol{\epsilon}, \tag{5}$$

$$\boldsymbol{RAD} = \boldsymbol{\alpha} + [\overline{CRT}, \overline{APM}, \overline{TOM}, \overline{Fem}]\,\boldsymbol{\beta} + \boldsymbol{\Omega}\,\boldsymbol{\omega} + \boldsymbol{\epsilon}, \tag{6}$$

$$\boldsymbol{RD_{\text{MAX}}} = \boldsymbol{\alpha} + [\overline{CRT}, \overline{APM}, \overline{TOM}, \overline{Fem}]\,\boldsymbol{\beta} + \boldsymbol{\Omega}\,\boldsymbol{\omega} + \boldsymbol{\epsilon}, \tag{7}$$

where $\boldsymbol{\alpha}$ denotes a unity vector for the constant; $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ are vectors of the coefficient estimates; and $\boldsymbol{\epsilon}$ indicates the residuals. $\overline{CRT}$, $\overline{APM}$, and $\overline{TOM}$ indicate the market-level means of cognitive reflection, fluid intelligence, and Theory of Mind scores, and $\overline{Fem}$ is the fraction of female traders per market; $\boldsymbol{\Omega}$ denotes a matrix of treatment indicators to account for condition fixed effects. The analysis is carried out at the market level ($n = 166$). We hypothesize negative signs for the four coefficient estimates comprised by $\boldsymbol{\beta}$.

The results are summarized in Table F6 in the Online Appendix; the regression analysis estimates pertaining to equations (5) through (7) are tabulated in Tables F6–F9. We find no evidence of an association between mean cognitive reflection, mean fluid intelligence, and mean Theory of Mind and either of the measures of mispricing. Hence, we do not find support for the claims put forth by Hefti, Heinke, and Schneider (2016) and Bosch-Rosa, Meissner, and Bosch-Domenech (2018). Likewise, we find no evidence of an association between the fraction of female traders and the three measures of overpricing. The suggestive evidence of an association between the fraction of female traders and overpricing in the opposite direction of the hypothesis in two of the secondary hypothesis tests of EF

[37] Testing for gender differences in cognitive reflection, we find statistically significant evidence of higher scores for men than women ($t(1{,}521) = 11.852$, $p < 0.001$; $n = 1{,}523$) (in line with the results of the meta-study by Brañas-Garza, Kujal, and Lenkei 2019). For fluid intelligence, we fail to reject the null of no difference between genders ($t(1{,}521) = 1.482$, $p = 0.139$; $n = 1{,}523$); for Theory of Mind scores, we find evidence for women outperforming men ($t(1{,}521) = 4.901$, $p < 0.001$; $n = 1{,}523$). These descriptive tests were not preregistered.

above is thus not robust to controlling for the markets' mean cognitive, analyzing, and mentalizing skills.

*Robustness Tests.*—As a preregistered robustness test, we estimate the nine replication hypothesis tests based on the average earnings across the two repetitions. In these robustness tests (reported alongside the main analyses in Tables F1–F5 in the Online Appendix), the evidence for replication hypothesis 3 strengthens from suggestive to statistically significant evidence; the remaining eight replication hypothesis tests are unaffected. Likewise, we report the results of robustness tests for the secondary hypothesis tests based on the average of the outcome measures (*RD*, $RD_{\text{MAX}}$, and *RAD*) across the two repetitions of the market. The results turn out to be robust, with the exception that the robustness tests indicate suggestive evidence of a negative association between the market-level mean cognitive reflection and mispricing (*RAD*).

### *E. Summary of Replication Hypothesis Tests*

Figure 7 summarizes the 17 replication hypothesis tests by plotting the estimates of the replication effect sizes alongside the original effect sizes in terms of the standardized effect size measures used to determine the relative effect sizes.

Our replications show a statistically significant effect ($p < 0.05$) in the same direction as the original claims for three (21.4%) of the 14 original results reported as statistically significant; the average effect size estimate in the replications is 2.9% of the original estimates. For the three original results reported as statistically insignificant, we do not find evidence of a significant association in our replication either. Weighing the replication rate for each of the four original studies equally, the average replication rate according to the statistical significance indicator is 12.5%, with an estimated average relative replication effect size of −4.8%.[38] For the 14 results deemed statistically significant in the original studies, the average replication rate of the three negative results is 100%.

[38] Note that especially for the replication hypothesis tests in AOL, KLS and EF, the replication results are likely to be correlated for the tests within each study (as the only difference between the replication tests within each study is using different outcome measures for mispricing). This provides an argument for also reporting the overall replication rate after weighting each of the four original studies equally.

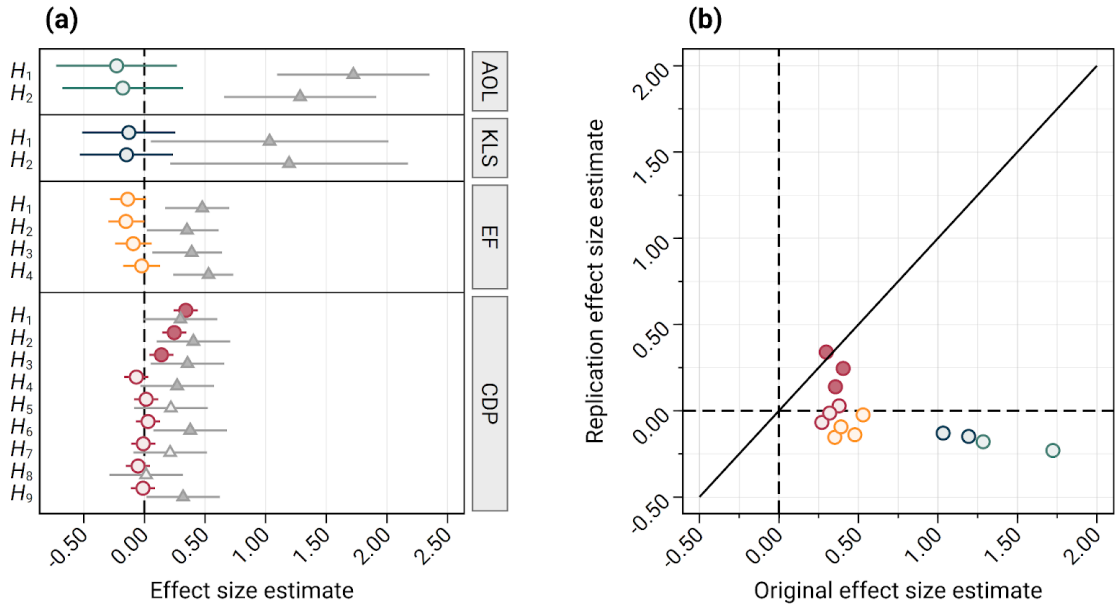


**Figure 7. Summary of the replication hypothesis tests. (a)** The figure plots the effect size estimates of the replication hypothesis tests (round markers) and the original estimates (triangular markers) with their 95% confidence intervals for the replications of Andrade, Odeon, and Lin (2016; AOL), Kocher, Lucks, and Schindler (2019; KLS), Eckel and Füllbrunn (2015; EF), and Corgnet, Desantis, and Porter (2018; CDP), respectively. Effect size estimates in AOL, KLS, and CDP are converted to Cohen's *d* units; estimates pertaining to EF are in correlation coefficient terms. Original estimates are assigned a positive sign; the replication estimates are normalized in signs so that positive (negative) values indicate effect size estimates in the same (opposite) direction as in the original study. For the original estimates, solid markers indicate results reported as statistically significant (with three results with $p < 0.10$ reported as significant); for the replications, solid markers indicate effect size estimates in the same direction as in the original study and $p < 0.05$ (the statistical significance indicator of replication). Three of the 14 original results reported as statistically significant (21.4%) successfully replicated according to the statistical significance indicator, and if each paper is weighted equally, the replication rate is 12.5%. None of the three original results reported as statistically insignificant had a *p*-value < 0.05 in the replications. **(b)** The figure illustrates the association between the effect size estimates in the replication and the original studies for the 14 results reported as statistically significant in the original studies. The solid line indicates equivalence (45°). The color coding in panel (b) is equivalent to the coding in panel (a). The relative effect size of the 14 hypotheses reported as statistically significant in the original studies varies between −43.7% and 114.7%, with a mean of 2.9%, and if each paper is weighted equally, the average relative effect size is −4.8%.

### *F. Conceptual Replication of the Experience Effect*

Apart from replicating the primary claims in AOL, KLS, EF, and CDP, our replication protocol has been designed to facilitate revisiting whether trading experience curbs the incidence and magnitude of bubbles—an effect that has been documented in several previous studies (e.g., Smith, Suchanek, and Williams 1988; Van Boening, Williams, and LaMaster 1993; Dufwenberg, Lindqvist, and Moore 2005; Haruvy, Lahav, and Noussair 2007; Hussam, Porter, and Smith 2008; Sutter, Huber, and Kirchler 2012). Thus, in contrast to the replication tests reported above, our tests of the experience effect are not directly tied to one specific study. We refer to these tests as conceptual replications with data from a similar population following the typology put forth by Dreber and Johannesson (2024).

*Replication Tests of the Experience Effect.*—Similar to Hussam, Porter, and Smith (2008) and Sutter, Huber, and Kirchler (2012), we examine whether trading experience mitigates the extent of mispricing by testing whether *RD*, $RD_{\text{MAX}}$, and *RAD* differ significantly between the first and second repetition of market trading. Figure 8 plots the empirical cumulative distribution functions of the three bubble measures. To test for an experience effect, we conducted three paired *t*-tests with $n = 166$ observations each; detailed test results are provided in Table G1 in the Online Appendix. We find statistically significant evidence in support of the experience hypothesis for all three bubble measures (relative deviation (*RD*): $t(165) = 8.236$, $p < 0.001$; peak overpricing ($RD_{\text{MAX}}$): $t(165) = 8.566$, $p < 0.001$; relative absolute deviation (*RAD*): $t(165) = 7.134$, $p < 0.001$), implying a replication rate of 100% for the experience hypothesis according to the statistical significance indicator.

*Tests for the Existence of Overpricing.*—As a complement to the tests for the existence of overpricing per condition reported in relation to the direct replications of AOL and KLS, we test for the presence of overpricing in repetitions 1 and 2, pooling the data across treatment conditions. One-sample *t*-tests for $RD = 0$ provide evidence for overpricing in both repetition 1 ($t(165) = 11.620$, $p < 0.001$; $n = 166$) and repetition 2 ($t(165) = 8.316$, $p < 0.001$; $n = 166$); see Table G2 in the Online Appendix. Hence, notwithstanding the significant decrease in overpricing from the first to the second repetition, experience does not eliminate overpricing in repetition 2.

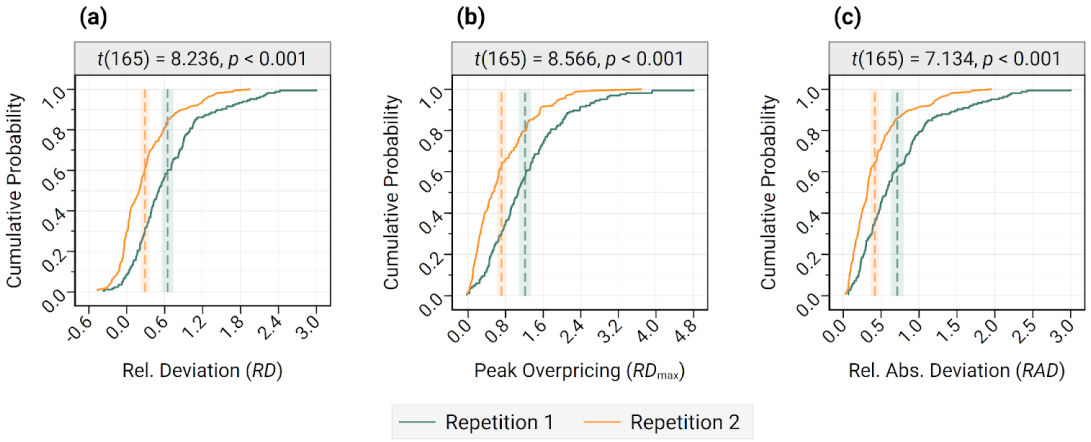


**Figure 8. Conceptual replication of the experience effect.** The figure plots the empirical cumulative distribution functions, separated for repetitions 1 and 2, for **(a)** relative deviation (*RD*), **(b)** peak overpricing ($RD_{MAX}$), and **(c)** relative absolute deviation (*RAD*). Vertical dashed lines and shaded areas indicate the means and corresponding 95% confidence intervals. The results of two-sided paired-sample *t*-tests, corresponding to experience hypothesis tests 1–3, are reported in the panel headers. As preregistered robustness tests, we report Wilcoxon matched-pairs signed-rank tests for *RD*, $RD_{MAX}$, and *RAD* between repetitions: *(a)* $z = 9.314$, $p < 0.001$; *(b)* $z = 9.591$, $p < 0.001$; and *(c)* $z = 8.867$, $p < 0.001$.

*Robustness Tests.*—As preregistered robustness analyses, we re-examine the hypothesis tests regarding the experience effect and the existence of overpricing using distribution-free tests. All tests reported above turn out to be robust in terms of both statistical significance and sign; see Tables G1–G2 in the Online Appendix for details.

## III. Conclusion

Replications are crucial for assessing the credibility of published findings and for updating beliefs regarding the strength of support for tested hypotheses and the magnitude of effects. We attempted to replicate 17 key findings from four experimental asset markets studies, relying on sample sizes that were, on average, about seven times larger than those in the original tests. All our replication tests were confirmatory, strictly following our preregistered protocol and adhering to our comprehensive pre-analysis plan.

We failed to replicate the focal claims of AOL and KLS that mispricing in asset markets is driven by emotions or low self-control, respectively, with replication effect sizes indistinguishable from zero but pointing in the opposite direction of the original studies.

Likewise, we failed to replicate the negative association between the fraction of female traders in a market and four bubble measures reported by EF, with statistically insignificant replication estimates pointing in the opposite direction of the original claims. With respect to the conceptual replication of CDP, we found support for the claims that traders' earnings are positively associated with cognitive reflection, fluid intelligence, and Theory of Mind. However, we did not find evidence for any of the six interaction effects examined in CDP, three of which were qualified as statistically significant in the original article. Finally, our replications corroborated the stylized fact that experience curbs the extent of bubbles documented repeatedly in the literature. In support of the experience hypothesis, we provided strong evidence of substantially reduced over- and mispricing in the second repetition of market trading. Despite a pronounced experience effect, however, market inefficiencies were not dashed out completely, with markets in a second repetition still exhibiting statistically significant over- and mispricing.

Pooling the results of the 17 replication hypothesis tests reveals a rather bleak picture regarding the credibility of positive results reported in the experimental asset market literature. For the 14 claims reported as statistically significant in the original articles, our estimated replication rate, according to the statistical significance indicator, stands at 21.4%, with an average relative replication effect size of only 2.9%. As the claims in each of the original studies—particularly in AOL, KLS, and EF—are likely to be highly correlated, it may be more appropriate to weigh each paper equally in aggregating the replication results across the four studies. Doing so results in an even lower replication rate (12.5%) and an even lower relative effect size (–4.8%).

Our results should be interpreted in light of some limitations and caveats. While our parameterization of the SSW market setting mirrors the experimental design of KLS, it differs somewhat from the implementation in AOL. The SSW parameterization in the 35 studies included in the meta-analysis in EF also slightly differs from the parameterization in our replication. We therefore consider it a borderline case whether to classify the replications of AOL and EF as direct or conceptual. However, we are not aware of any empirical evidence pointing at systematic effects of the market parameters we altered in our replication, such as the period length or the number of traders per market, that could account for the replication failures. While we cannot rule out the possibility that the

replicability of AOL's and EF's claims hinges on the specific market parameterization used, at a minimum, our replication results suggest that the respective claims do not generalize to arguably pertinent market settings. Relying on the SSW paradigm rather than the environment put forth by Plott and Sunder (1988) in the replication of CDP, on the other hand, implies more substantive and conceptual differences in the study design. While the failures to replicate the interaction effects in CDP could possibly stem from differences in the employed market paradigms, the credibility of the replication results, however, appears to be strengthened by the fact that we successfully replicated CDP's main effects. Finally, since our selection of claims put to a replication test is not necessarily representative of the experimental asset market literature, caution should be exercised in generalizing our findings to the entire field.

The credibility of empirical claims hinges on their replicability using new data. Various challenges—including publication bias, inadequate statistical power, and questionable research practices (see, e.g., Bishop 2019)—contribute to low replicability, ultimately eroding trust in empirical research and hindering the progress of scientific knowledge accumulation. In light of costly data collection, the literature on experimental asset markets appears to be prone to these challenges. Given the large variability in market-level outcomes commonly observed in market experiments, a major concern is that studies, on average, tend to be underpowered, increasing the risk of reporting false positives and exaggerated effect size estimates (see, e.g., Ioannidis 2005; 2008; Andrew Gelman and Carlin 2014). Moreover, the share of negative findings in the published literature on experimental asset markets appears to be too small in view of the small-sample settings most empirical claims are based on (see, e.g., Powell and Shestakova 2016 for a review), raising concerns about publication bias (Maniadis, Tufano, and List 2014; Benjamin et al. 2018).

To facilitate an effective accumulation of knowledge, claims in the experimental asset market literature should be revisited in systematic replication attempts to substantiate likely true effects and sort out likely false results. For yet-to-be-established empirical findings, experimental asset market studies should strive toward substantially larger samples to bolster statistical power, proper randomization to ensure unbiased inference, and confirmatory research practices through preregistration and pre-analysis plans to enhance the credibility of empirical findings.

# References

Altman, D. G., and J. M. Bland. 1995. "Statistics Notes: Absence of Evidence Is Not Evidence of Absence." *BMJ* 311 (7003): 485. https://doi.org/10/cdvzz9.

Andrade, E. B., T. Odean, and S. Lin. 2016. "Bubbling with Excitement: An Experiment." *Review of Finance* 20 (2): 447–66. https://doi.org/10/gs6tb5.

Azrieli, Y., C. P. Chambers, and P. J. Healy. 2018. "Incentives in Experiments: A Theoretical Analysis." *Journal of Political Economy* 126 (4): 1472–1503. https://doi.org/10/gd73q6.

Baron-Cohen, S., T. Jolliffe, C. Mortimore, and M. Robertson. 1997. "Another Advanced Test of Theory of Mind: Evidence from Very High Functioning Adults with Autism or Asperger Syndrome." *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 38 (7): 813–22. https://doi.org/10/c2p9xn.

Baron-Cohen, S., S. Wheelwright, J. Hill, Y. Raste, and I. Plumb. 2001. "The 'Reading the Mind in the Eyes' Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-Functioning Autism." *Journal of Child Psychology and Psychiatry* 42 (2): 241–51. https://doi.org/10/bng8xf.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, et al. 2018. "Redefine Statistical Significance." *Nature Human Behaviour* 2 (1): 6–10. https://doi.org/10/cff2.

Bishop, D. 2019. "Rein in the Four Horsemen of Irreproducibility." *Nature* 568 (7753): 435–435. https://doi.org/10/gfztcz.

Bloomfield, R., and A. Anderson. 2010. "Experimental Finance." In *Behavioral Finance*, edited by H. K. Baker and J. R. Nofsinger, 113–30. Hoboken, NJ: John Wiley & Sons, Ltd.

Bosch-Rosa, C., T. Meissner, and A. Bosch-Domènech. 2018. "Cognitive Bubbles." *Experimental Economics* 21 (1): 132–53. https://doi.org/10/gc2kcv.

Brañas-Garza, P., P. Kujal, and B. Lenkei. 2019. "Cognitive Reflection Test: Whom, How, When." *Journal of Behavioral and Experimental Economics* 82:101455. https://doi.org/10/ggs3c6.

Breaban, A., and C. N. Noussair. 2018. "Emotional State and Market Behavior." *Review of*

*Finance* 22 (1): 279–309. https://doi.org/10/gdbctx.

Brodeur, A., N. Cook, and A. Heyes. 2020. "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–60. https://doi.org/10/ghg83w.

Brunnermeier, M. K., and M. Oehmke. 2013. "Bubbles, Financial Crises, and Systemic Risk." In *Handbook of the Economics of Finance*, edited by G. M. Constantinides, M. Harris, and R. M. Stulz, 2:1221–88. Amsterdam, NL: Elsevier.

Brunnermeier, M. K., and I. Schnabel. 2016. "Bubbles and Central Banks." In *Central Banks at a Crossroads: What Can We Learn from History?*, edited by M. D. Bordo, O. Eitrheim, M. Flandreau, and J. F. Qvigstad, 493–562. Cambridge, UK: Cambridge University Press.

Caginalp, G., D. Porter, and V. L. Smith. 1998. "Initial Cash/Asset Ratio and Asset Prices: An Experimental Study." *Proceedings of the National Academy of Sciences* 95 (2): 756–61. https://doi.org/10/fd65vw.

———. 2001. "Financial Bubbles: Excess Cash, Momentum, and Incomplete Information." *Journal of Psychology and Financial Markets* 2 (2): 80–99. https://doi.org/10/bsnvwb.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–36. https://doi.org/10/bdps.

Camerer, C. F., A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour* 2 (9): 637–44. https://doi.org/10/gd3v2n.

Chen, D. L., M. Schonger, and C. Wickens. 2016. "oTree—An Open-Source Platform for Laboratory, Online, and Field Experiments." *Journal of Behavioral and Experimental Finance* 9:88–97. https://doi.org/10/bj42.

Corgnet, B., C. Deck, M. DeSantis, K. Hampton, and E. O. Kimbrough. 2023. "When Do Security Markets Aggregate Dispersed Information?" *Management Science* 69 (6): 3697–3729. https://doi.org/10/gtm33c.

Corgnet, B., M. Desantis, and D. Porter. 2018. "What Makes a Good Trader? On the Role of Intuition and Reflection on Trader Performance." *The Journal of Finance* 73 (3): 1113–37. https://doi.org/10/ggtmrc.

Cubitt, R. P., C. Starmer, and R. Sugden. 1998. "On the Validity of the Random Lottery Incentive System." *Experimental Economics* 1 (2): 115–31. https://doi.org/10/cb82rf.

Cueva, C., and A. Rustichini. 2015. "Is Financial Instability Male-Driven? Gender and Cognitive Skills in Experimental Asset Markets." *Journal of Economic Behavior & Organization* 119:330–44. https://doi.org/10/ggwnfd.

Dang, J., P. Barker, A. Baumert, M. Bentvelzen, E. Berkman, N. Buchholz, J. Buczny, et al. 2021. "A Multilab Replication of the Ego Depletion Effect." *Social Psychological and Personality Science* 12 (1): 14–24. https://doi.org/10/ggtpft.

Davis, A. M., B. Flicker, K. Hyndman, E. Katok, S. Keppler, S. Leider, X. Long, and J. D. Tong. 2023. "A Replication Study of Operations Management Experiments in Management Science." *Management Science* 69 (9): 4977–91. https://doi.org/10/gtm3h6.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3): 522–50. https://doi.org/10/d3cbfz.

Dreber, A., and M. Johannesson. 2024. "A Framework for Evaluating Reproducibility and Replicability in Economics." *Economic Inquiry* online first. https://doi.org/10/gt3vmw.

Dreber, A., T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, and M. Johannesson. 2015. "Using Prediction Markets to Estimate the Reproducibility of Scientific Research." *Proceedings of the National Academy of Sciences* 112 (50): 15343–47. https://doi.org/10/f738kx.

Dufwenberg, M., T. Lindqvist, and E. Moore. 2005. "Bubbles and Experience: An Experiment." *American Economic Review* 95 (5): 1731–37. https://doi.org/10/bm83fr.

Eckel, C. C., and S. C. Füllbrunn. 2015. "Thar SHE Blows? Gender, Competition, and Bubbles in Experimental Asset Markets." *American Economic Review* 105 (2): 906–20. https://doi.org/10/ggwnd5.

———. 2017. “Hidden vs. Known Gender Effects in Experimental Asset Markets.” *Economics Letters* 156:7–9. https://doi.org/10/gbm5n2.

Fama, E. F. 1970. “Efficient Capital Markets: A Review of Theory and Empirical Work.” *The Journal of Finance* 25 (2): 383–417. https://doi.org/10/b3kfdr.

Farago, A., M. Holmén, F. Holzmeister, M. Kirchler, and M. Razen. 2022. “Cognitive Skills and Economic Preferences in the Fund Industry.” *Economic Journal* 132 (645): 1737–64. https://doi.org/10/gpt24h.

Frederick, S. 2005. “Cognitive Reflection and Decision Making.” *Journal of Economic Perspectives* 19 (4): 25–42. https://doi.org/10/b98rhb.

Friese, M., D. D. Loschelder, K. Gieseler, J. Frankenbach, and M. Inzlicht. 2019. “Is Ego Depletion Real? An Analysis of Arguments.” *Personality and Social Psychology Review* 23 (2): 107–31. https://doi.org/10/gfxx79.

Galbraith, J. K. 1994. *A Short History of Financial Euphoria*. New York, NY: Penguin Books.

Gelman, A., and E. Loken. 2014. “The Statistical Crisis in Science.” *American Scientist* 102 (6): 460. https://doi.org/10/gc3f2j.

Gelman, Andrew, and John Carlin. 2014. “Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors.” *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 9 (6): 641–51. https://doi.org/10/b2h3.

Guerron-Quintana, P. A., T. Hirano, and R. Jinnai. 2023. “Bubbles, Crashes, and Economic Growth: Theory and Evidence.” *American Economic Journal: Macroeconomics* 15 (2): 333–71. https://doi.org/10/gt4zmc.

Hagger, M. S., N. L. D. Chatzisarantis, H. Alberts, C. O. Anggono, C. Batailler, A. R. Birt, R. Brand, et al. 2016. “A Multilab Preregistered Replication of the Ego-Depletion Effect.” *Perspectives on Psychological Science* 11 (4): 546–73. https://doi.org/10/f3tc5n.

Haruvy, E., Y. Lahav, and C. N. Noussair. 2007. “Traders’ Expectations in Asset Markets: Experimental Evidence.” *American Economic Review* 97 (5): 1901–20. https://doi.org/10/ccsdng.

Haruvy, E., and C. N. Noussair. 2006. “The Effect of Short Selling on Bubbles and Crashes in

Experimental Spot Asset Markets." *The Journal of Finance* 61 (3): 1119–57. https://doi.org/10/fp783q.

Haruvy, E., C. N. Noussair, and O. Powell. 2014. "The Impact of Asset Repurchases and Issues in an Experimental Market." *Review of Finance* 18 (2): 681–713. https://doi.org/10/gjp89b.

Hefti, A., S. Heinke, and F. Schneider. 2016. "Mental Capabilities, Trading Styles, and Asset Market Bubbles: Theory and Experiment." Working Paper. https://doi.org/10/m383.

Holzmeister, F., M. Johannesson, C. F. Camerer, Y. Chen, T.-H. Ho, S. Hoogeveen, J. Huber, et al. 2024. "Examining the Replicability of Online Experiments Selected by a Decision Market." *Nature Human Behaviour*, 1–15. https://doi.org/10/g8rfkm.

Hori, T., and R. Im. 2023. "Asset Bubbles, Entrepreneurial Risks, and Economic Growth." *Journal of Economic Theory* 210:105663. https://doi.org/10/gt4zmd.

Huber, J., and M. Kirchler. 2012. "The Impact of Instructions and Procedure on Reducing Confusion and Bubbles in Experimental Asset Markets." *Experimental Economics* 15 (1): 89–105. https://doi.org/10/b5njr7.

Hussam, R. N., D. Porter, and V. L. Smith. 2008. "Thar She Blows: Can Bubbles Be Rekindled with Experienced Subjects?" *American Economic Review* 98 (3): 924–37. https://doi.org/10/cvp77k.

Ioannidis, J. P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124. https://doi.org/10/chhf6b.

———. 2008. "Why Most Discovered True Associations Are Inflated." *Epidemiology* 19 (5): 640. https://doi.org/10/cst2h8.

Jaeggi, S. M., B. Studer-Luethi, M. Buschkuehl, Y.-F. Su, J. Jonides, and W. J. Perrig. 2010. "The Relationship between N-Back Performance and Matrix Reasoning — Implications for Training and Transfer." *Intelligence* 38 (6): 625–35. https://doi.org/10/ddjp5n.

James, D., and R. M. Isaac. 2000. "Asset Markets: How They Are Affected by Tournament Incentives for Individuals." *American Economic Review* 90 (4): 995–1004. https://doi.org/10/fcz37x.

Kindleberger, C. P., and R. Z. Aliber. 2011. *Manias, Panics and Crashes: A History of Financial Crises*. 6th ed. Basingstoke, UK: Palgrave Macmillan.

Kirchler, M., C. Bonn, J. Huber, and M. Razen. 2015. "The 'Inflow-Effect'—Trader Inflow and Price Efficiency." *European Economic Review* 77:1–19. https://doi.org/10/f3n47b.

Kirchler, M., J. Huber, and T. Stöckl. 2012. "Thar She Bursts: Reducing Confusion Reduces Bubbles." *American Economic Review* 102 (2): 865–83. https://doi.org/10/gg584h.

Kocher, M. G, K. E Lucks, and D. Schindler. 2019. "Unleashing Animal Spirits: Self-Control and Overpricing in Experimental Asset Markets." *The Review of Financial Studies* 32 (6): 2149–78. https://doi.org/10/gjb95r.

Kocher, M. G., P. Martinsson, and D. Schindler. 2017. "Overpricing and Stake Size: On the Robustness of Results from Experimental Asset Markets." *Economics Letters* 154:101–4. https://doi.org/10/f96sf4.

Kopányi-Peuker, A., and M. Weber. 2021. "Experience Does Not Eliminate Bubbles: Experimental Evidence." *The Review of Financial Studies* 34 (9): 4450–85. https://doi.org/10/ghf2zm.

Lei, V., C. N. Noussair, and C. R. Plott. 2001. "Nonspeculative Bubbles in Experimental Asset Markets: Lack of Common Knowledge of Rationality vs. Actual Irrationality." *Econometrica* 69 (4): 831–59. https://doi.org/10/c32zvz.

Maniadis, Z., F. Tufano, and J. A. List. 2014. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects." *The American Economic Review* 104 (1): 277–90. https://doi.org/10/f5npvq.

———. 2017. "To Replicate or Not to Replicate? Exploring Reproducibility in Economics through the Lens of a Model and a Pilot Study." *Economic Journal* 127 (605): F209–35. https://doi.org/10/gtq9fc.

Marini, M. M. 2023. "Emotions and Financial Risk-Taking in the Lab: A Meta-Analysis." *Journal of Behavioral Decision Making* 36 (4): e2342. https://doi.org/10/gtb8gv.

Miao, J., and P. Wang. 2018. "Asset Bubbles and Credit Constraints." *American Economic Review* 108 (9): 2590–2628. https://doi.org/10/gfgwhn.

Miller, R. M. 2002. “Can Markets Learn to Avoid Bubbles?” *Journal of Psychology and Financial Markets* 3 (1): 44–52. https://doi.org/10/fkckwz.

Noussair, C. N., and S. Tucker. 2016. “Cash Inflows and Bubbles in Asset Markets with Constant Fundamental Values.” *Economic Inquiry* 54 (3): 1596–1606. https://doi.org/10/g8qw8r.

Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251): aac4716. https://doi.org/10/68c.

Palan, S. 2013. “A Review of Bubbles and Crashes in Experimental Asset Markets.” *Journal of Economic Surveys* 27 (3): 570–88. https://doi.org/10/ggtd4f.

Patil, P., R. D. Peng, and J. T. Leek. 2016. “What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science.” *Perspectives on Psychological Science* 11 (4): 539–44. https://doi.org/10/f825k4.

Pawel, S., R. Heyard, C. Micheloud, and L. Held. 2024. “Replication of ‘Null Results’ – Absence of Evidence or Evidence of Absence?” *eLife* 12. https://doi.org/10/gtwg6k.

Plott, C. R., and S. Sunder. 1988. “Rational Expectations and the Aggregation of Inverse Information in Laboratory Security Markets.” *Econometrica* 56 (5): 1085–1118. https://doi.org/10/c3vwh2.

Porter, D., and V. L. Smith. 2008. “Price Bubbles.” In *Handbook of Experimental Economics Results*, edited by C. R. Plott and V. L. Smith, 1:247–55. Amsterdam, NL: Elsevier. https://doi.org/10/dsg9n3.

Powell, O., and N. Shestakova. 2016. “Experimental Asset Markets: A Survey of Recent Developments.” *Journal of Behavioral and Experimental Finance* 12:14–22. https://doi.org/10/gf794m.

Raven, J. C. 1941. “Standardization of Progressive Matrices.” *British Journal of Medical Psychology* 19 (1): 137–50. https://doi.org/10/dv6vk8.

Raven, J., and J. C. Raven. 2008. *Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven’s Quest for Non-Arbitrary Metrics*. Unionville, NY: Royal Fireworks Press.

Raven, J., J. C. Raven, and J. H. Court. 1998. *Manual for Raven’s Progressive Matrices and*

*Vocabulary Scales*. San Antonio, TX: Pearson.

Razen, M., J. Huber, and M. Kirchler. 2017. "Cash Inflow and Trading Horizon in Asset Markets." *European Economic Review* 92:359–84. https://doi.org/10/f9wzhf.

Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. https://doi.org/10/bxbw3c.

Smith, A., T. Lohrenz, J. King, P. R. Montague, and C. F. Camerer. 2014. "Irrational Exuberance and Neural Crash Warning Signals during Endogenous Experimental Market Bubbles." *Proceedings of the National Academy of Sciences* 111 (29): 10503–8. https://doi.org/10/f6brbd.

Smith, V. L., G. L. Suchanek, and A. W. Williams. 1988. "Bubbles, Crashes, and Endogenous Expectations in Experimental Spot Asset Markets." *Econometrica* 56 (5): 1119–51. https://doi.org/10/drhxwq.

Stöckl, T., J. Huber, and M. Kirchler. 2010. "Bubble Measures in Experimental Asset Markets." *Experimental Economics* 13 (3): 284–98. https://doi.org/10/djfr59.

Stroop, J. R. 1935. "Studies of Interference in Serial Verbal Reactions." *Journal of Experimental Psychology* 18 (6): 643–62. https://doi.org/10/b77m95.

Sutter, M., J. Huber, and M. Kirchler. 2012. "Bubbles and Information: An Experiment." *Management Science* 58 (2): 384–93. https://doi.org/10/bkdjpx.

Szucs, D., and J. P. A. Ioannidis. 2017. "Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature." *PLoS Biology* 15 (3): e2000797. https://doi.org/10/b4r4.

Thomson, K. S., and D. M. Oppenheimer. 2016. "Investigating an Alternate Form of the Cognitive Reflection Test." *Judgment and Decision Making* 11 (1): 99–113. https://doi.org/10/grzxgq.

Toplak, M. E., R. F. West, and K. E. Stanovich. 2014. "Assessing Miserly Information Processing: An Expansion of the Cognitive Reflection Test." *Thinking & Reasoning* 20 (2): 147–68. https://doi.org/10/gd3vqj.

Van Boening, M. V., A. W. Williams, and S. LaMaster. 1993. “Price Bubbles and Crashes in Experimental Call Markets.” *Economics Letters* 41 (2): 179–85. https://doi.org/10/bkqzmg.

Wake, S., J. Wormwood, and A. B. Satpute. 2020. “The Influence of Fear on Risk Taking: A Meta-Analysis.” *Cognition and Emotion* 34 (6): 1143–59. https://doi.org/10/gjqm3b.

Weitzel, U., C. Huber, J. Huber, M. Kirchler, F. Lindner, and J. Rose. 2020. “Bubbles and Financial Professionals.” *The Review of Financial Studies* 33 (6): 2659–96. https://doi.org/10/gh7pmd.

Zhang, L., and A. Ortmann. 2013. “Exploring the Meaning of Significance in Experimental Economics.” Working Paper. https://doi.org/10/m39f.