Replications and Crowd Research

實驗經濟學最新發展: 重現性與群衆研究

Joseph Tao-yi Wang (王道一)
National Taiwan University

Are Canonical Experimental Results Robust?

- People Assumed this was true, until:
- Bem (JPSP 2011): 9 experiments / total N>1000
  - "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect"
  - ESP is true?! Or, "standard" statistical practices flawed?
- Did X result replicate in follow-ups? Google-scholar it!
  - Shih, Pittinsky and Ambady (Psych Science 1999)
    - Asian-American women math scores increase* if activate ethnic identity, but decrease* if activate gender identity

Are Canonical Experimental Results Robust?

- Open Science Collaboration (Science 2015)
  - "Estimating the Reproducibility of Psychological Science"
  - Replicate all 100 Experiments published 2008 in top-3 experimental psychology journals:
    - Psychological Science/JPSP/J Experimental Psychology
  - Only 36 out of 100 replicate
  - Mean Effect Size = Slightly less than Half of Original
- Fail to Replicate due to bad incentives:
  - Publication Bias, Novelty Seeking, Budget Constraint

Is Economics Immune/Better?

- Camerer et al. (Science 2016)
  - "Evaluating Replicability of Laboratory Experiments in Economics"
  - Replicate all 18 Lab Experiments published 2011-14 in top-4 economics journals (AER/QJE/JPE/ECMA)
  - 7 out of 18 Fail, including one on Robin Williams effect
  - Mean Effect Size = 66% of Original
    - Replication: 61% vs. 36%; Effect Size 66% vs. <50%
- Better Replication due to better incentives?
  - Transparency? Good Econometrics? Less Novelty Seeking?

Is Economics Immune/Better?

- Brodeur et al. (AEJ-applied 2016)
  - "Star Wars: The Empirics Strike Back"
  - Z-stat Humps right before 1.96 for AER/QJE/JPE
  - p-values missing from 0.25-0.10, retrieved just after 0.05
  - Humps of lab experiments (slightly) lower/better
- "All have sinned and fall short of the glory of God..."
  - Bad Incentives are Everywhere!
- Need Incentives for Replication
  - Encourage more empirical tests on result robustness

Possible Solutions to the Replication Crisis

1. Open Science (Voluntary!)
   - More transparency in data collection and analysis
   - Shared experimental software/instructions allow replication
   - Shared data and code allow reproduction/robustness checks
     - Reinhart and Rogoff (AER 2010): Average real economic growth drops by 4% once country debt >90% of GDP
       - Fiscal Cliff at 90%, but due to an excel error (+ some others)!
       - Herndon, Ash and Pollin (2014), "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff," Cambridge Journal of Economics, 38 (2): 257–279.

Possible Solutions to the Replication Crisis

2. Data Policy (Mandatory!)
   - Journals require data and code packages prior to acceptance
     - Since 2005, AEA journals (AER/JEL/JEP/AEJs) accept papers conditional on authors providing replication packages
   - Fisar et al. (MS 2024) find new data policy of Management Science (since 2019) significantly improved reproducibility
3. Data Editor (Dedicated Enforcer!)
   - Journals hire data editors to enforce their data policy
     - AEA appointed its first data editor Lars Vilhuber in 2018

Possible Solutions to the Replication Crisis

4. Pre-Registration (just PR) without Pre-Analysis Plan
   - Register your study to avoid p-hacking/novelty-seeking
     - AEA RCT Registry: Required for field experiments in 2018
5. Pre-Registration with Pre-Analysis Plan (PAP)
   - Register your PAP (and even code!) before data collection
   - Brodeur, Cook, Hartley and Heyes (JPE-micro, 2024): 44% PR for RCTs published in 15 economics journals lack PAP
     - Just PR has no effect on p-hacking/publication bias
     - But, PR with PAP reduces p-hacking and publication bias

Possible Solutions to the Replication Crisis

6. Registered Report
   - Referees review PAP before seeing the results
   - Journals conditionally accept (until data collected/reported)
     - Journal of Development Economics (JDE)
     - European Economic Review (EER)
7. Top Journals Publish Replication Studies
   - Open Science Collaboration (Science 2015)
   - Camerer et al. (Science 2016), Lin et al. (NHB 2020)
   - Fisar et al. (MS 2024), etc.

Possible Solutions to the Replication Crisis

8. Replication Reports
   - In Special Issues, Replication Sections, or entire Journals
   - Journal of Applied Econometrics (JAE) Replication Section
     - Publish RR of its own since 2003, but now for 9 top economics journals (Top-5 + RESTAT, JoE, EJ, JBES)
   - European Economic Review (EER): Replication Section
   - Journal of the Economic Science Association (JESA)
- Can we require journals to publish RR of its articles?
  - Where a paper is published, so will its replication report!

Possible Solutions to the Replication Crisis

9. More Studies in non-WEIRD societies
   - Avoid focusing too much on WEIRD societies
     - Western, Educated, Industrialized, Rich, Democratic
   - Worldwide attempts of replications

These Solutions Still Face Challenges

- Novelty-seeking remains
  - Perform Replications = Lack of Novelty/Creativity
    - Can each master thesis archive a replication report?
    - Economists replicate past experiments as control group
- Pre-Registration without Pre-Analysis Plan (just PR!)
  - Cannot prevent p-hacking/publication bias unless with PAP
  - (Even if done right) creates barrier of entry against newcomer
    - Field experiments have to pre-register on AEA RCT Registry
  - Do reviewers check, or registry editor has to check all entries?

These Solutions Still Face Challenges

- Pre-Registration (PR) with Pre-Analysis Plan (PAP)
  - Reduces p-hacking and publication bias in published RCTs
  - But, cannot prevent p-hacking unless reviewers check them
    - Can PAP self-enforce, or needs registry editor to (like data editors enforcing data policy)?
  - (Even if done right) creates barrier of entry against newbies
    - But, voluntary PAPs are very helpful for experimental design!
- How to review registered PAP/Registered Reports?
- Non-WEIRD studies are under-funded...%

A New Hope: Crowd Research

- Though few have funding to conduct large scale studies,
- Hundreds of researchers can jointly
  - Conduct large scale replications
  - Computationally reproduce thousands of papers
    - Open Science Collaboration (2015) done by Many Labs
    - Fisar et al. (2024) invited 700+ authors to form the Management Science Reproducibility Collaboration
    - Camerer et al. (2016) and follow-up studies employ (expert) prediction markets to predict replicability

A New Hope: Crowd Research

- Hopefully, the Replication Crisis can be solved soon...
- By the Wisdom of the Crowd!
  - Crowd Research
  - Data Policy with Data Editors enforcing
  - Pre-Registration with Pre-Analysis Plan
  - Registered Report

- And Better Incentives for replications!

Evidence of General Economic Principles of Bargaining and Trade from 2,000 Classroom Experiments

Po-Hsuan Lin, Alexander L. Brown,
Taisuke Imai, Joseph Tao-yi Wang,
Stephanie W. Wang and Colin F. Camerer

Are Canonical Experimental Results Robust?

- Test Many Places with Standard Design
- Budget/Time Constraints Limit Scope
  - Oosterbeek et al. (2004): (Lack Standard Design)
    Meta-analysis of ultimatum game
  - Roth et al. (1991): (Fewer Places)
    Ultimatum game and market game in 4 countries
- MobLab: Ultimatum Game/Double Auction
  - Data From 10 Regions/Countries
  - Online Classroom Experiment Platform

MobLab Classroom Experimental Data

Cons:

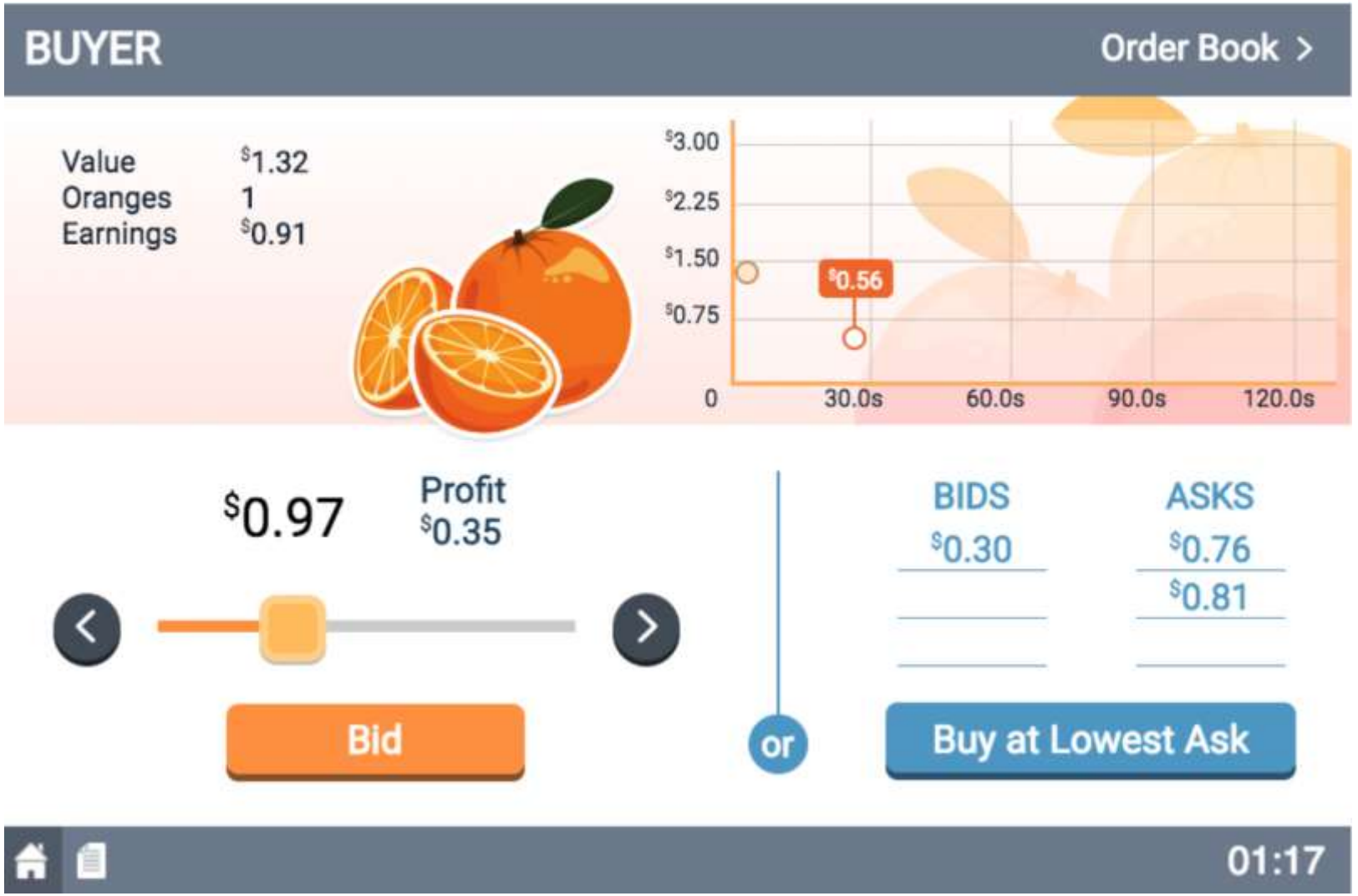
1. Most are not incentivized by money
   - Other incentives: Class Grade, Internal Motivation
2. Students in courses NOT representative
   - But exactly as Chamberlin (1948) and Smith (1962)

- Pros:

1. No Publication Bias
2. Demographic Variation: 10 regions/countries
3. Exact Same Interface/Language/Design

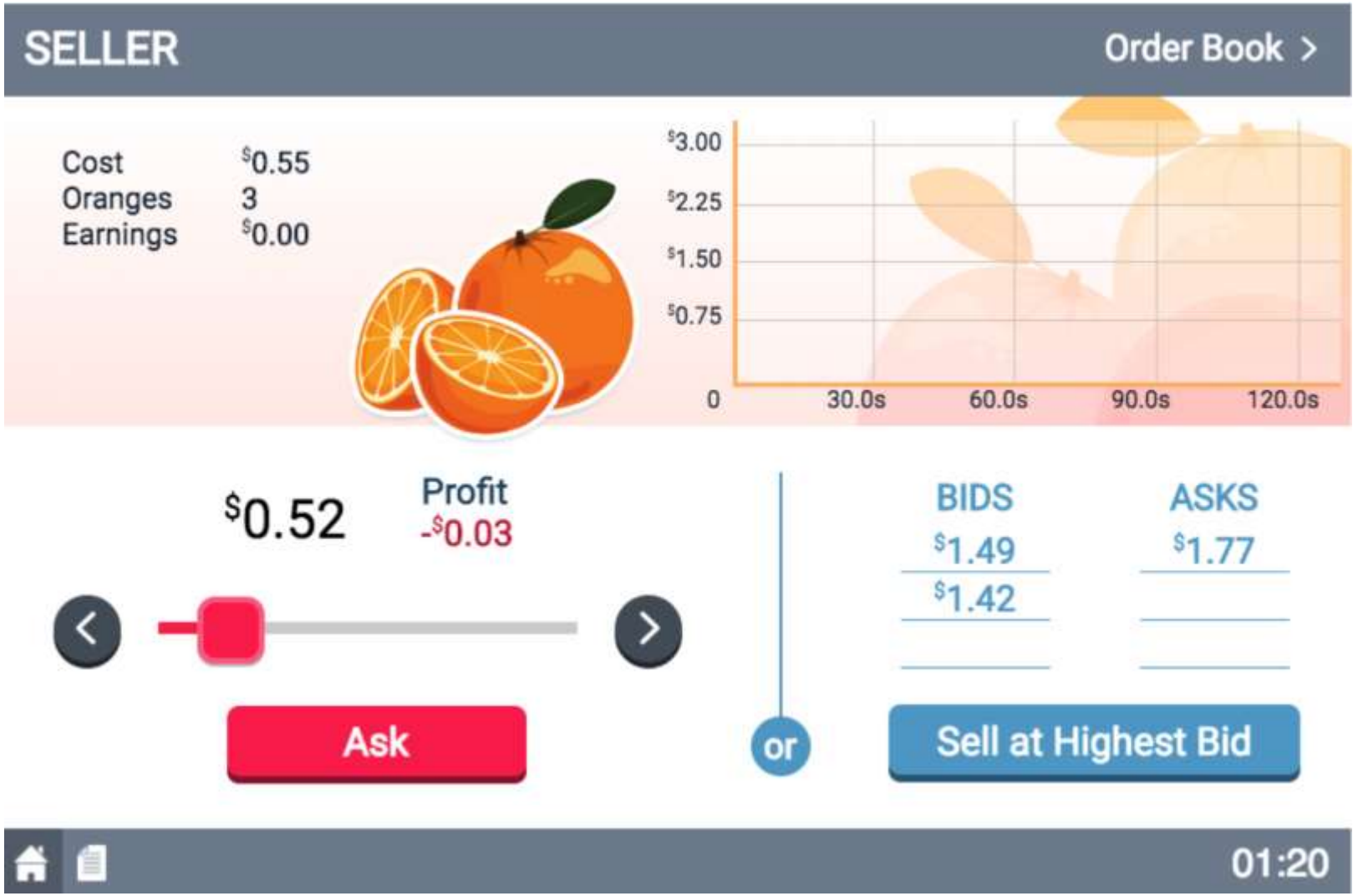
Double Auction:

Buyers



Double Auction:

Sellers



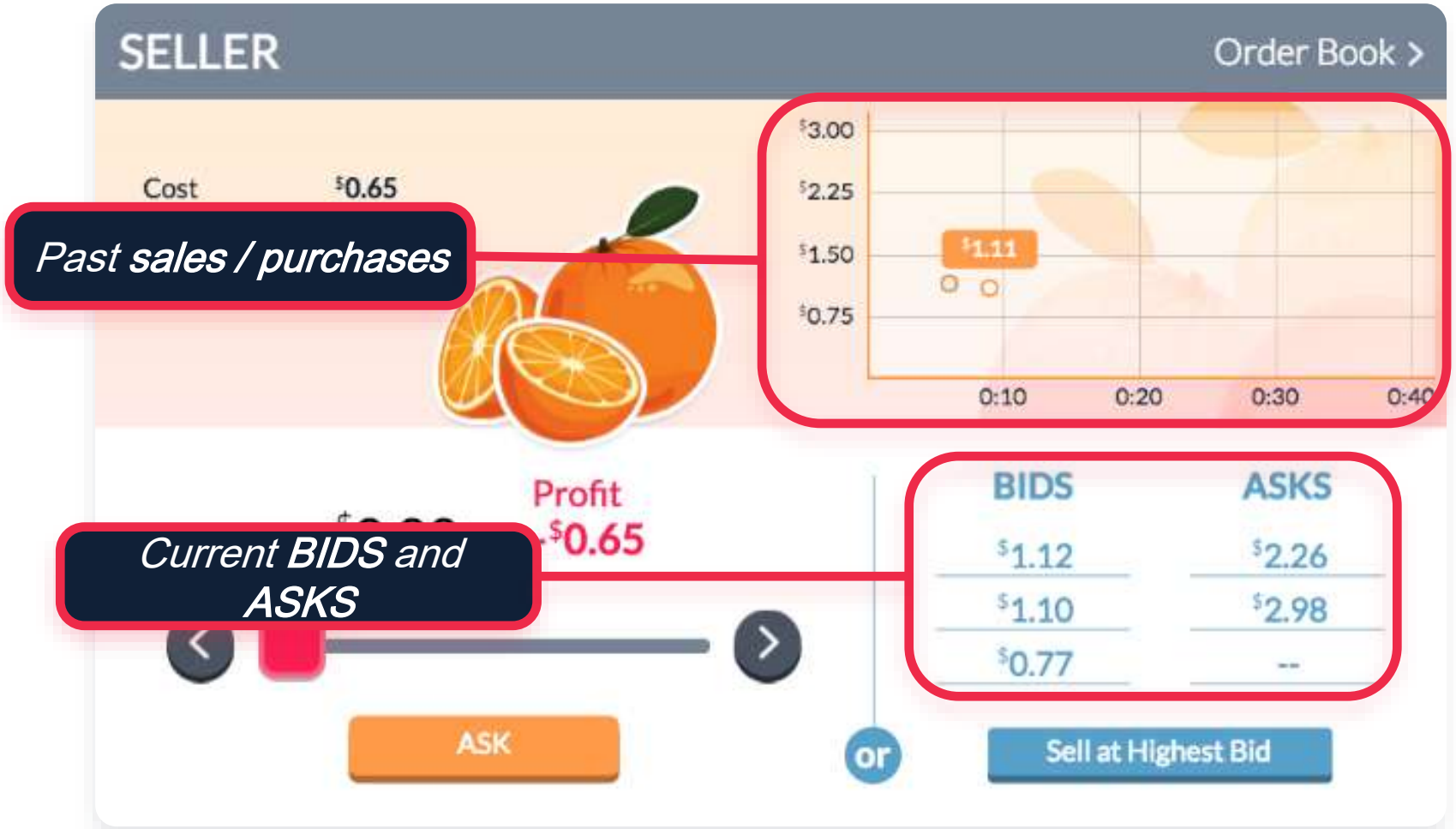
Game Instructions

Groups of 5 sellers and 5 buyers.
Trade to maximize your profits!

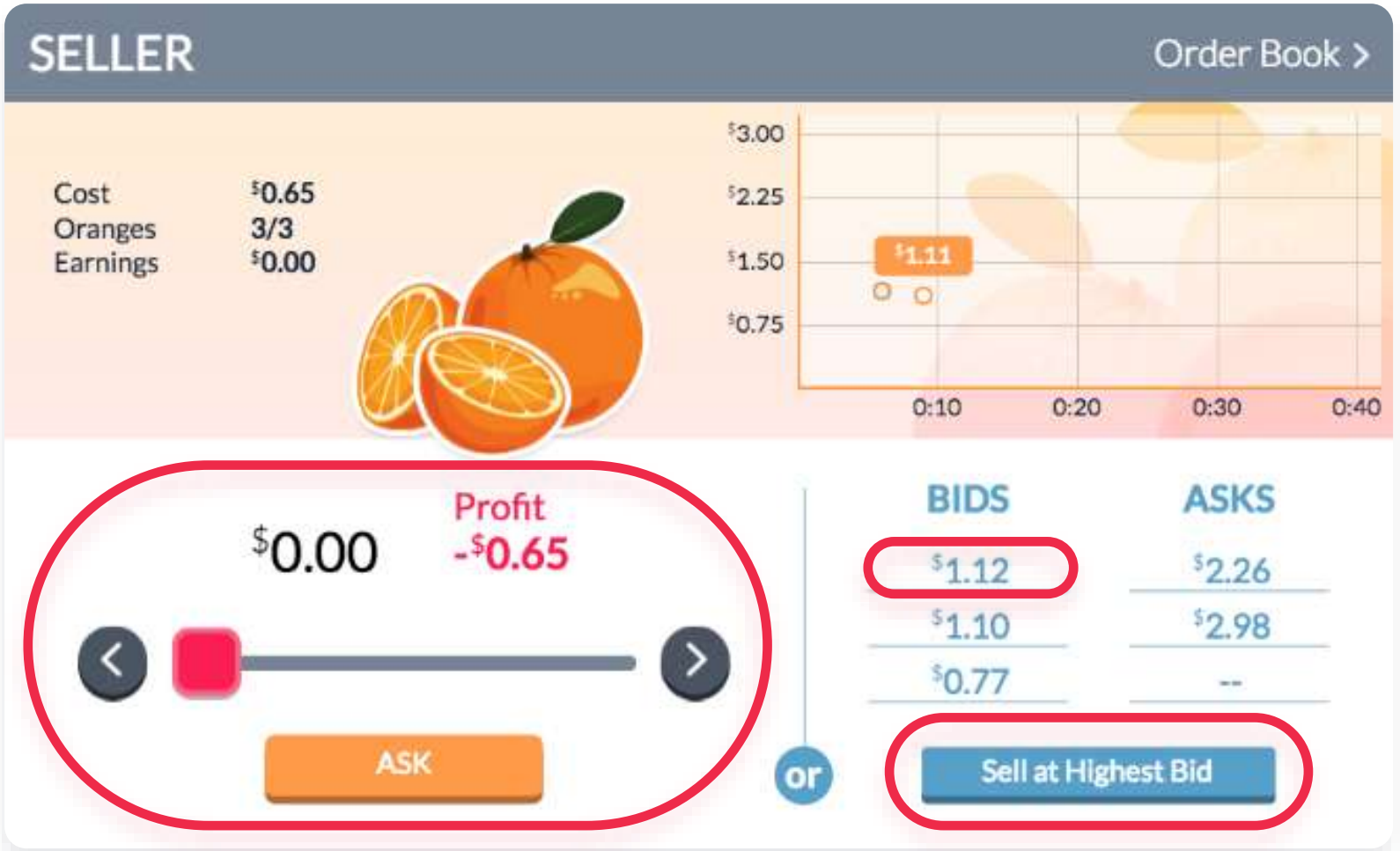
Orange producer, selling oranges

Hungry consumer, buying oranges

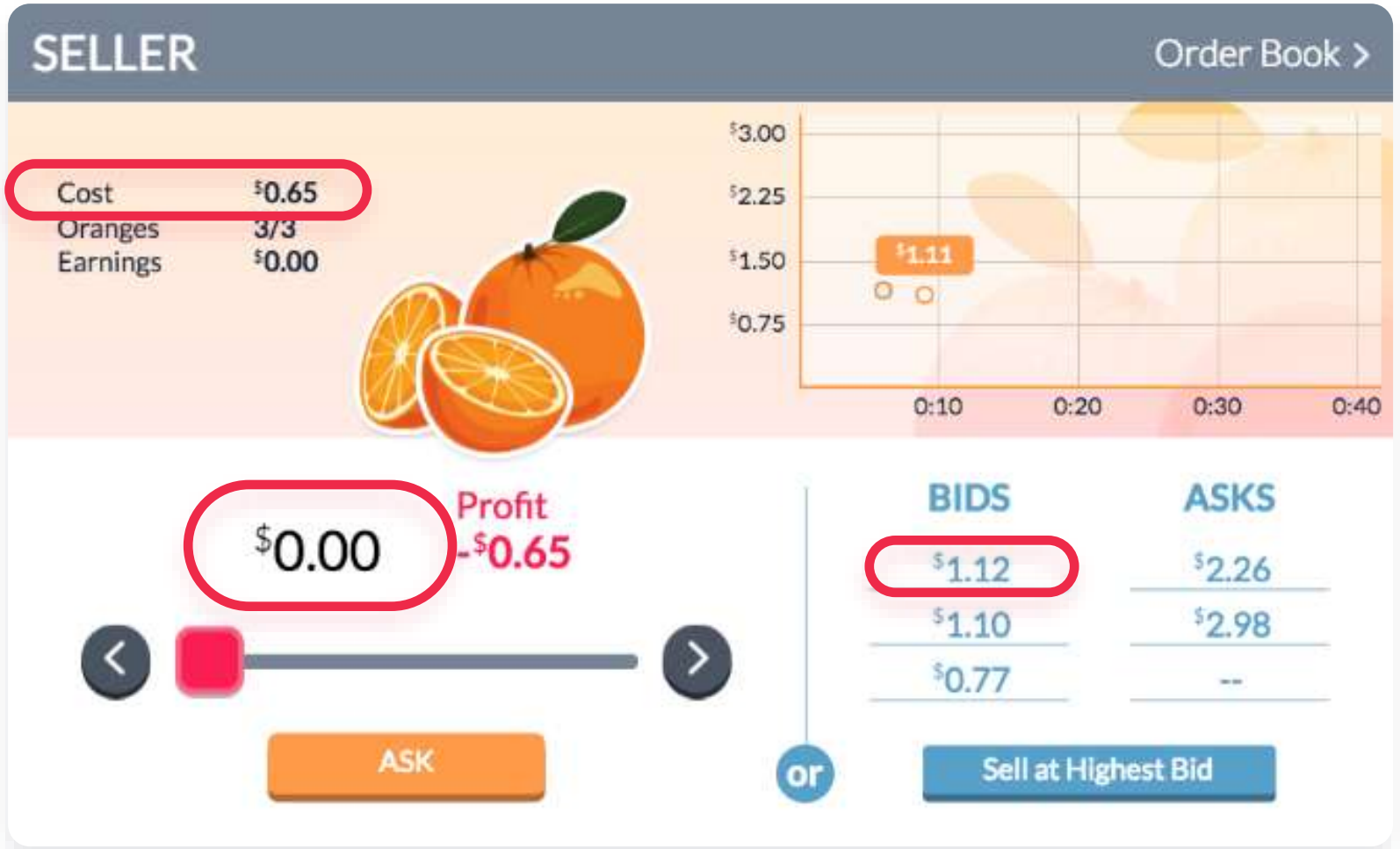
Bids are offers to buy



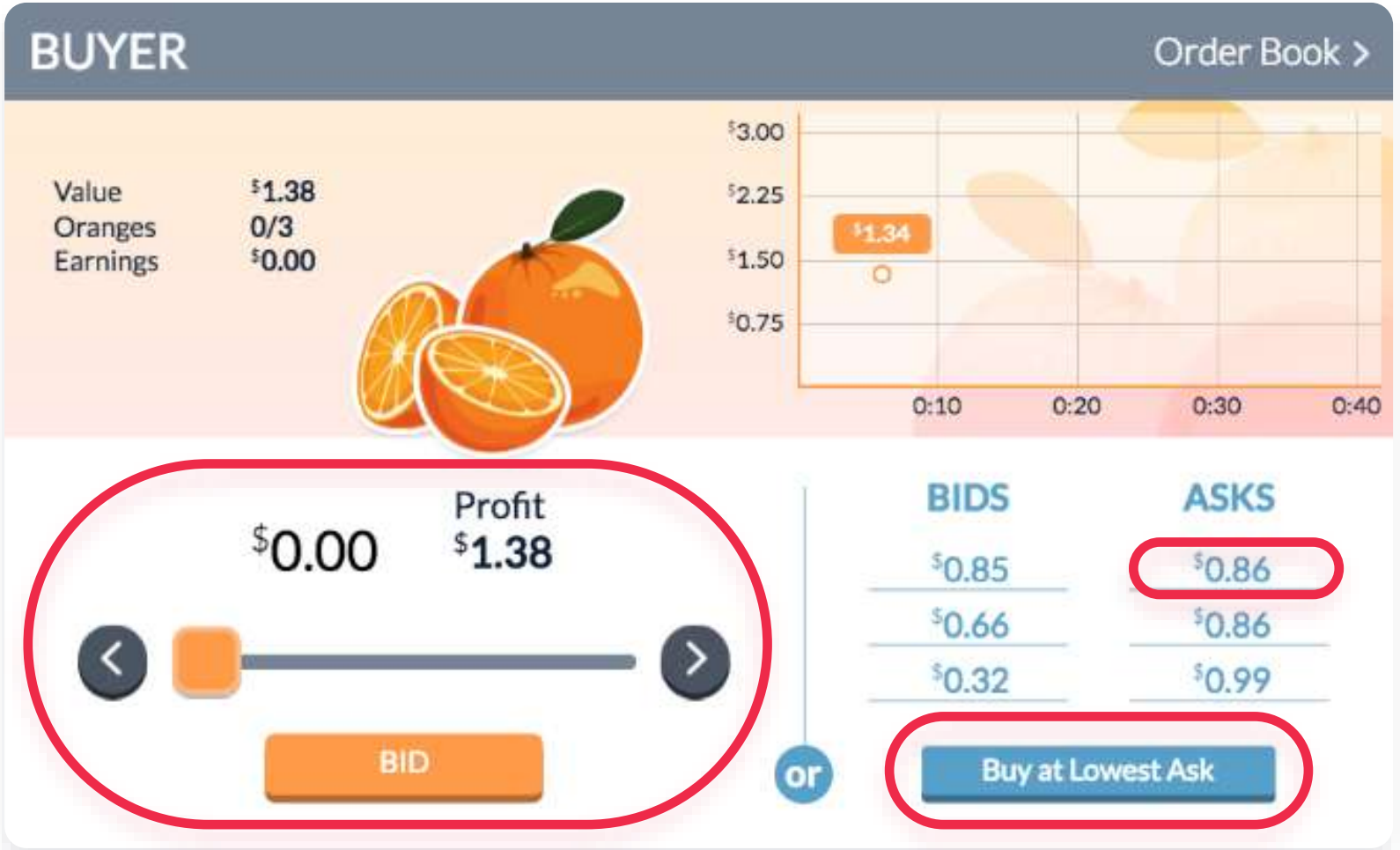
Submit an **ASK** using the slider, or **Sell at Highest Bid**



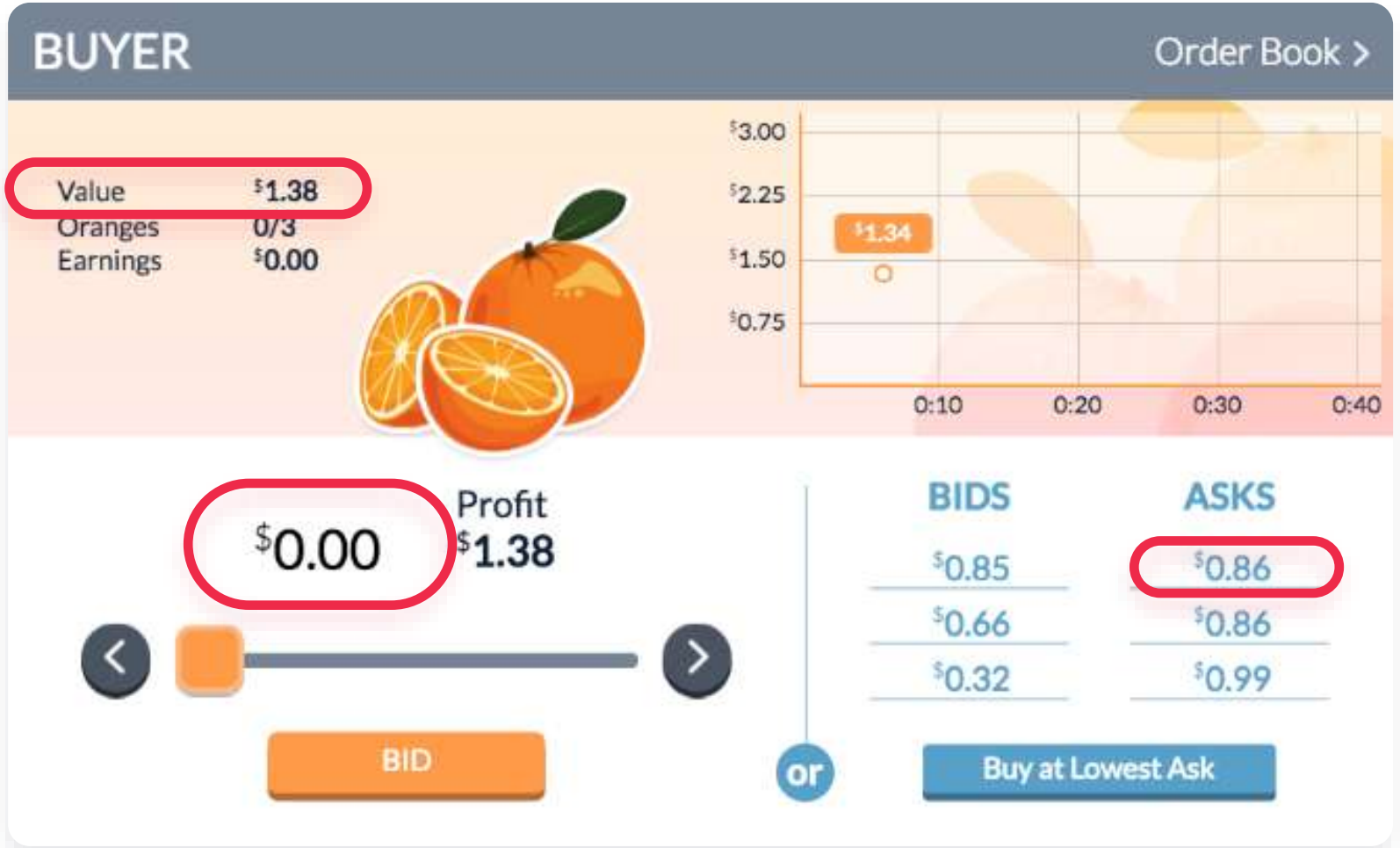
Seller's Profit = Sale Price – Cost of Production



Submit a **BID** using the slider, or **Buy at Lowest Ask**



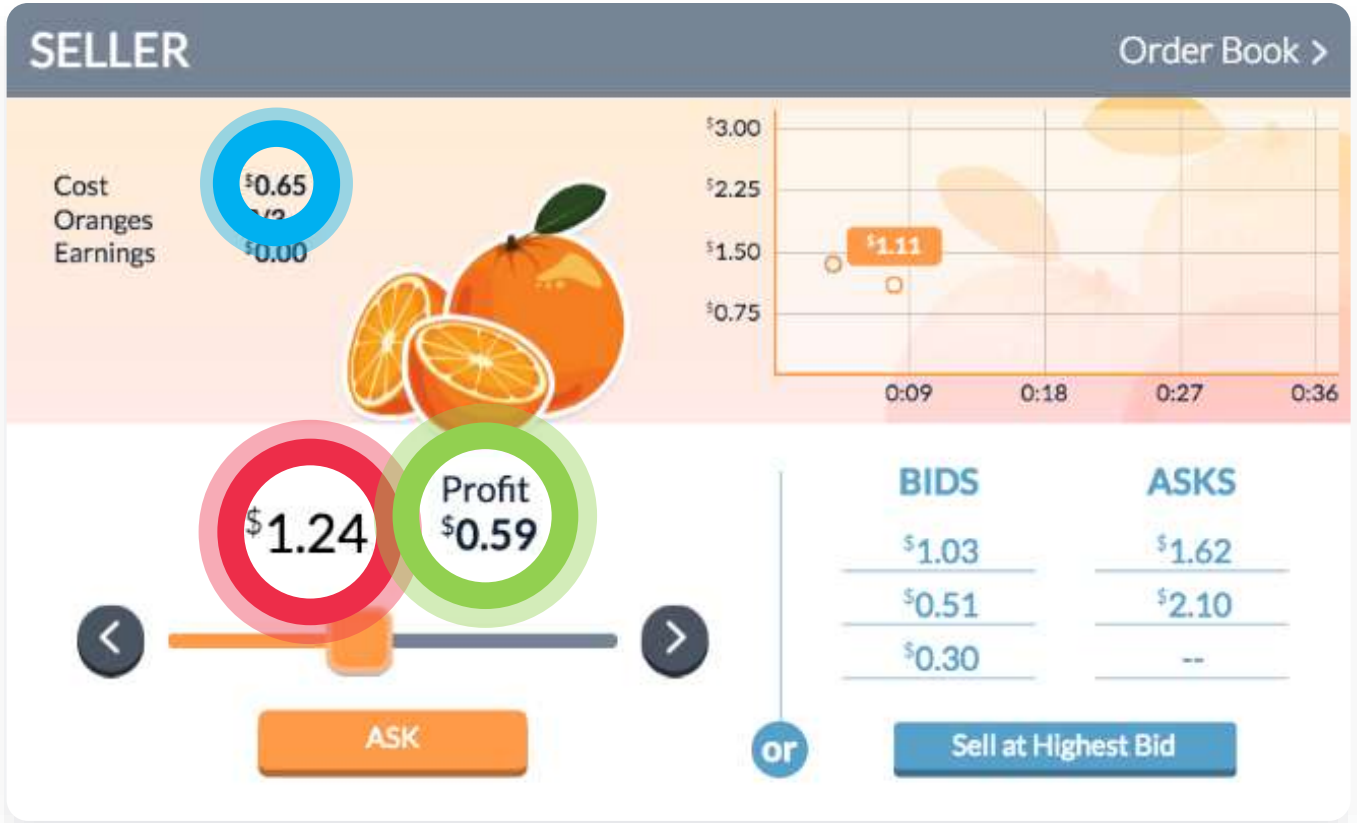
Buyer's Profit = Value of consumption – Purchase Price



When does a transaction occur?

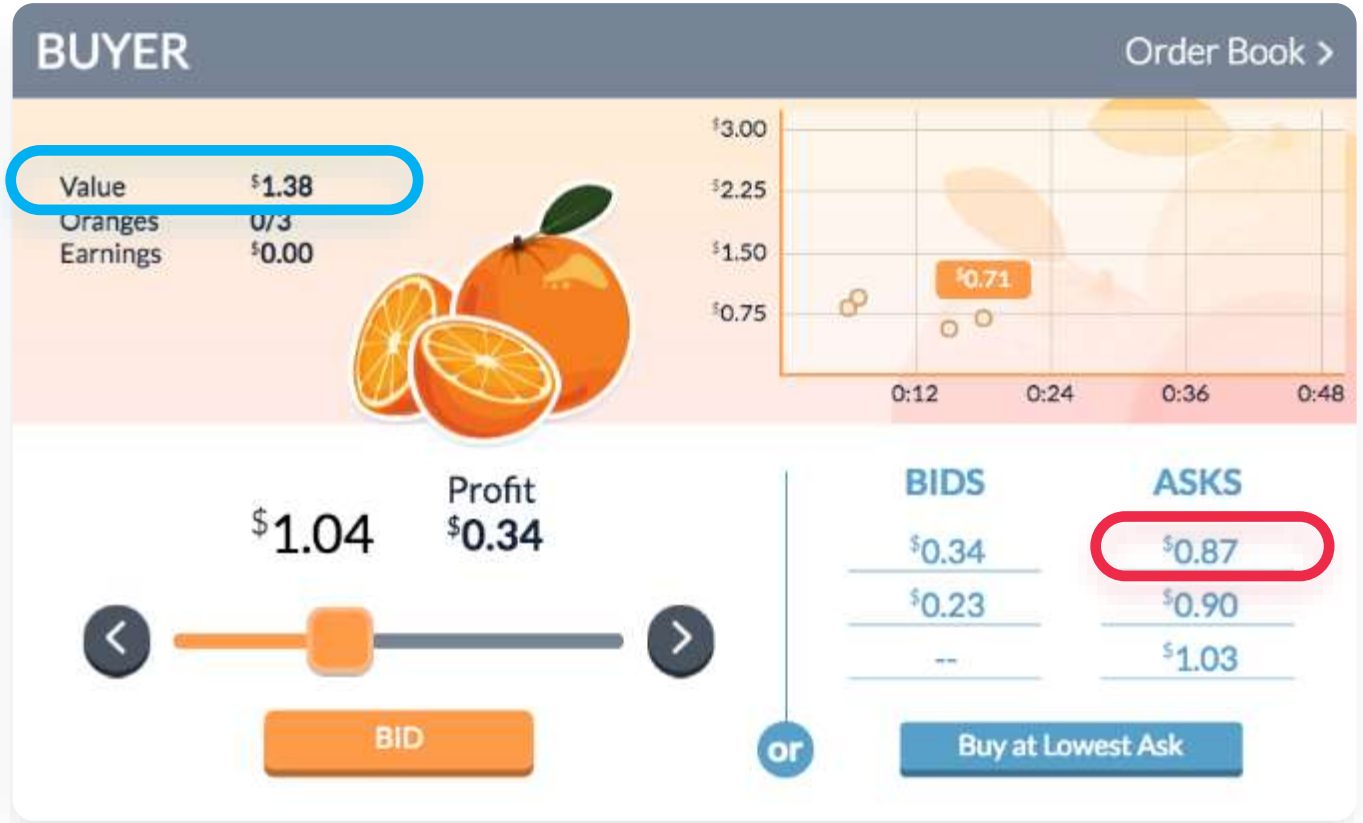
- Someone uses Buy at Lowest Ask Sell at Highest Bid
- A Buyer places a BID than the lowest outstanding Ask
- A Seller places an lower than the highest outstanding Bid ASK

As a **seller**, what is your profit if someone accepts your ask of $1.24?



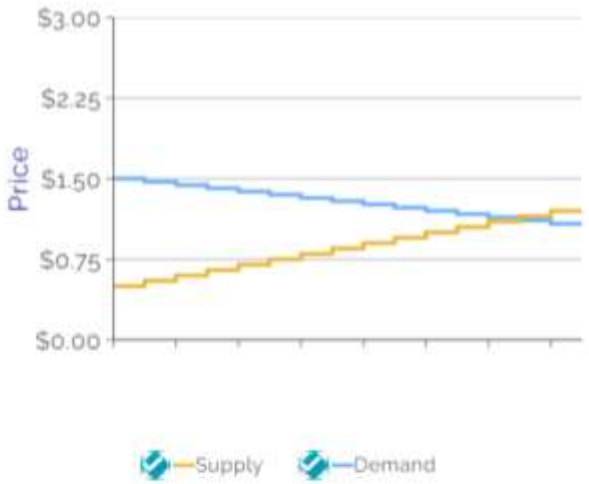
Sale Price – Cost = $1.24 - $0.65 = $0.59

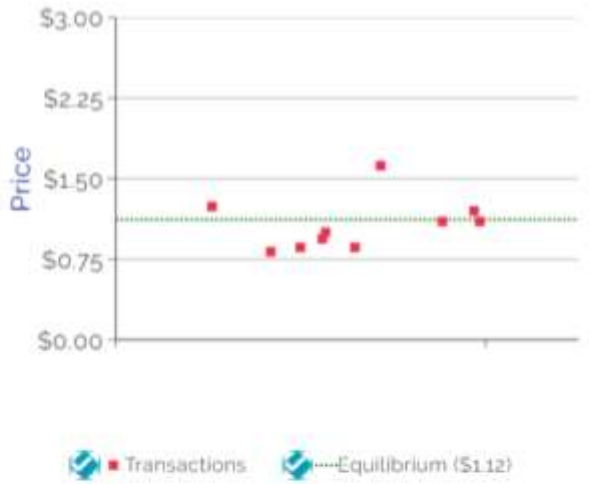
As a **buyer**, what is your profit if you "Buy at Lowest Ask"?

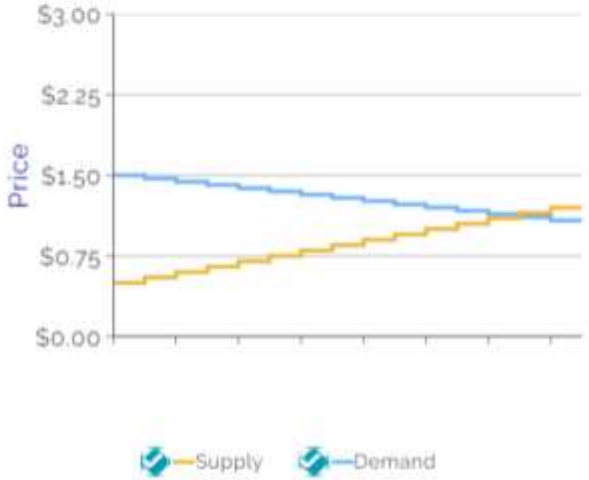


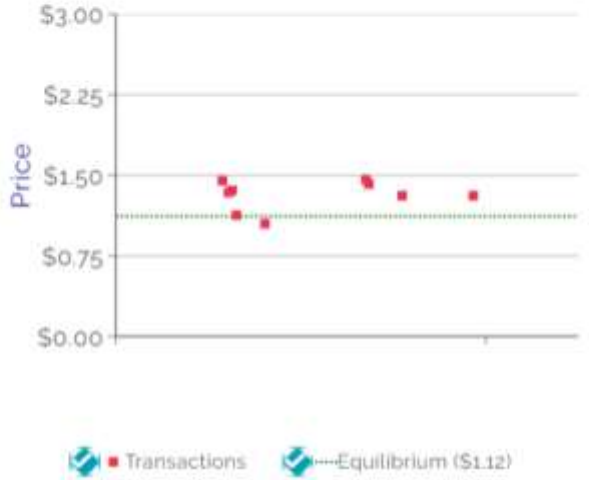
Value – Purchase Price = $1.38 - $0.87 = $0.51

MobLab Double Auction:
EE-BGT 21S Results: Round 1

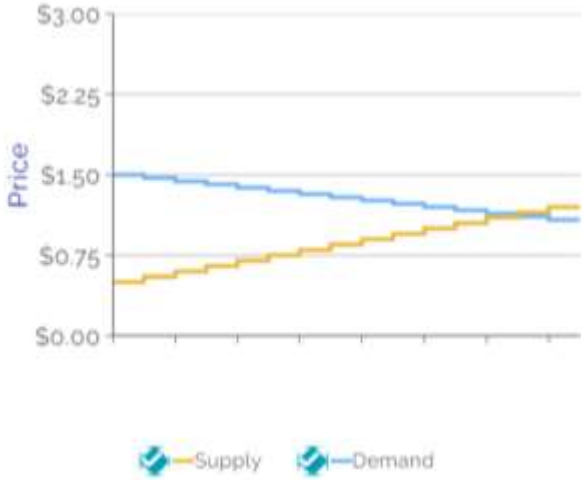


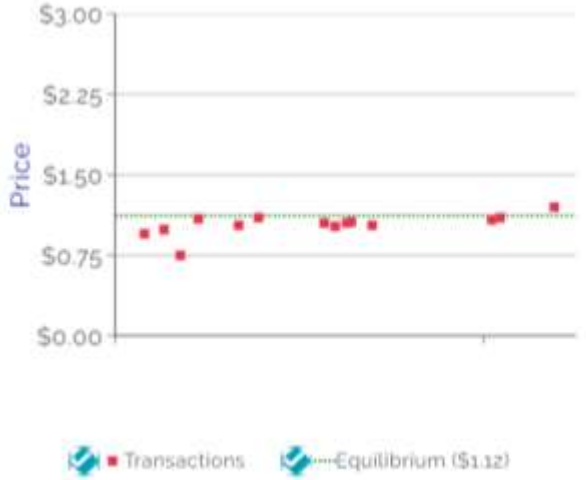


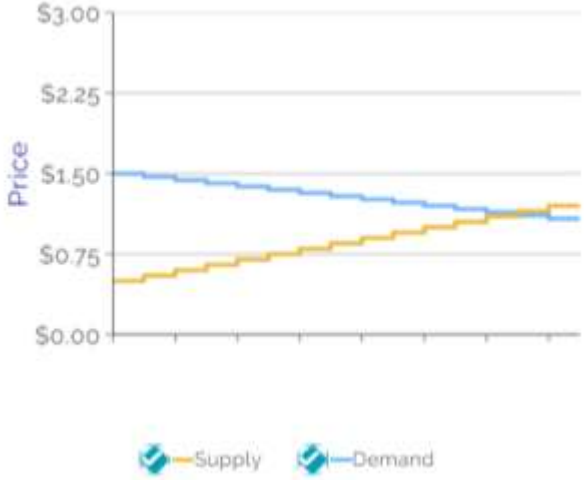




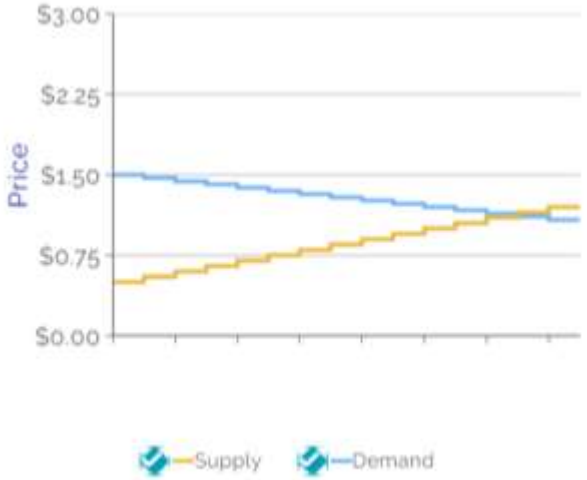
MobLab Double Auction:
EE-BGT 21S Results: Round 2

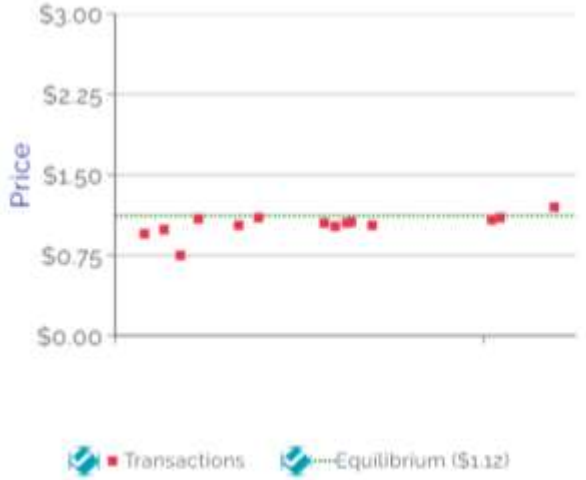


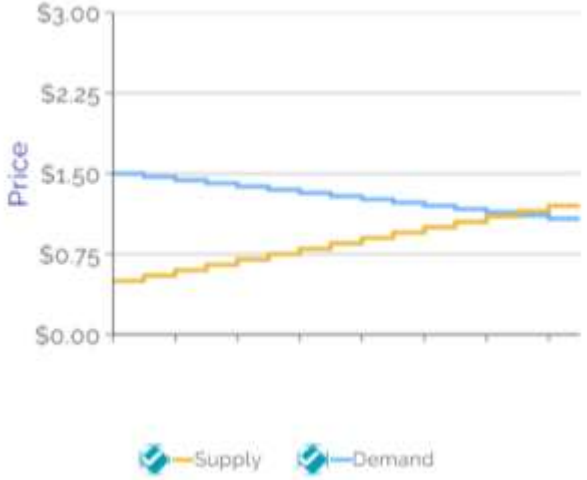


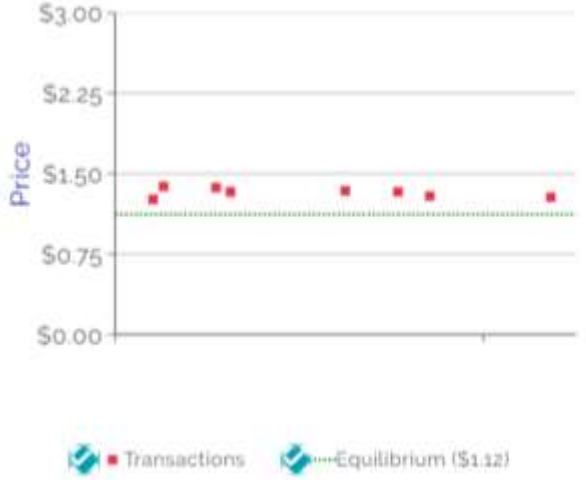


MobLab Double Auction:
EE-BGT 21S Results: Round 3

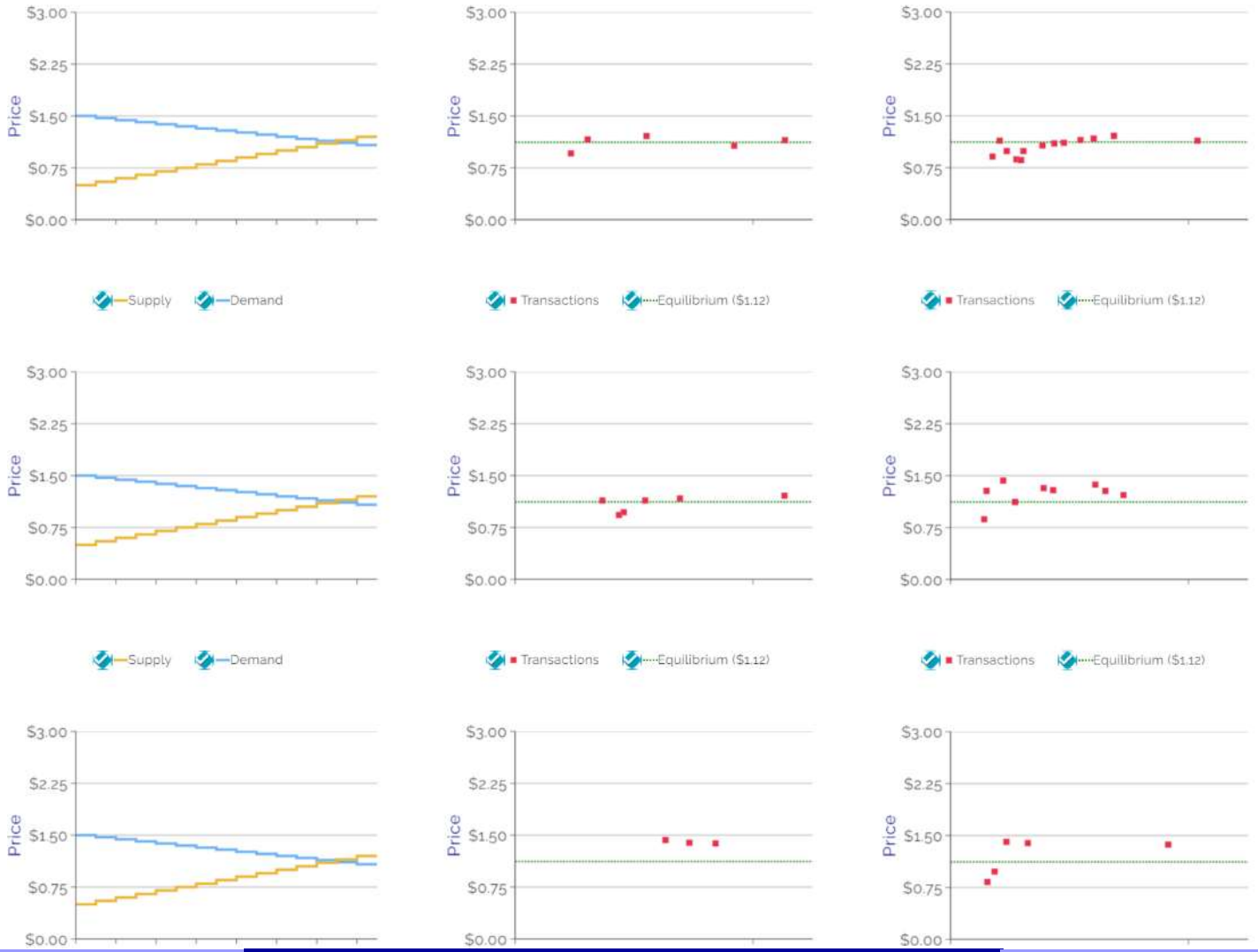




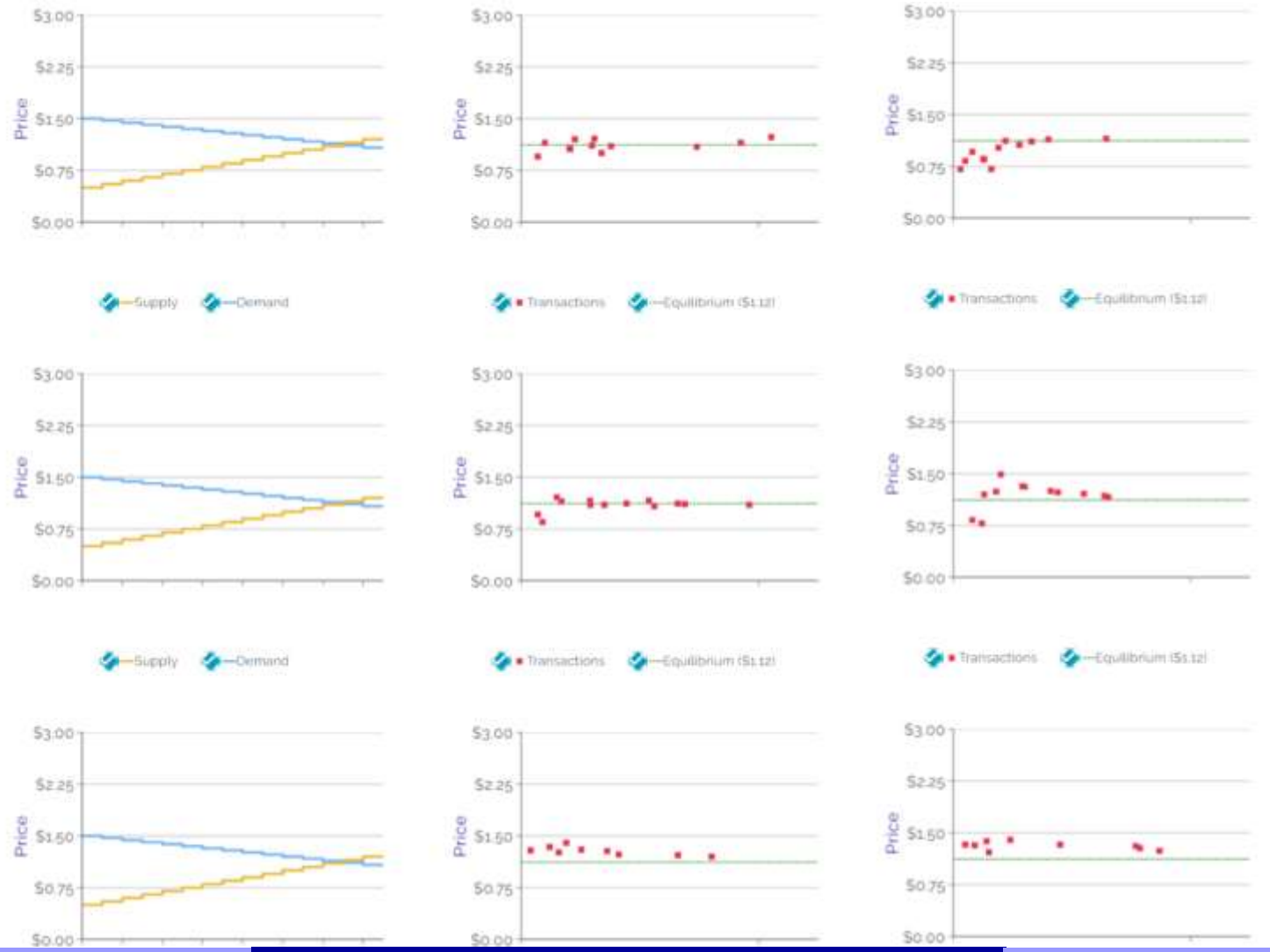




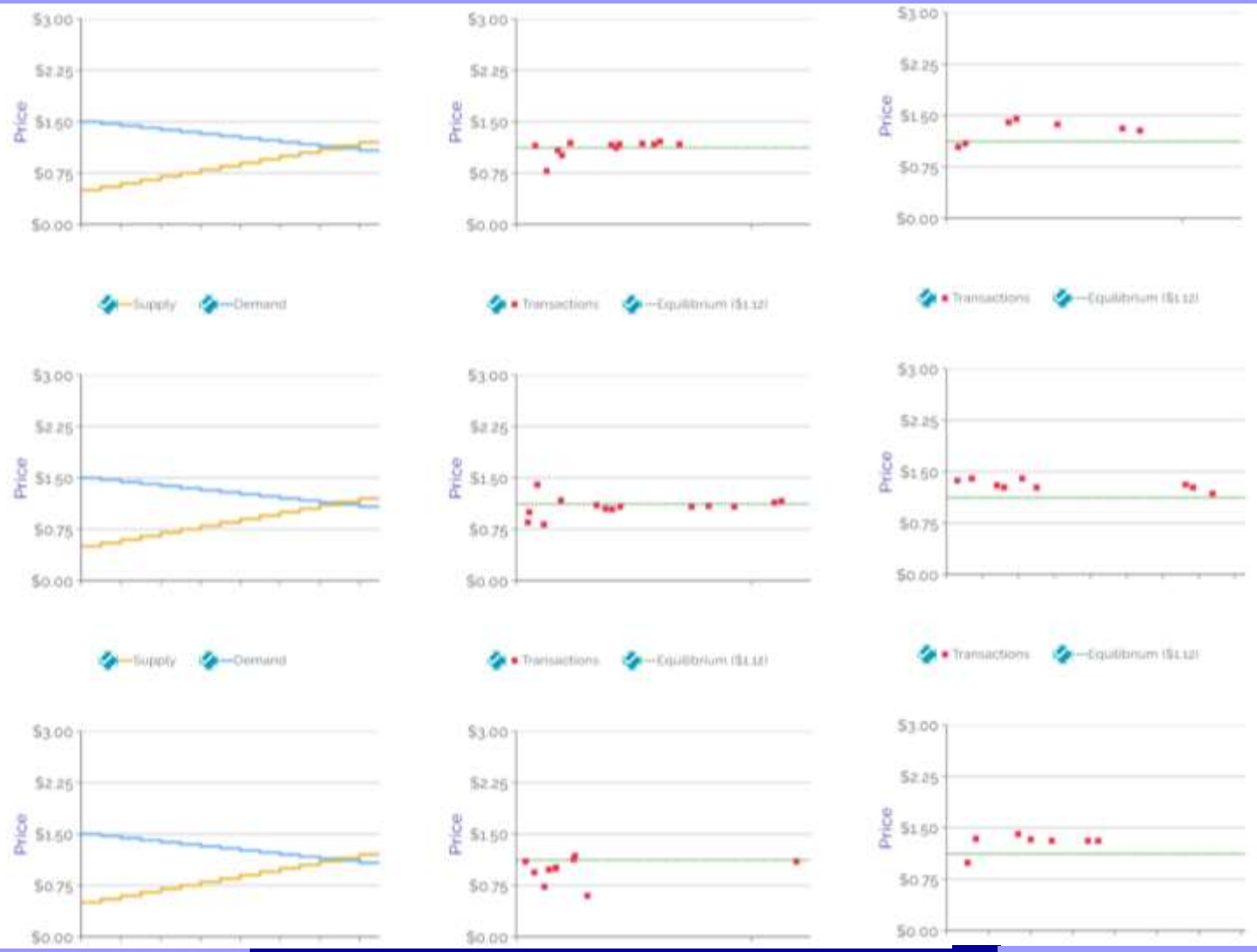
MobLab Double Auction:
CCU Results: Round 1



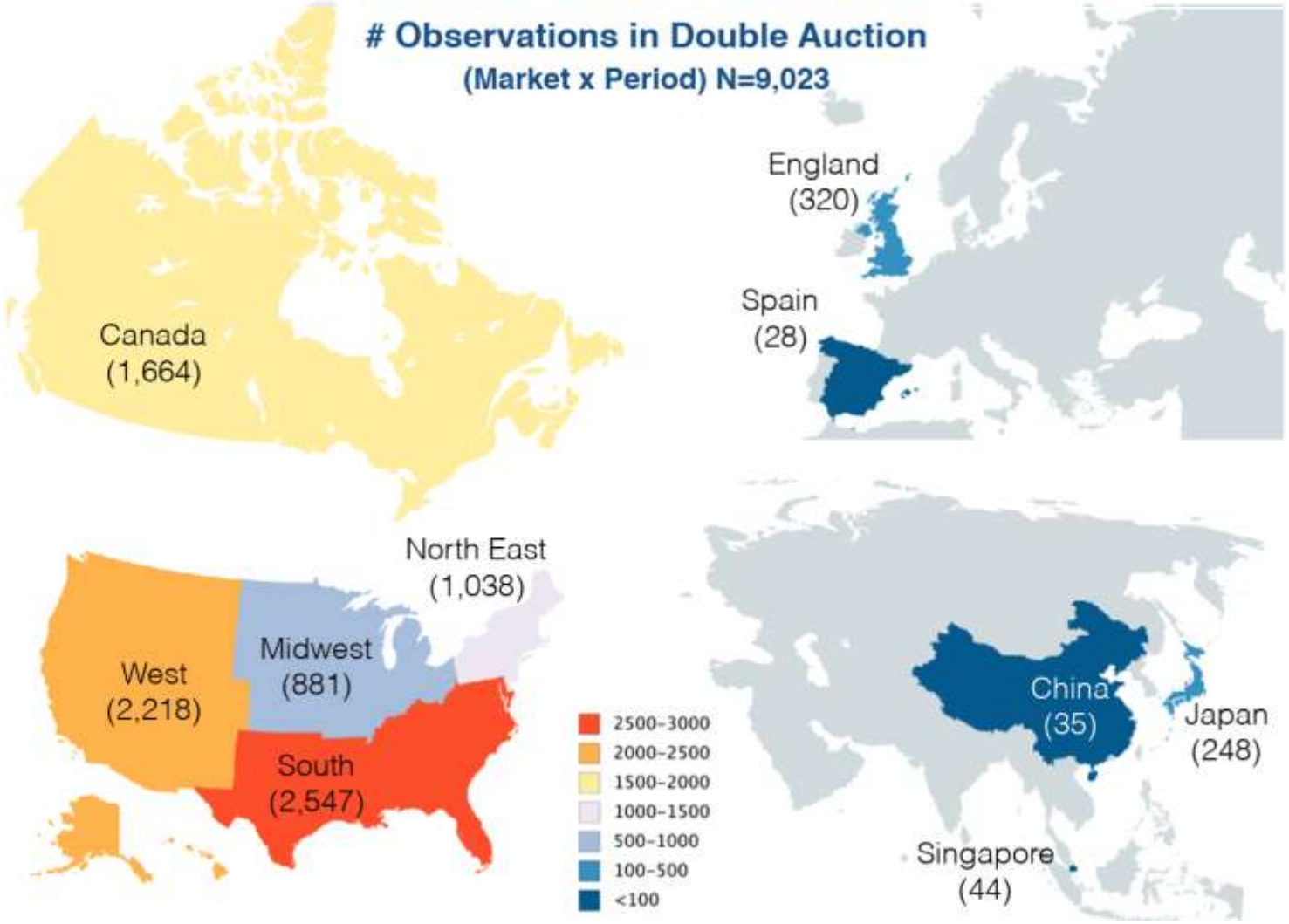
MobLab Double Auction:
CCU Results: Round 2



MobLab Double Auction:
CCU Results: Round 3

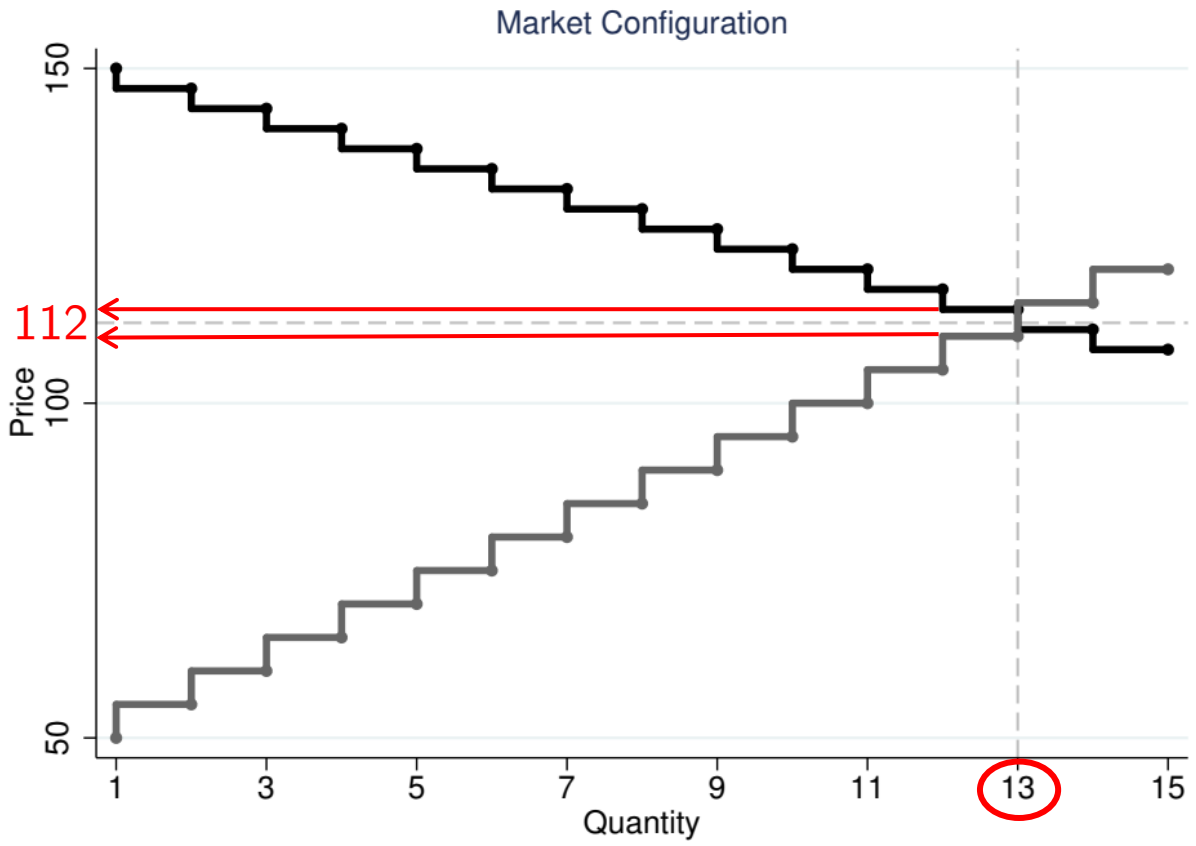


Observation in Different Regions/Countries



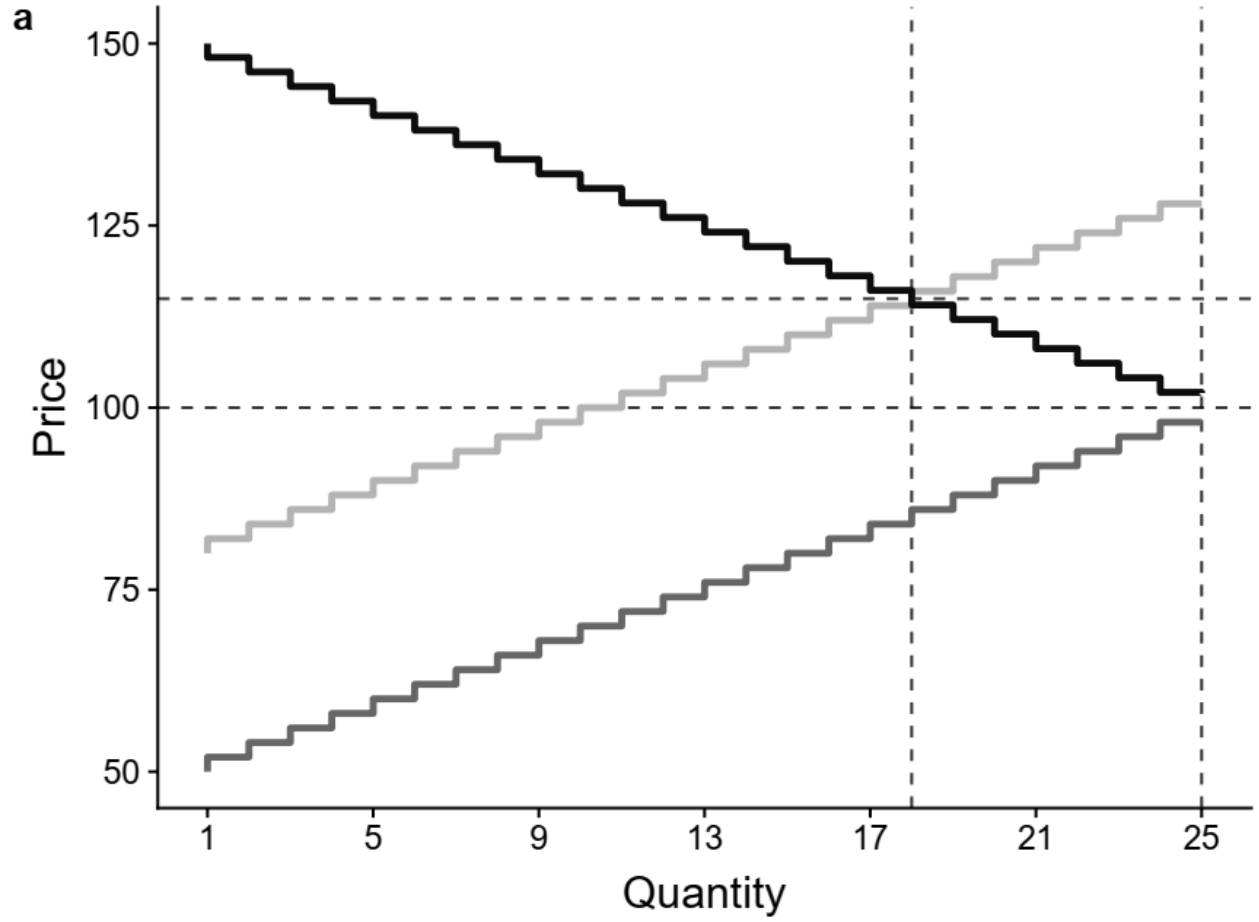
Default Configurations

- Ultimatum: Default pie size is 100
  - 82% used this
- Double Auction:
  - 5 buyers,
  - 5 sellers,
- each player has 3 cost/value
  - 36% used this



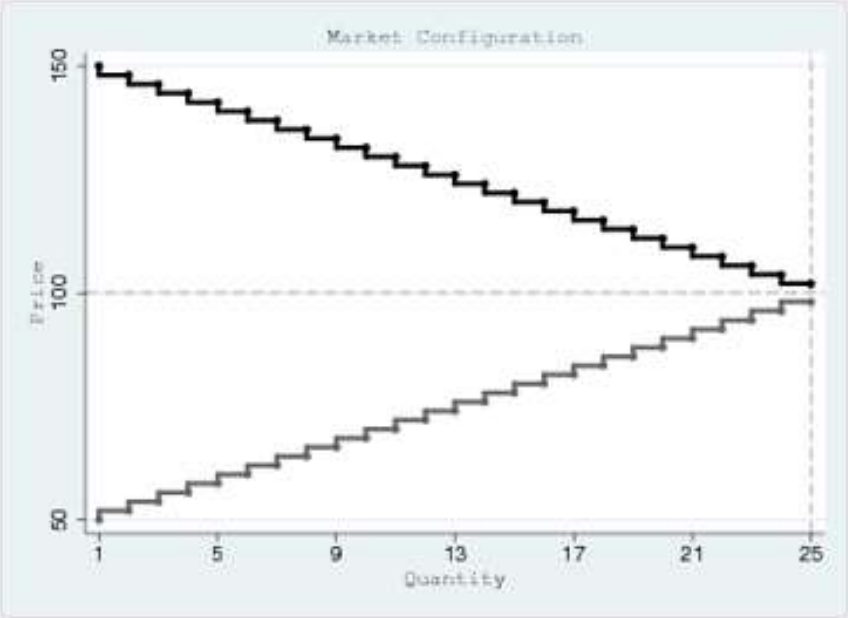
Fewer Default in Double Auction Since...

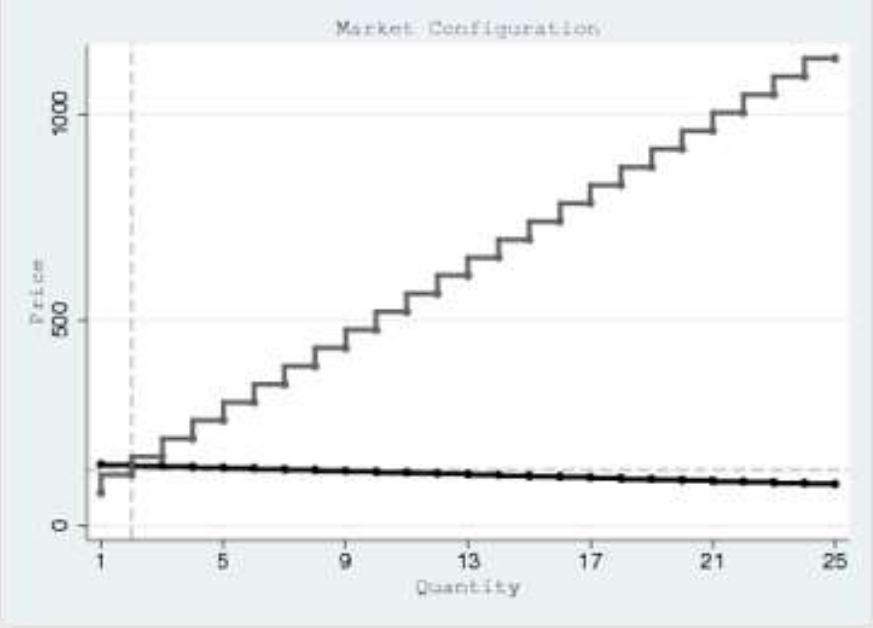
Instructors Frequently Shift S-D when Teaching



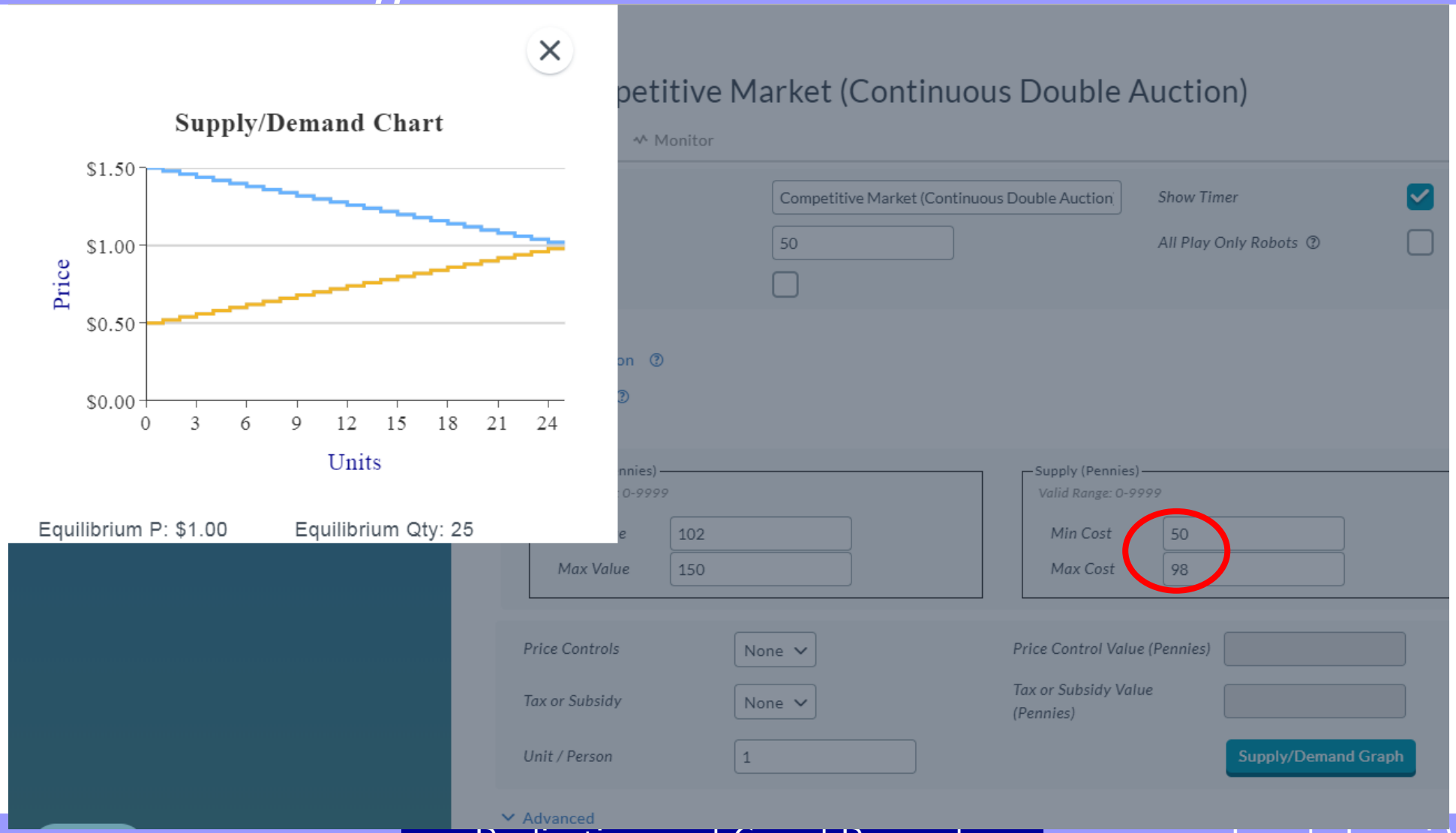
One Outlier Excluded in Double Auction

Original Market #228 vs. Outlier Market #1750

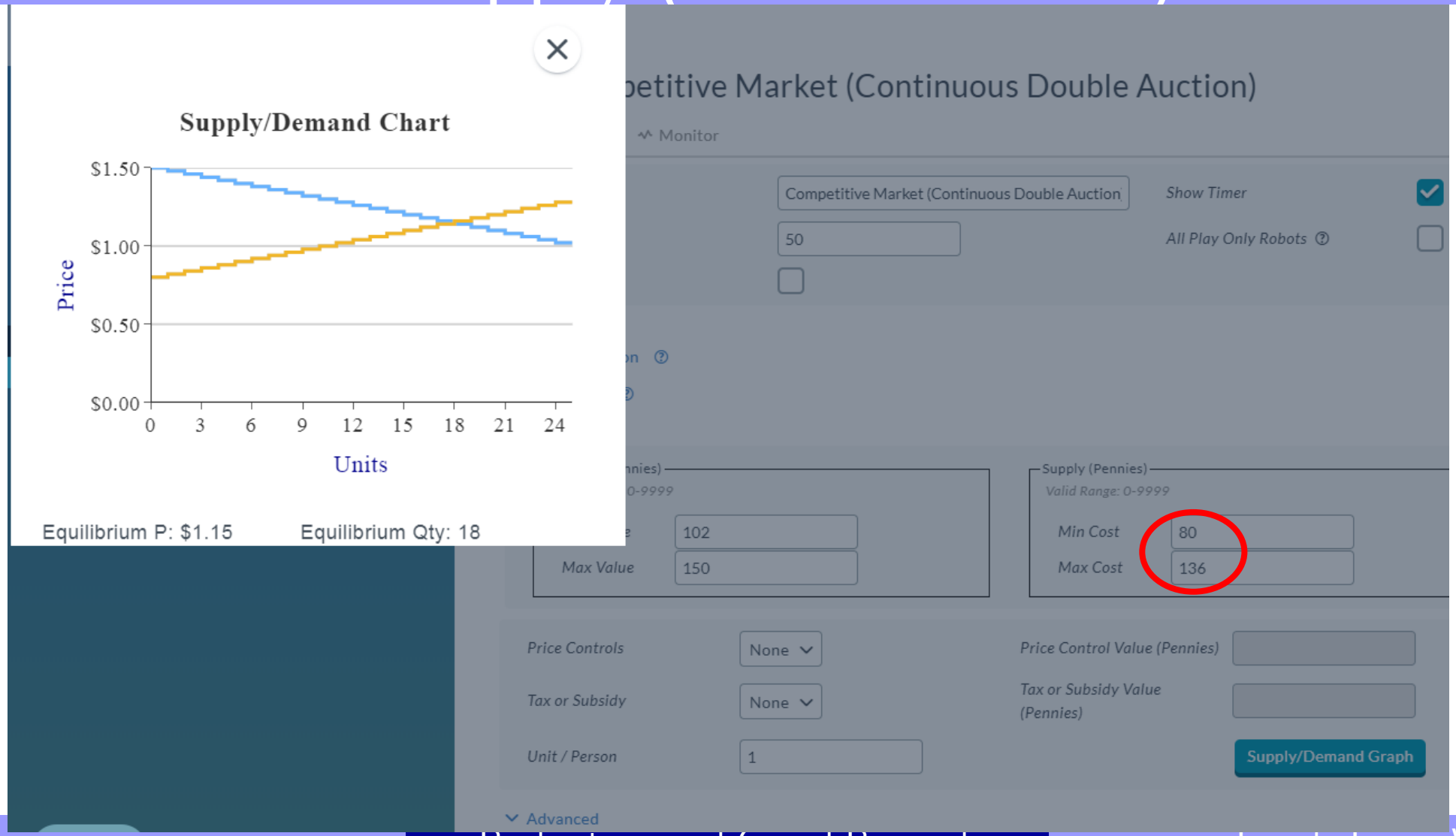




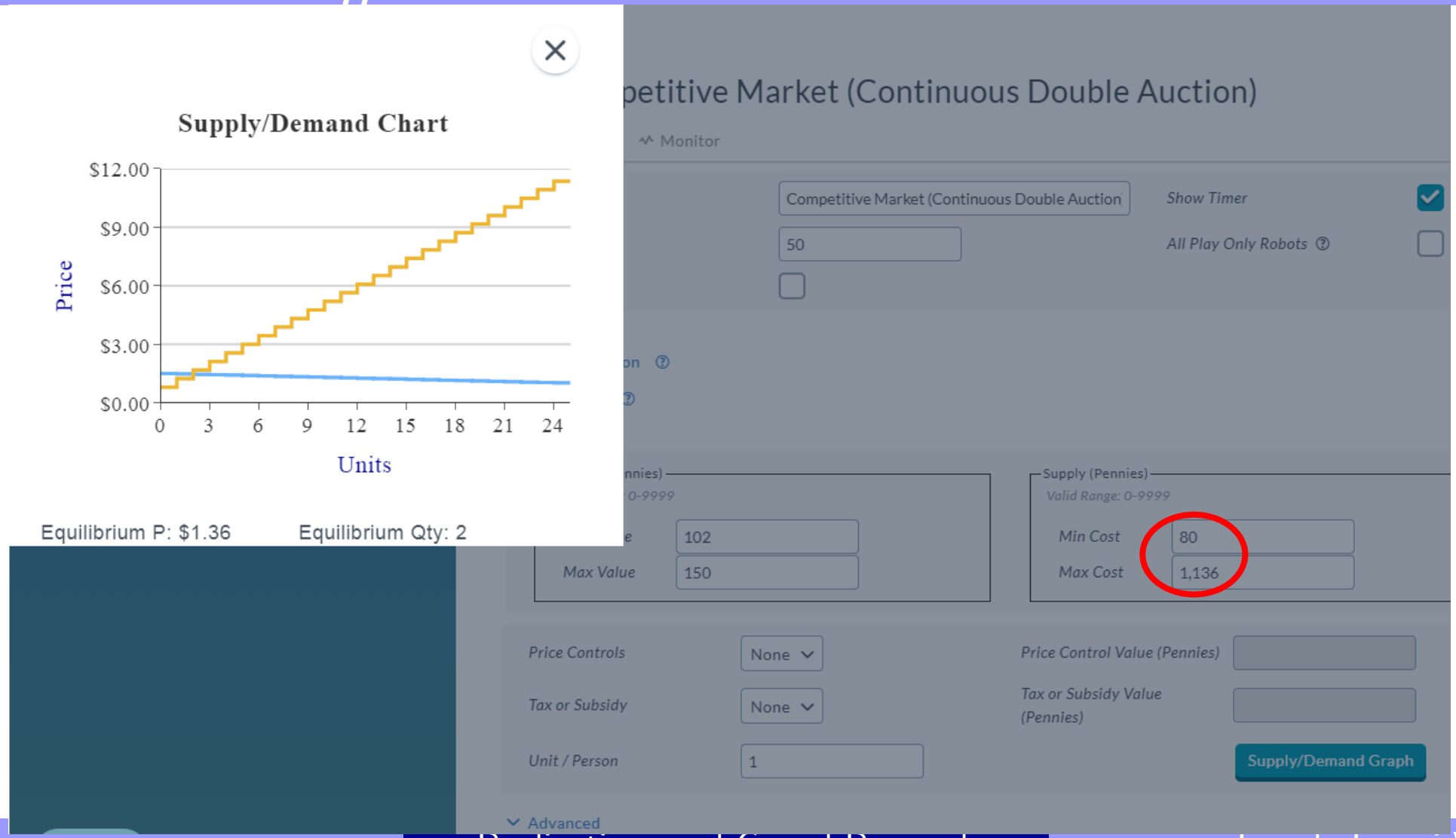
Original Market #228



Intended Shift in Supply (Didn't Occur)

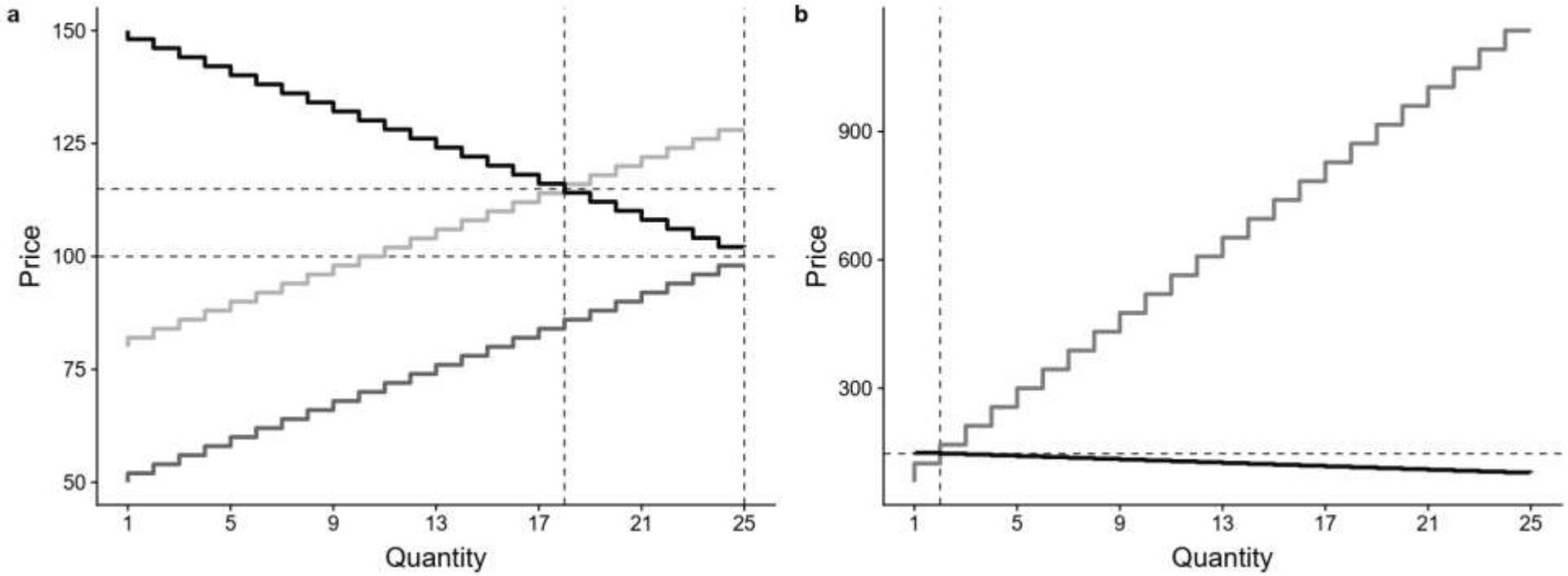


Outlier Market #1750



One Outlier Excluded in Double Auction

Original Market #228 vs. Outlier Market #1750



Summary Statistics

	Mean	(s. d.)
Double Auction	(5,809 Markets)	
MED δ - Accuracy	0.070	(0.280)
Smith's α - Fluctuation	0.279	(0.294)
Efficiency	81.5%	(25.8%)
Ultimatum Game	(6,505 Pairs)	
Proposal Offer	36.82	(18.16)
Acceptance Rate	64.0%	(48.0%)

Mean Error Deviation (MED):

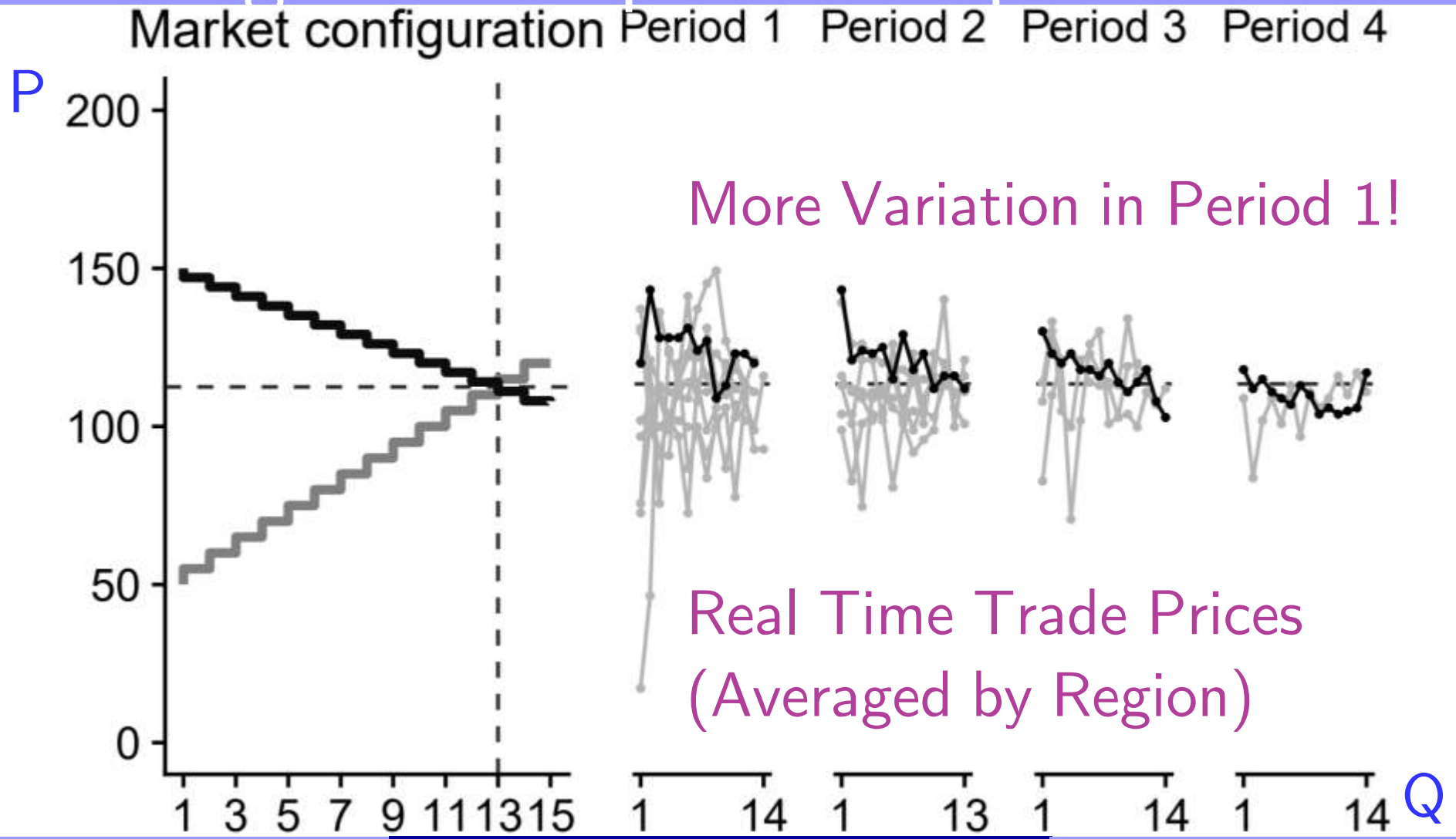
$$\delta = \frac{1}{Q} \sum_{q=1}^{Q} \frac{P_q - P_{CE}}{P_{CE}}$$

Smith's Alpha:

$$\alpha = \frac{\sqrt{\frac{1}{Q} \sum_{q=1}^{Q} (P_q - P_{CE})^2}}{P_{CE}}$$

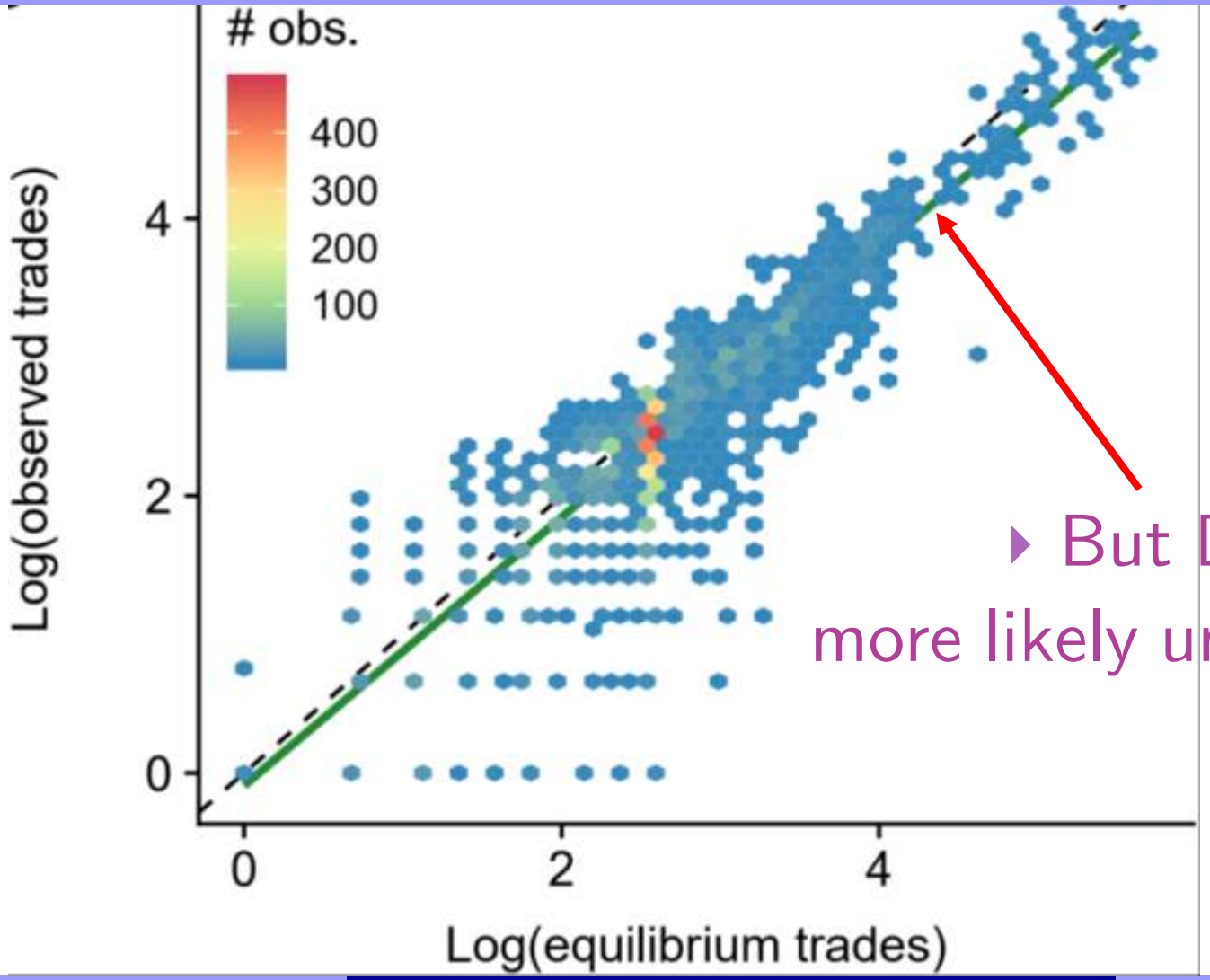
MobLab Double Auction: Lin et al. (2020)

Prices Converge to Competitive Equilibrium



Double Auction:

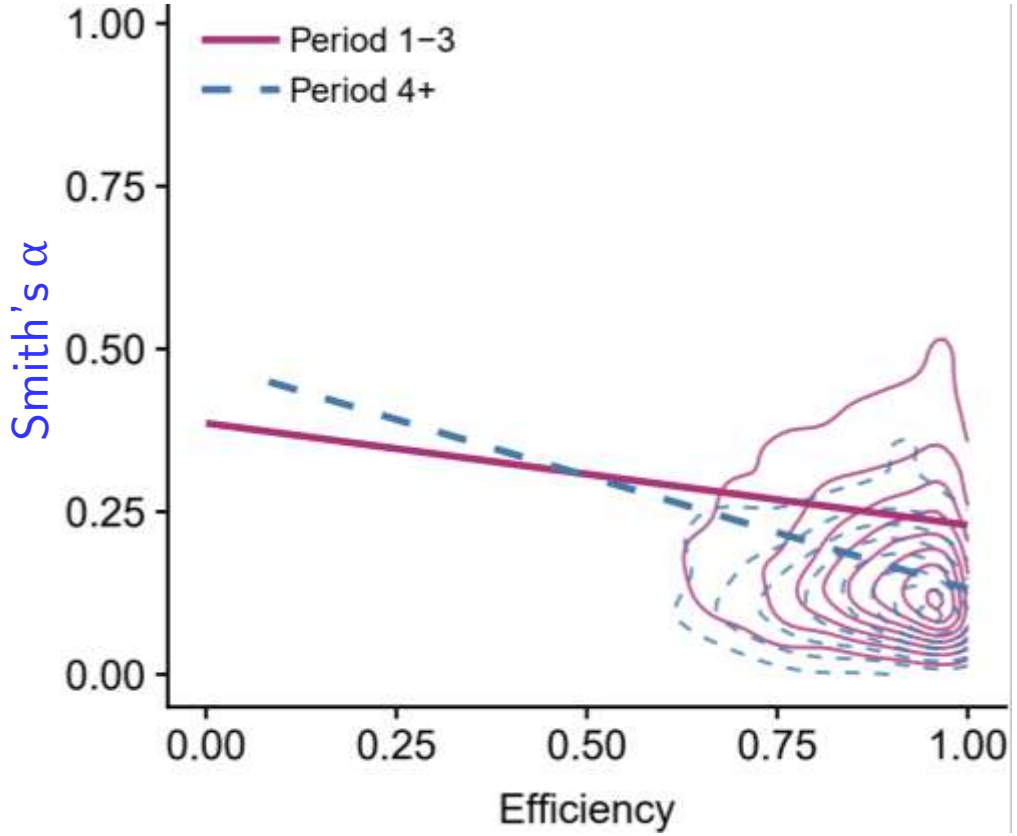
Trade Volume Close to CE!



- But Deviations more likely under-trade

MobLab Double Auction:
Between-Period Price Convergence to CE

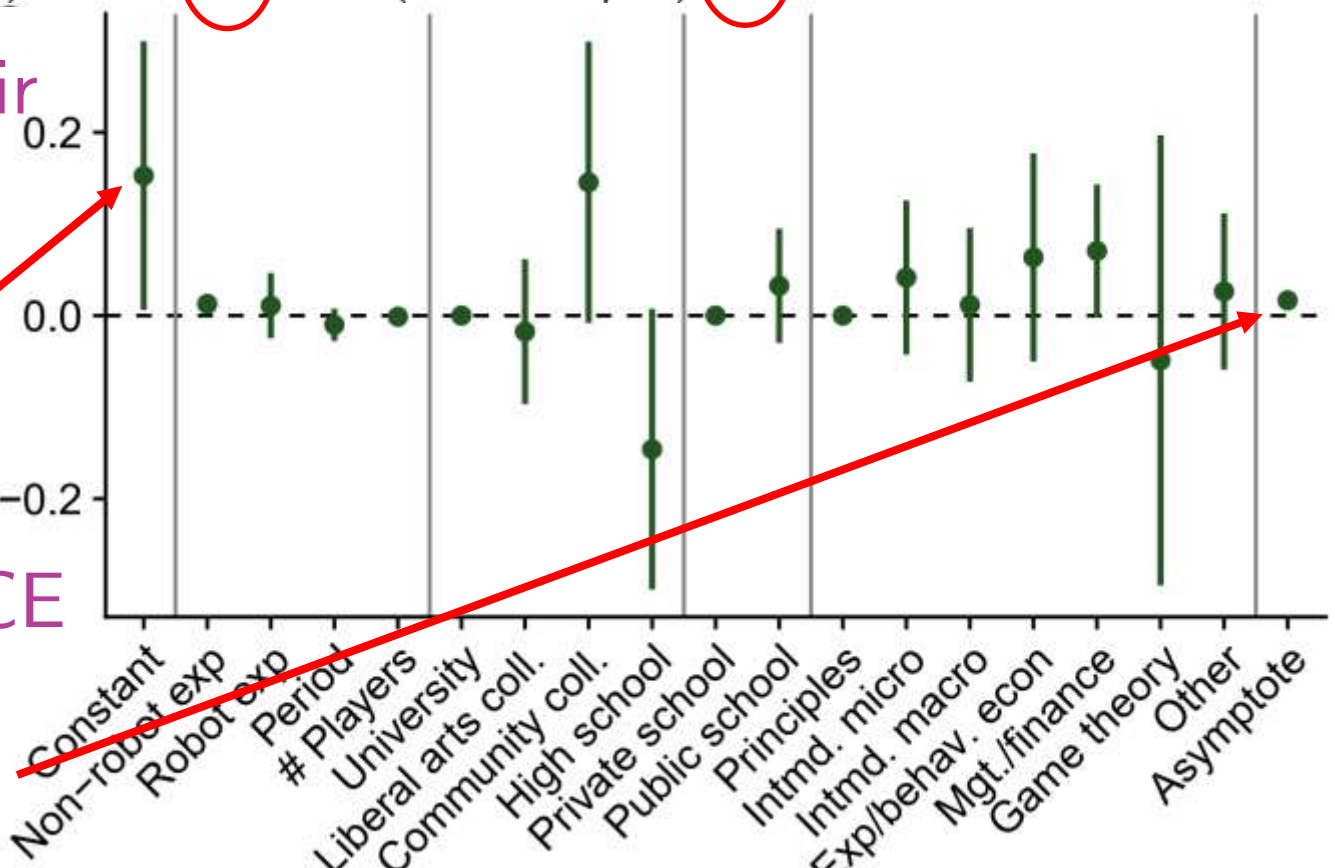
- Negative Relation Between:
- Smith's α
  - Converge from 20.6% to 8.6% (in 25 rounds)
- Efficiency
  - Stable at 92%
- Benchmark:
  - Ketcham et al. (1984)
    - Asymptotic Smith's α = 5.9%; Efficiency: around 95.89%



MobLab Double Auction:
Within-Period Price Convergence to CE

$$y_{it} = (1/t)\mathbb{X}_i \cdot \beta_1 + (1 - 1/t)\beta_2 + \epsilon_{it},$$

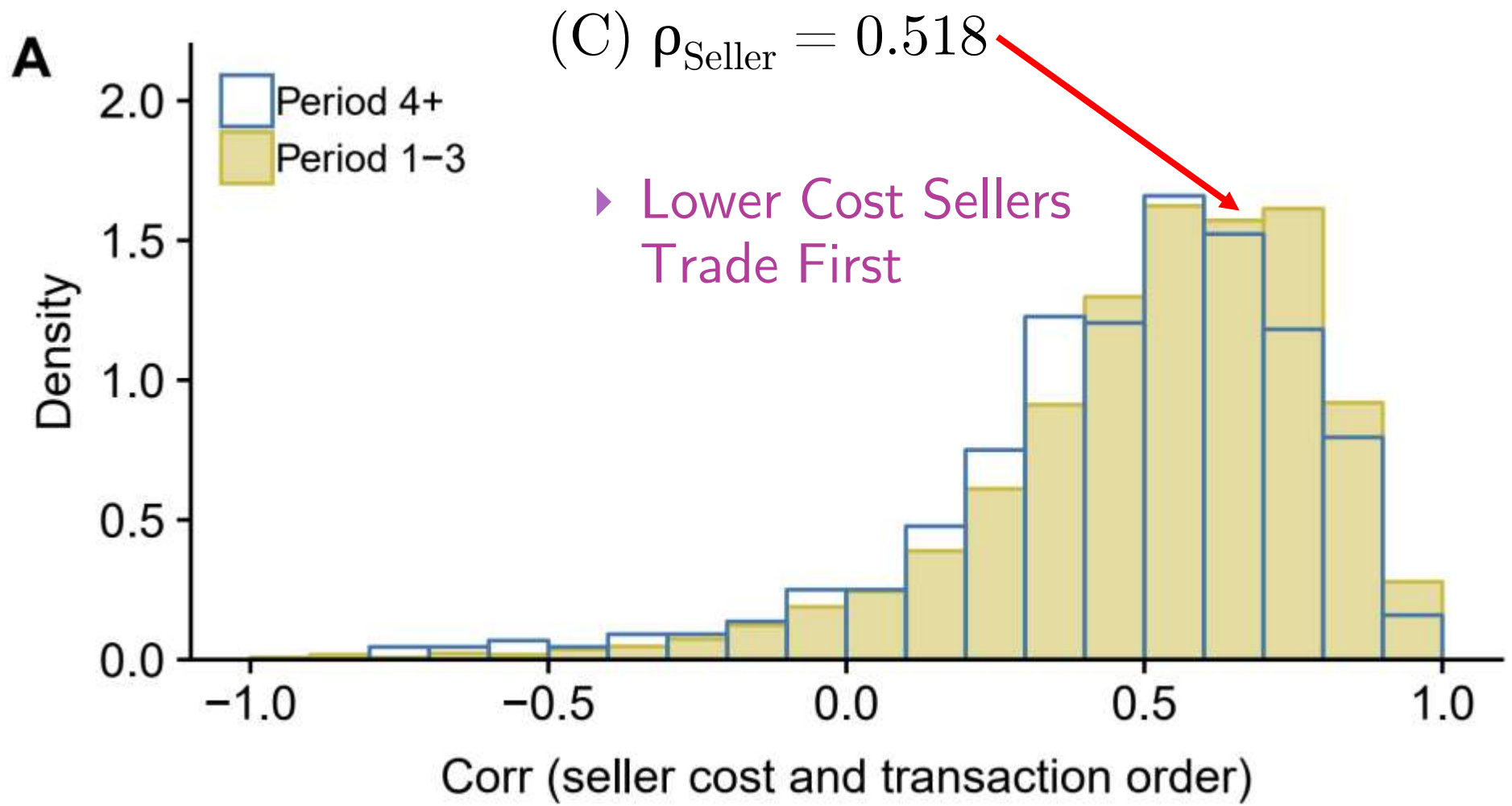
- as in Noussair et al. (1995)



- 1st Trade:
  - On average 15.3% above CE
    - (t=1)
- Converge to:
  - 1.7% above CE

MobLab Double Auction: Seller Rank-Order Correlation

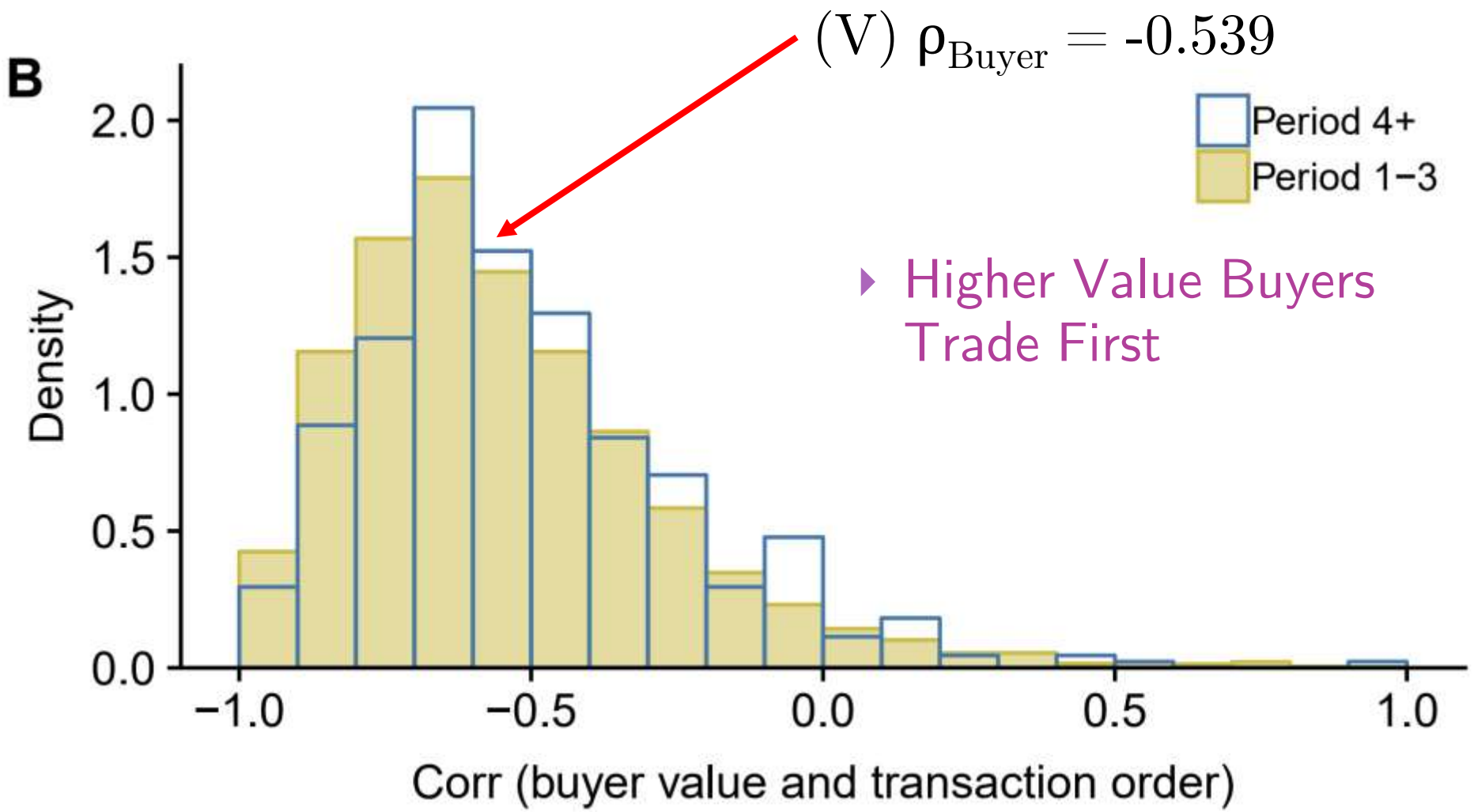
Correlation(Transaction Order, Seller Cost)



- Lower Cost Sellers Trade First

MobLab Double Auction: Buyer Rank-Order Correlation

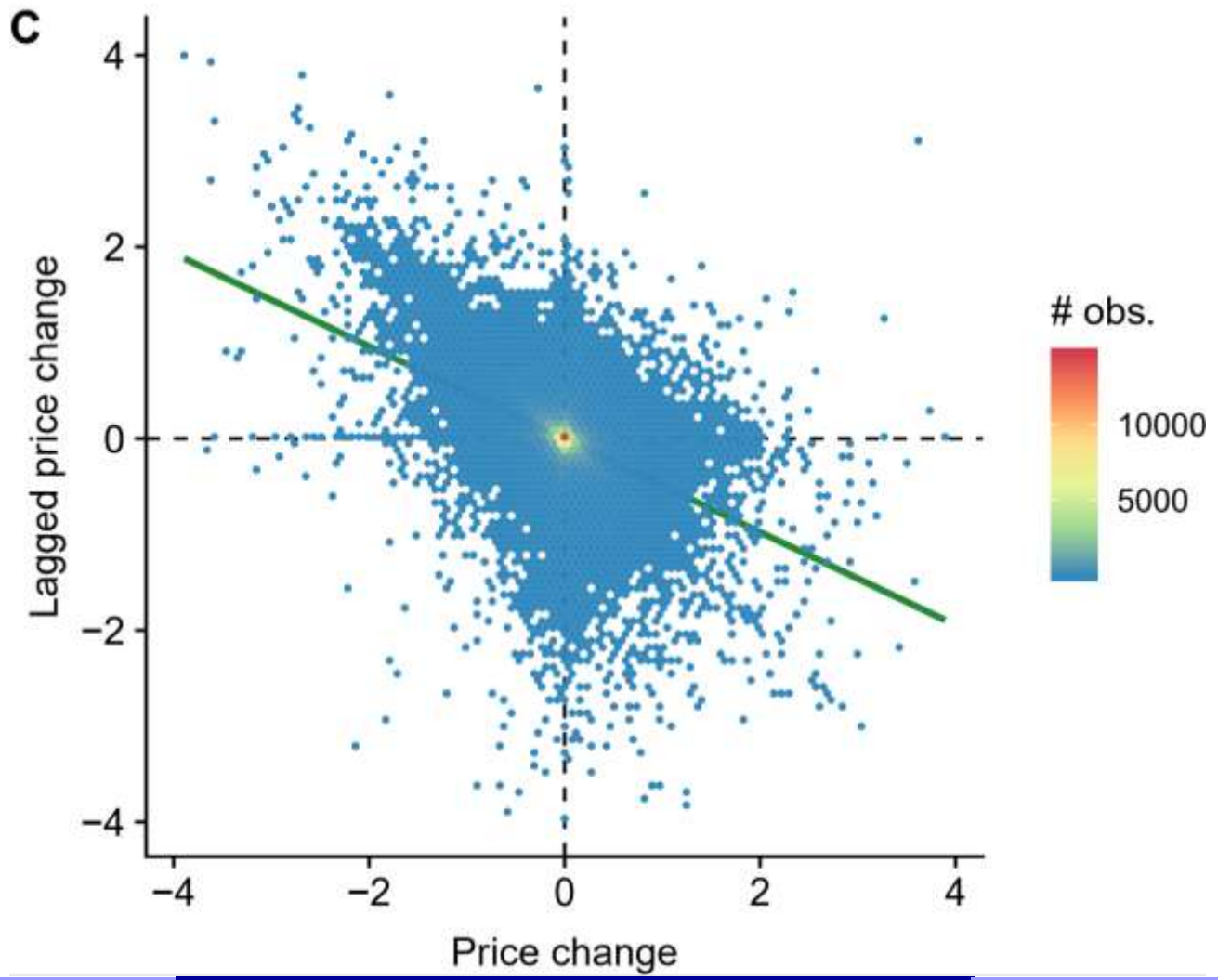
Correlation(Transaction Order, Buyer Value)



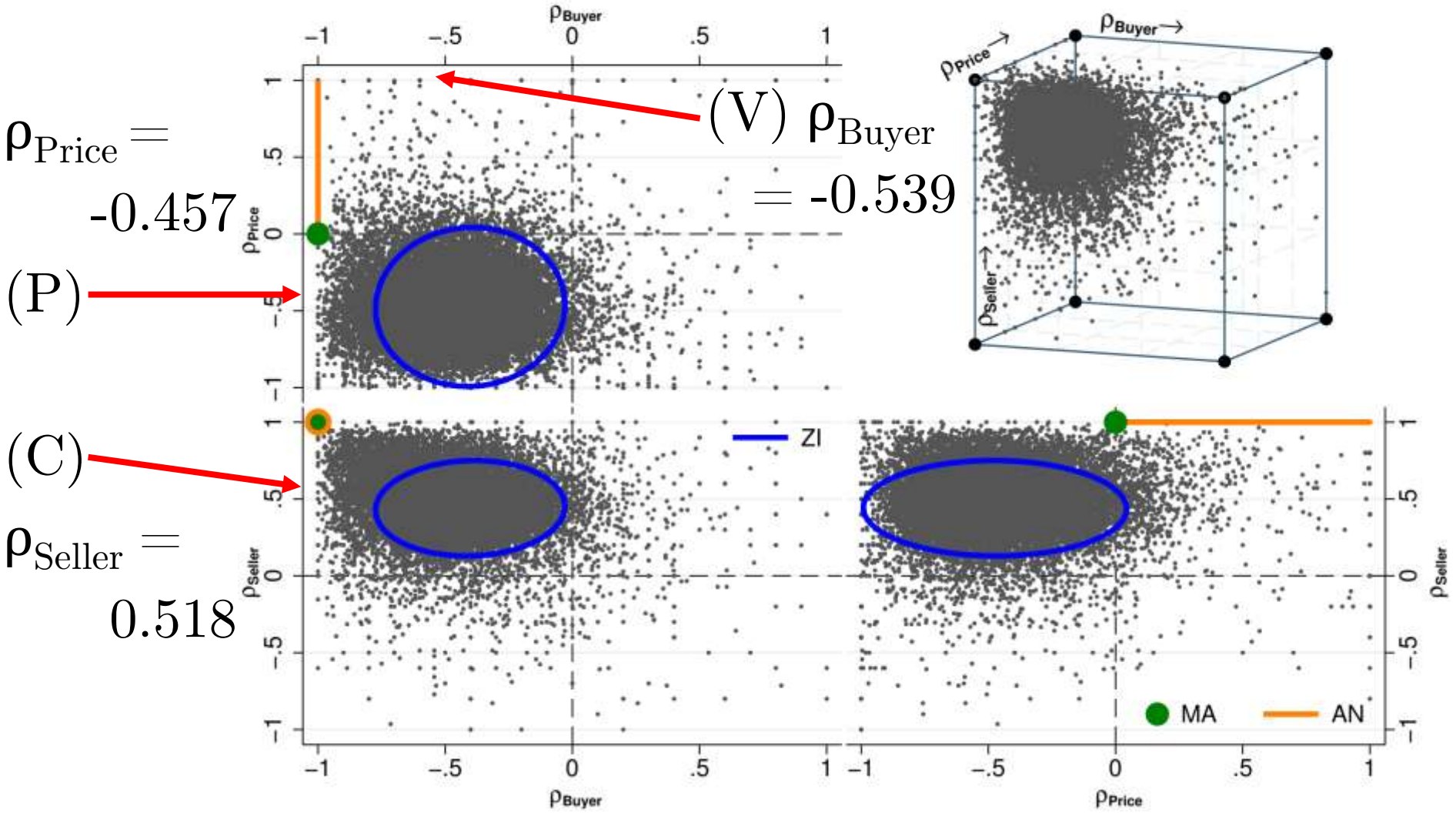
- Higher Value Buyers Trade First

MobLab Double Auction:

Price Change Autocorrelation = -0.457



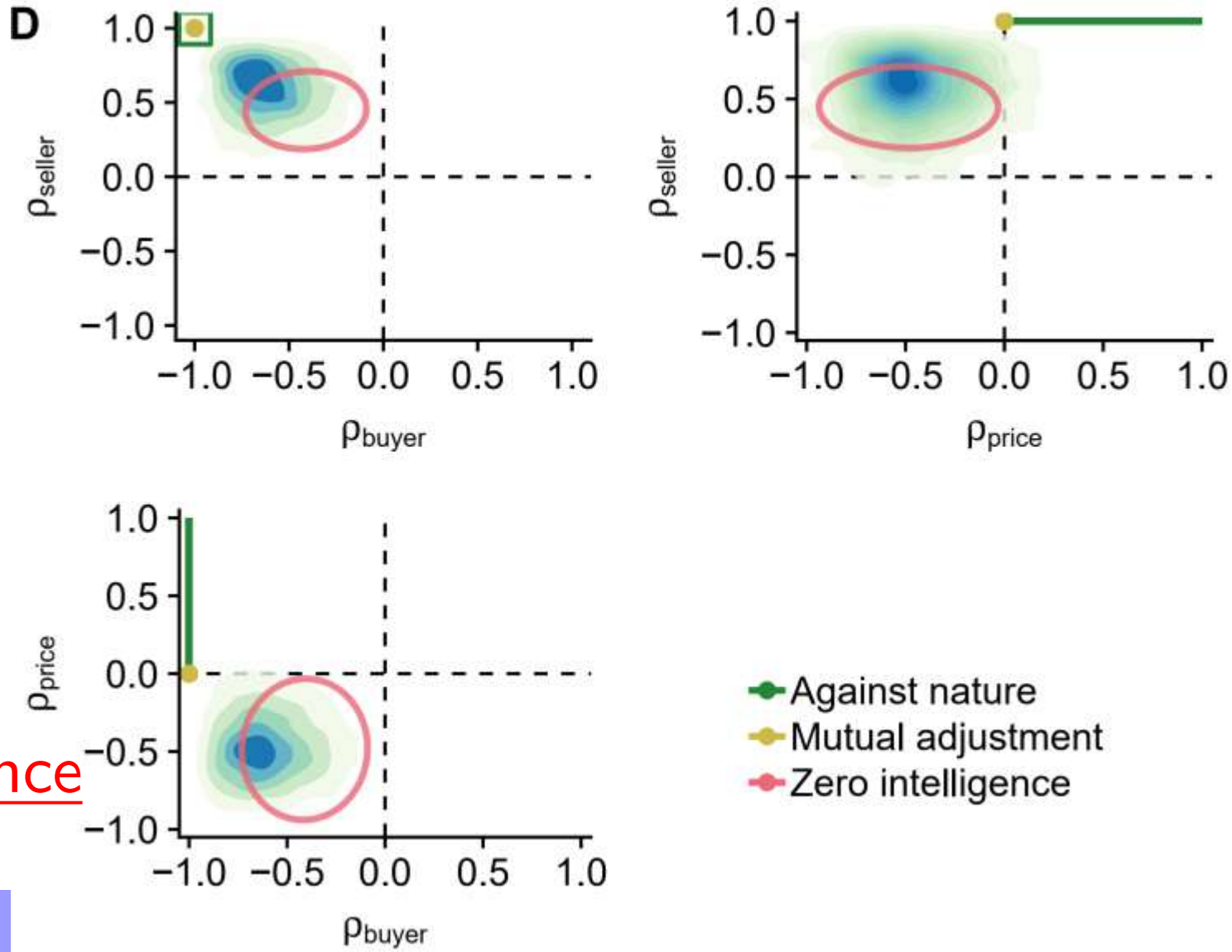
MobLab Double Auction:
Correlation Between Order and P/V/C



MobLab Double Auction:

Testing Theories of Price Formation

- MA:
  Wilson
  (1987)
- AN:
  Friedman
  (1991)
- ZI:
  0-intelligence



Robustness:

Small vs. Large Markets: ZI or Not!!!

(N = 172)

Favor AN/MA!

(traders≤6)

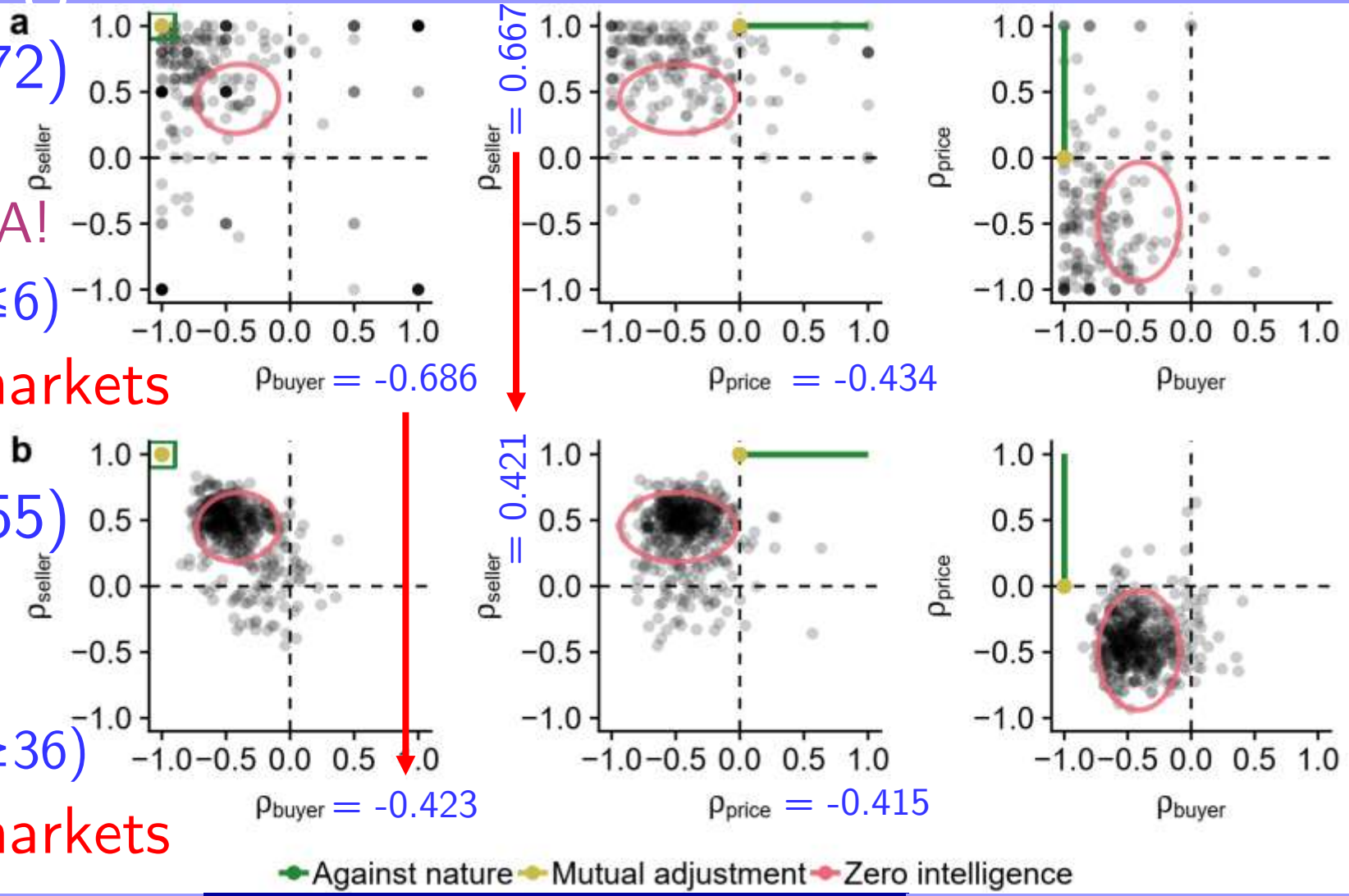
Small markets

(N = 455)

Favor ZI!

(traders≥36)

Large markets



MobLab Ultimatum Game:

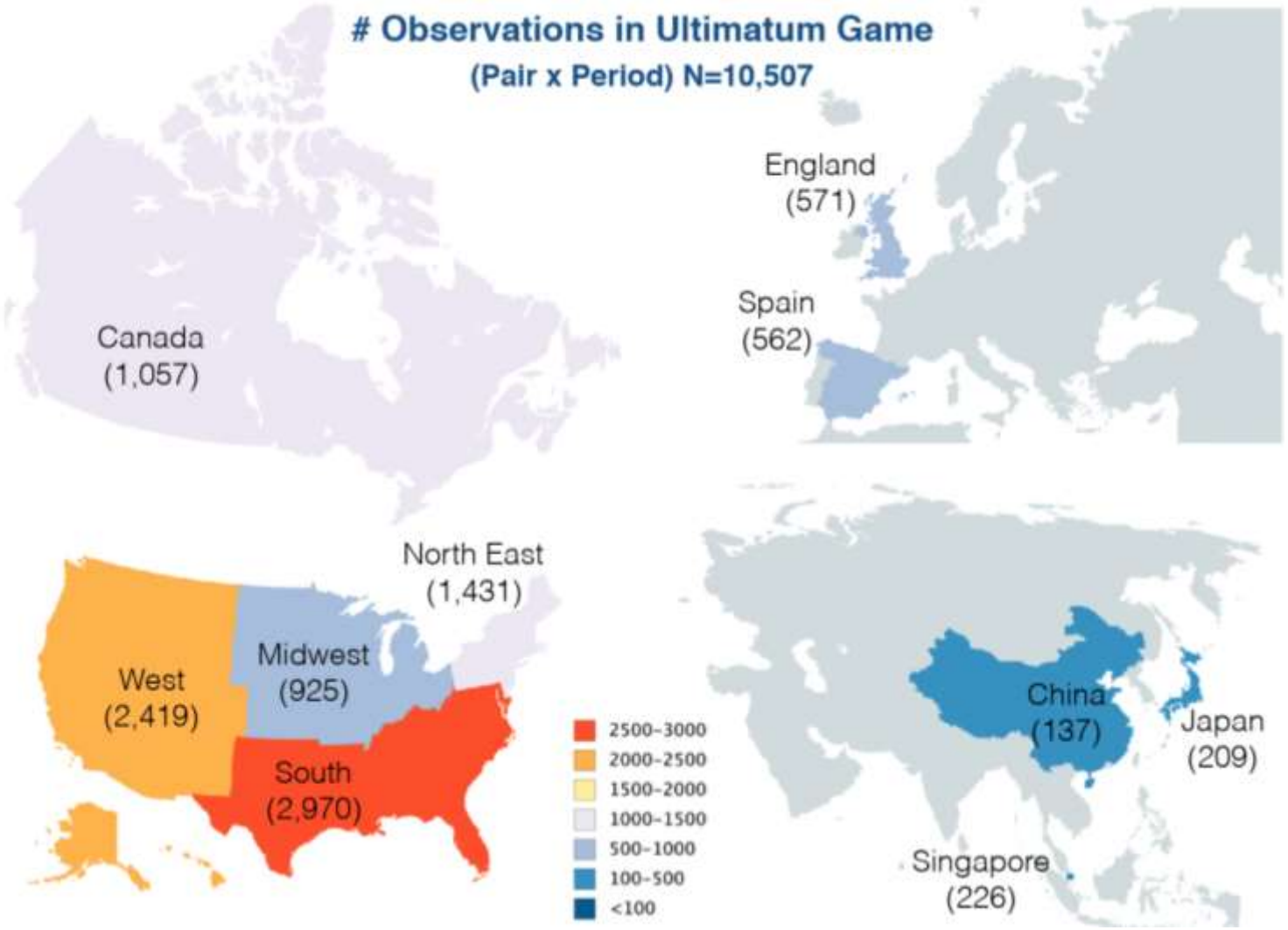
Proposer

MobLab Ultimatum Game:

Respondent

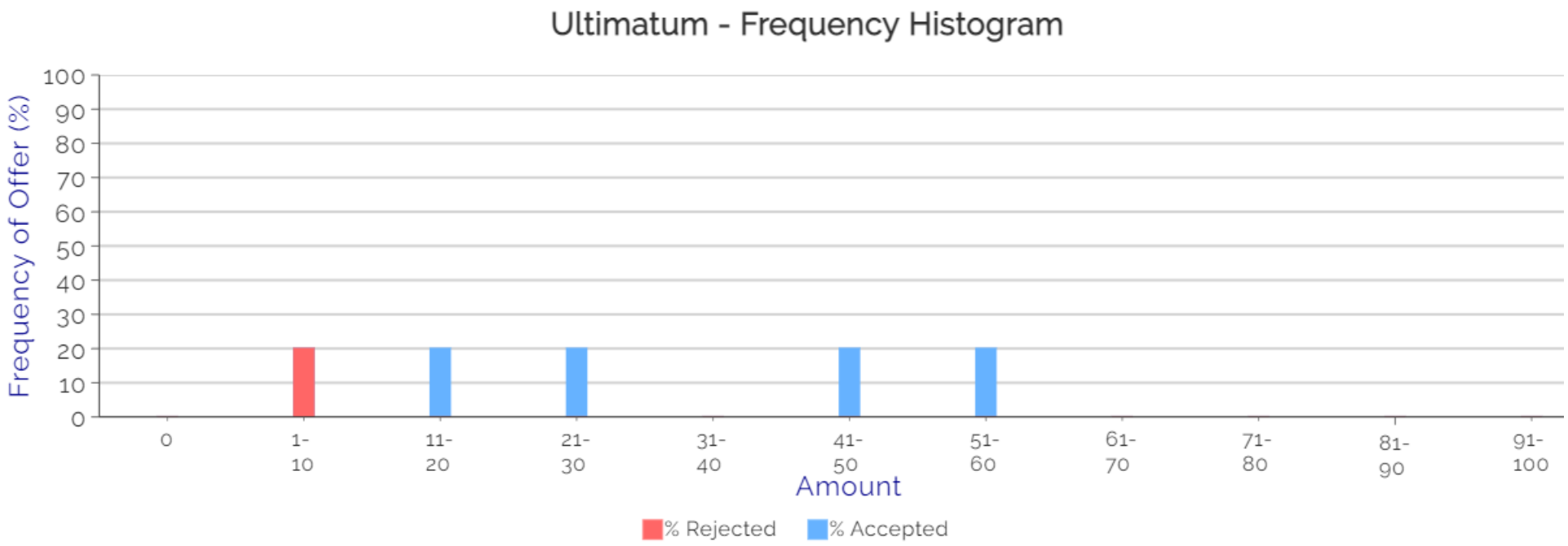
MobLab Ultimatum Game:

Observation in Different Regions/Countries



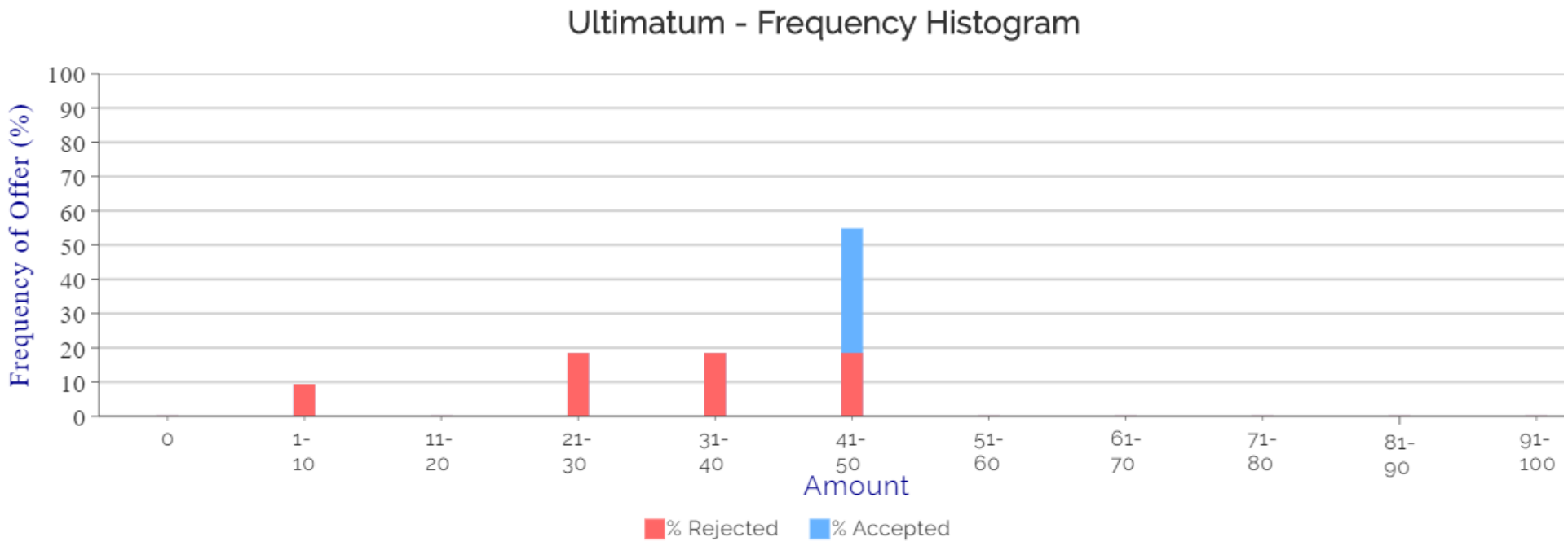
MobLab Ultimatum Game:
EE-BGT 21S Results:

# of Groups	Total Pie	Avg. Offer	Avg. Accepted Offer	Avg. Rejected Offer	Mode Offer	Rejection %
5	100	33.20	39.00	10.00	10	20.00



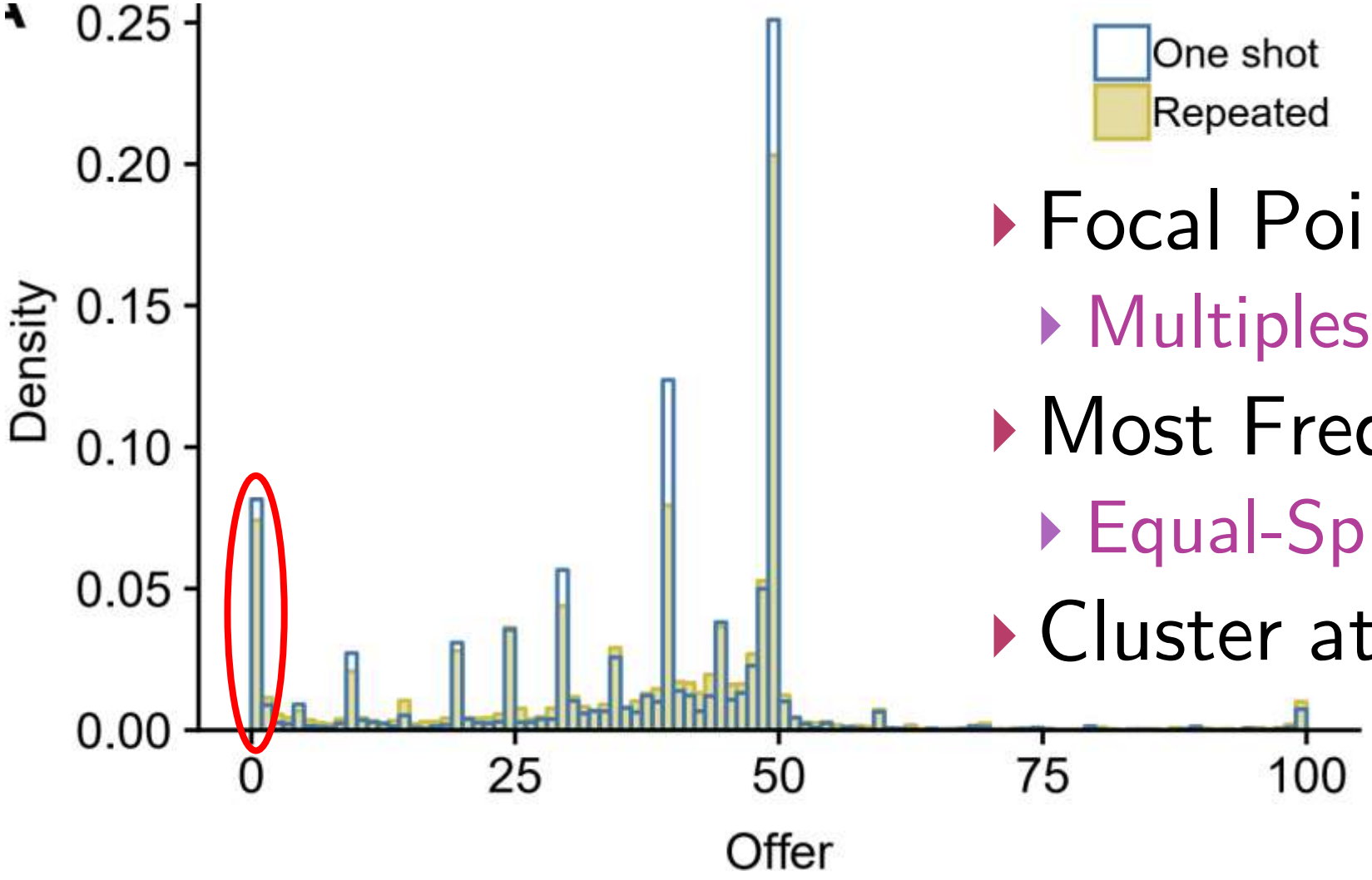
MobLab Ultimatum Game:
CCU Results:

# of Groups	Total Pie	Avg. Offer	Avg. Accepted Offer	Avg. Rejected Offer	Mode Offer	Rejection %
11	100	37.55	49.75	30.57	50	63.64



MobLab Ultimatum Game:

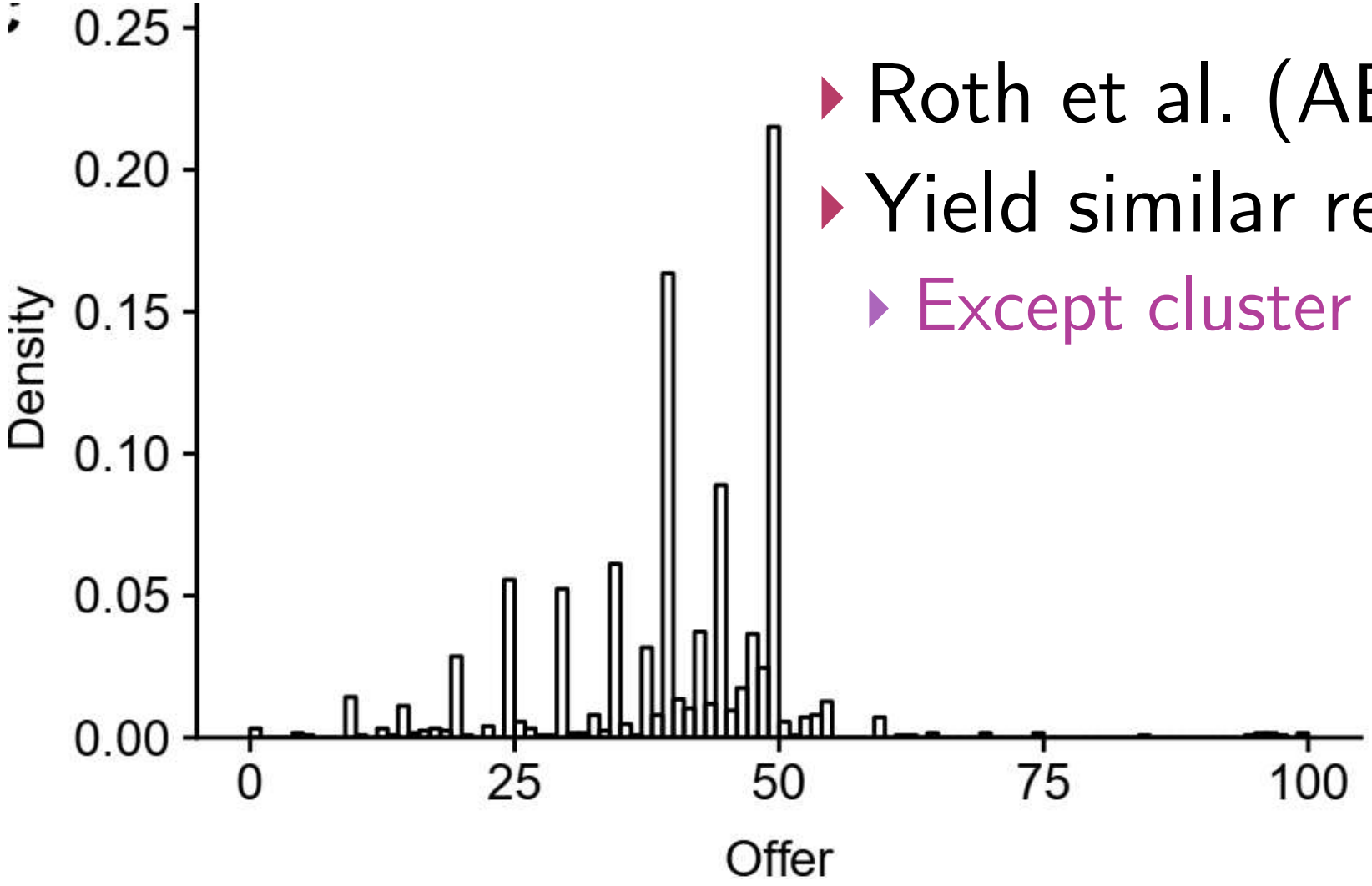
Proposal Offers



- Focal Points:
  - Multiples of Tens
- Most Frequent:
  - Equal-Split
- Cluster at 0

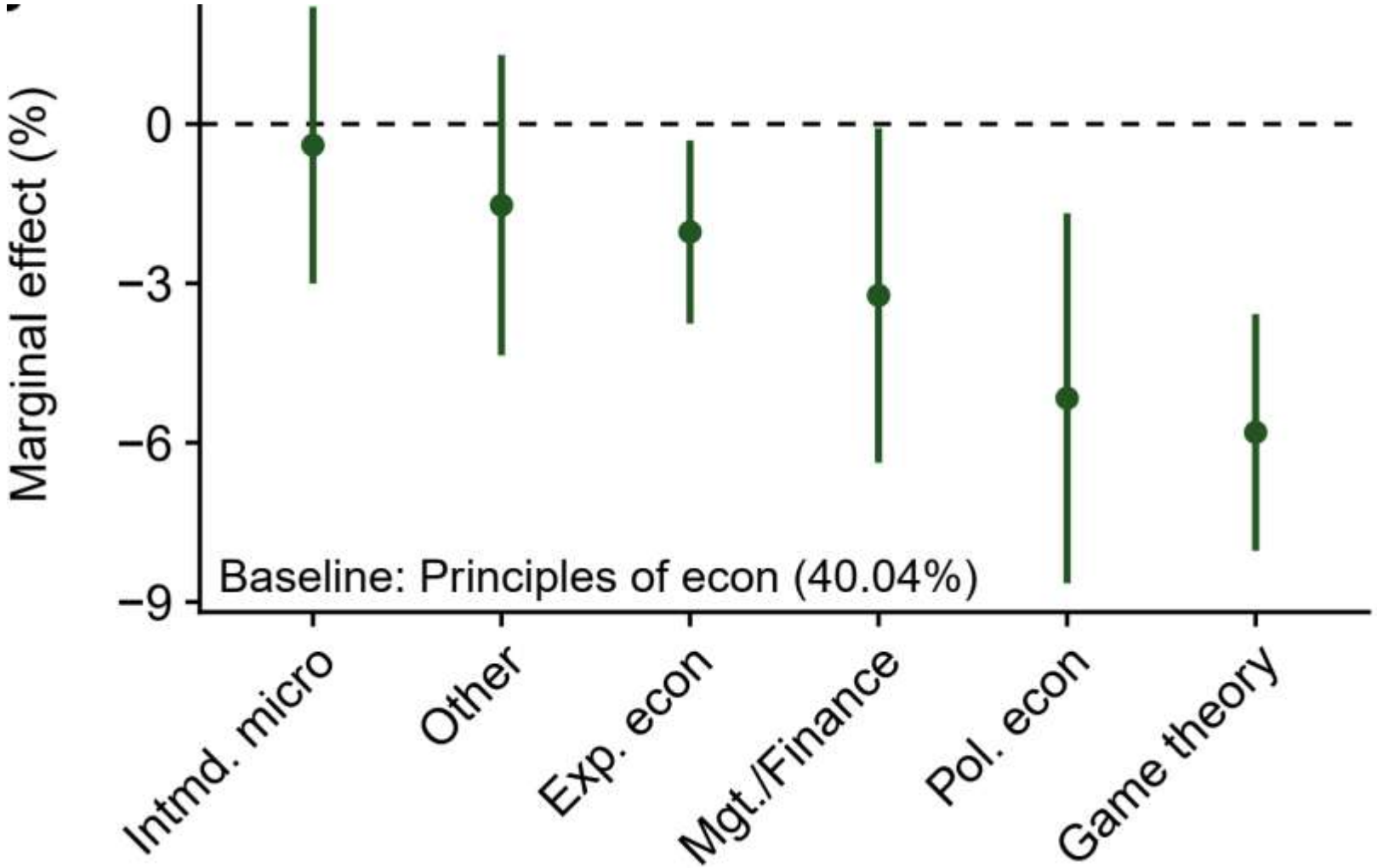
Ultimatum Game in the Lab:

Proposal Offers of Roth et al. (AER 1991)

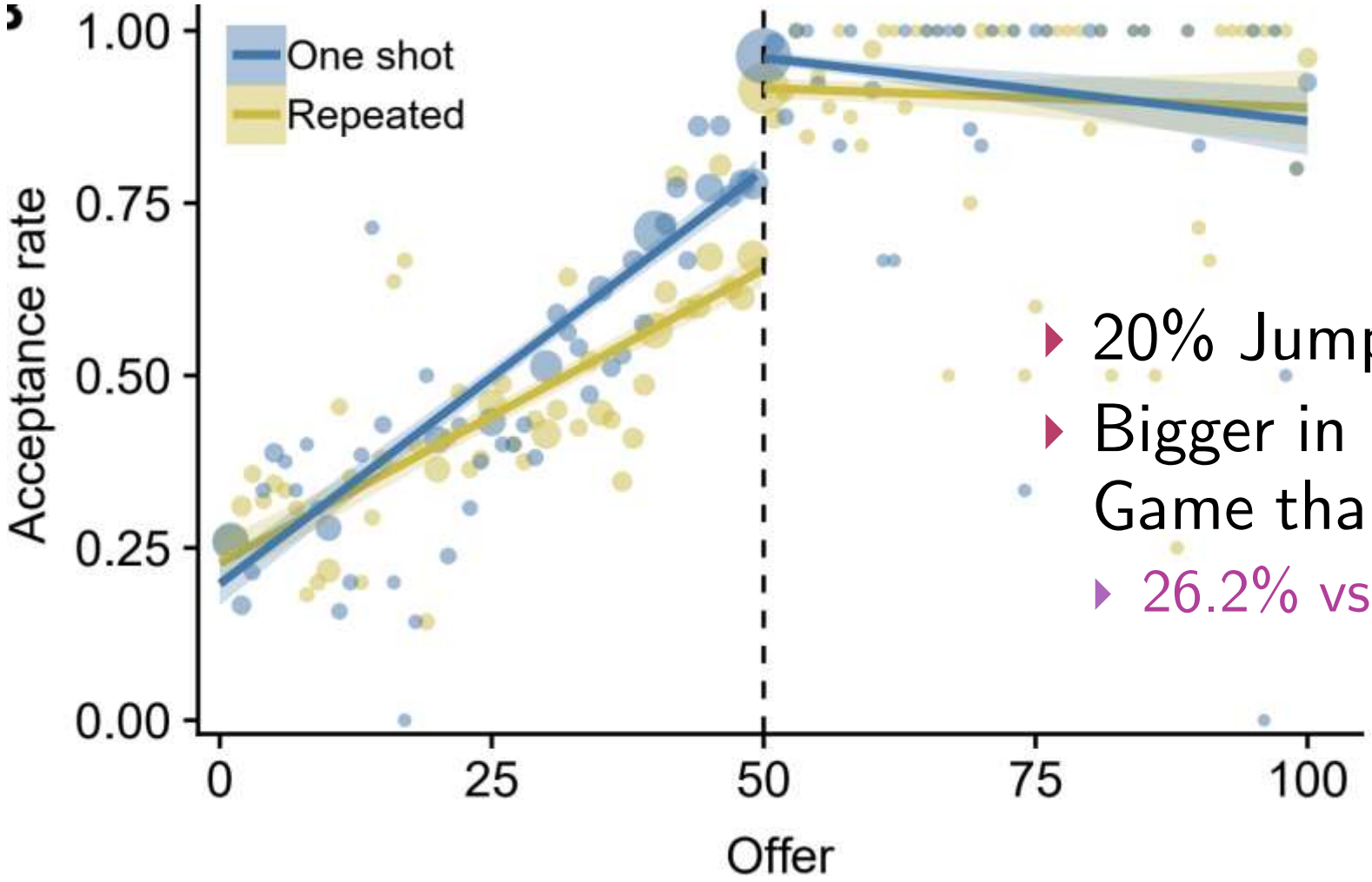


- Roth et al. (AER 1991)
- Yield similar results
  - Except cluster at 0

MobLab Ultimatum Game:
Proposal Offer - Class Effect



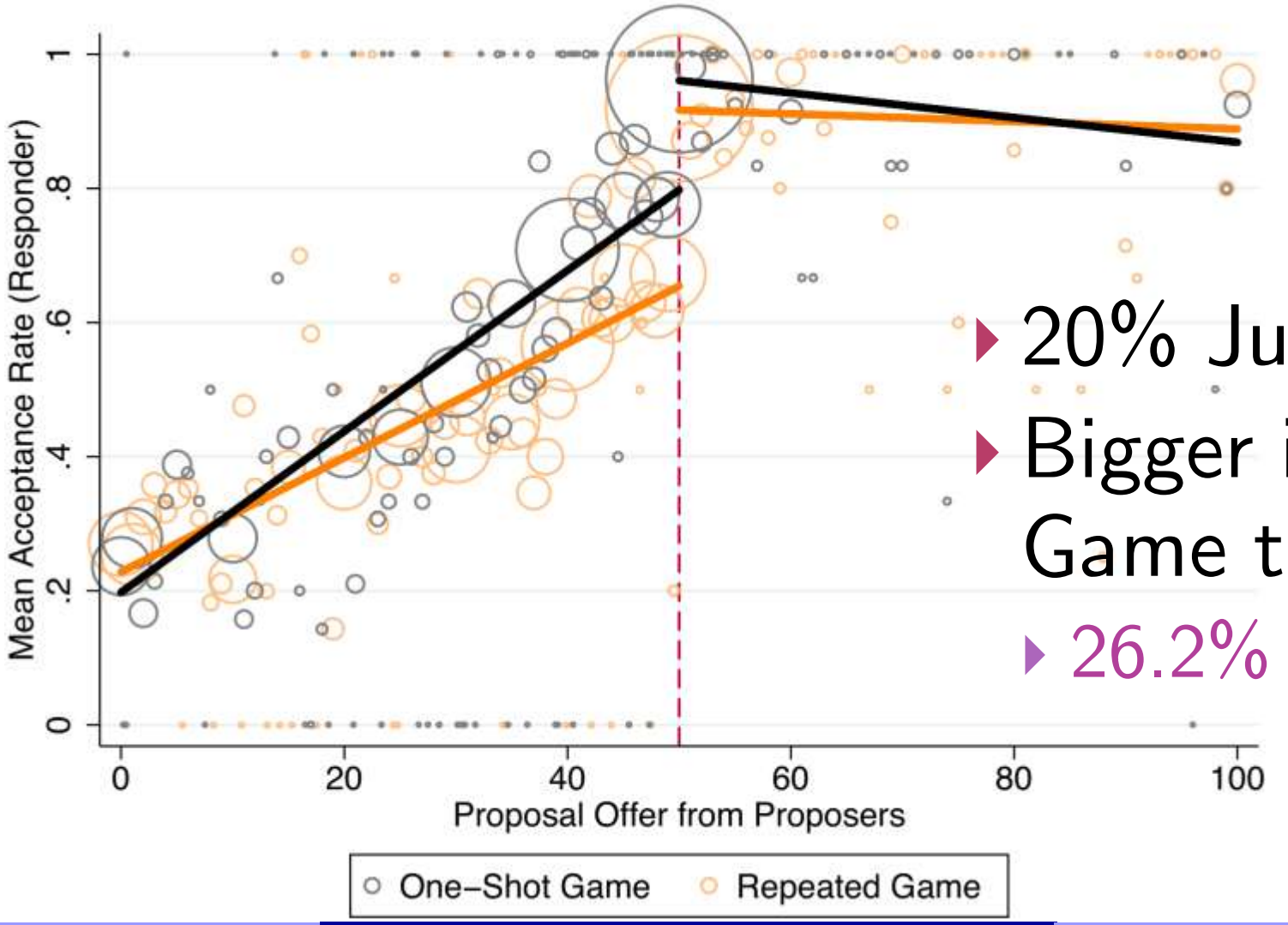
MobLab Ultimatum Game:
Acceptance Rate (Fit 2-part Regression)



- 20% Jump at 50
- Bigger in Repeated Game than 1-shot
  - 26.2% vs. 16.3%

Ultimatum:

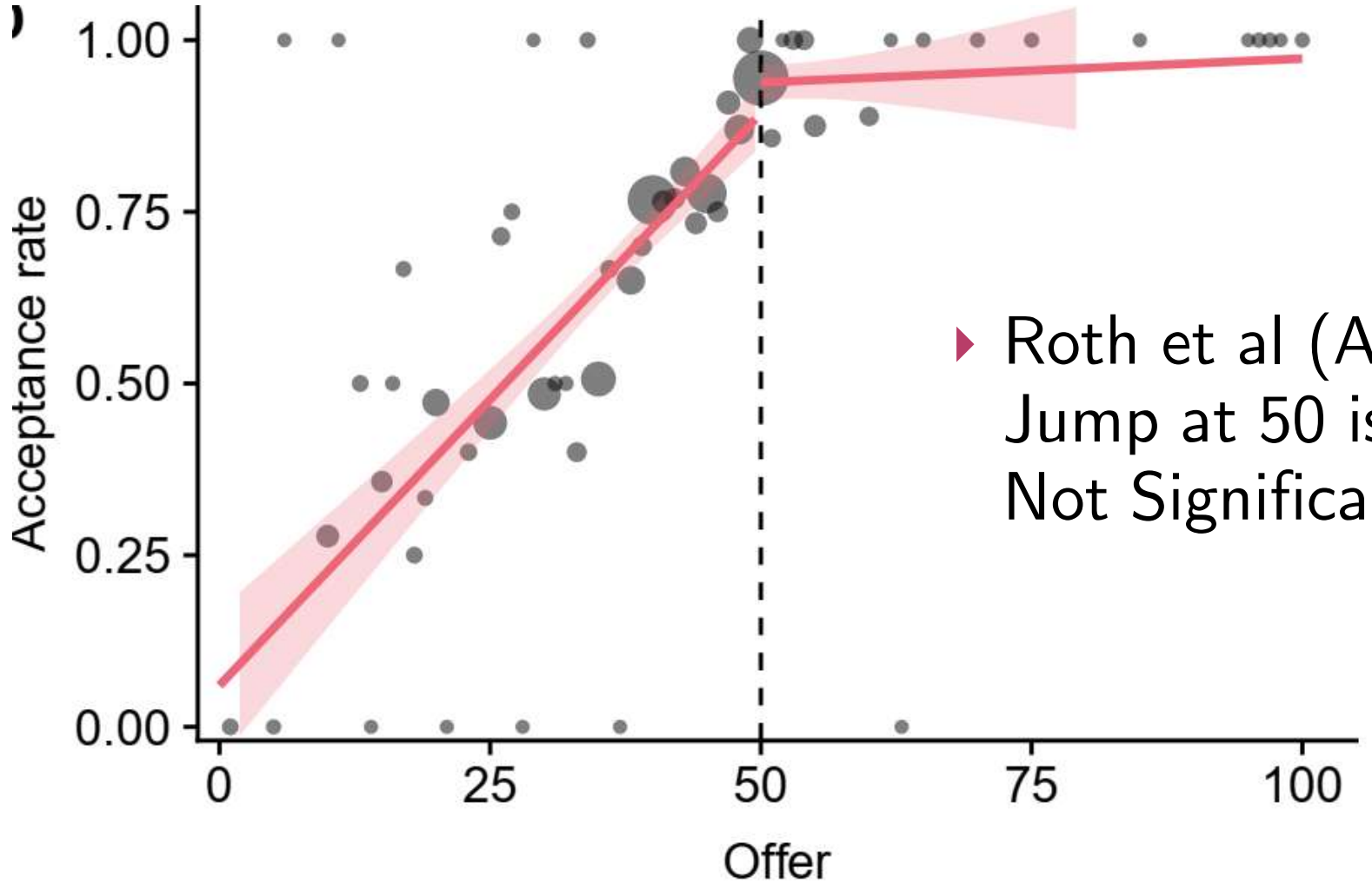
Acceptance Rate (Fit 2-part Regression)



- 20% Jump at 50
- Bigger in Repeated Game than 1-shot
  - 26.2% vs. 16.3%

Ultimatum:

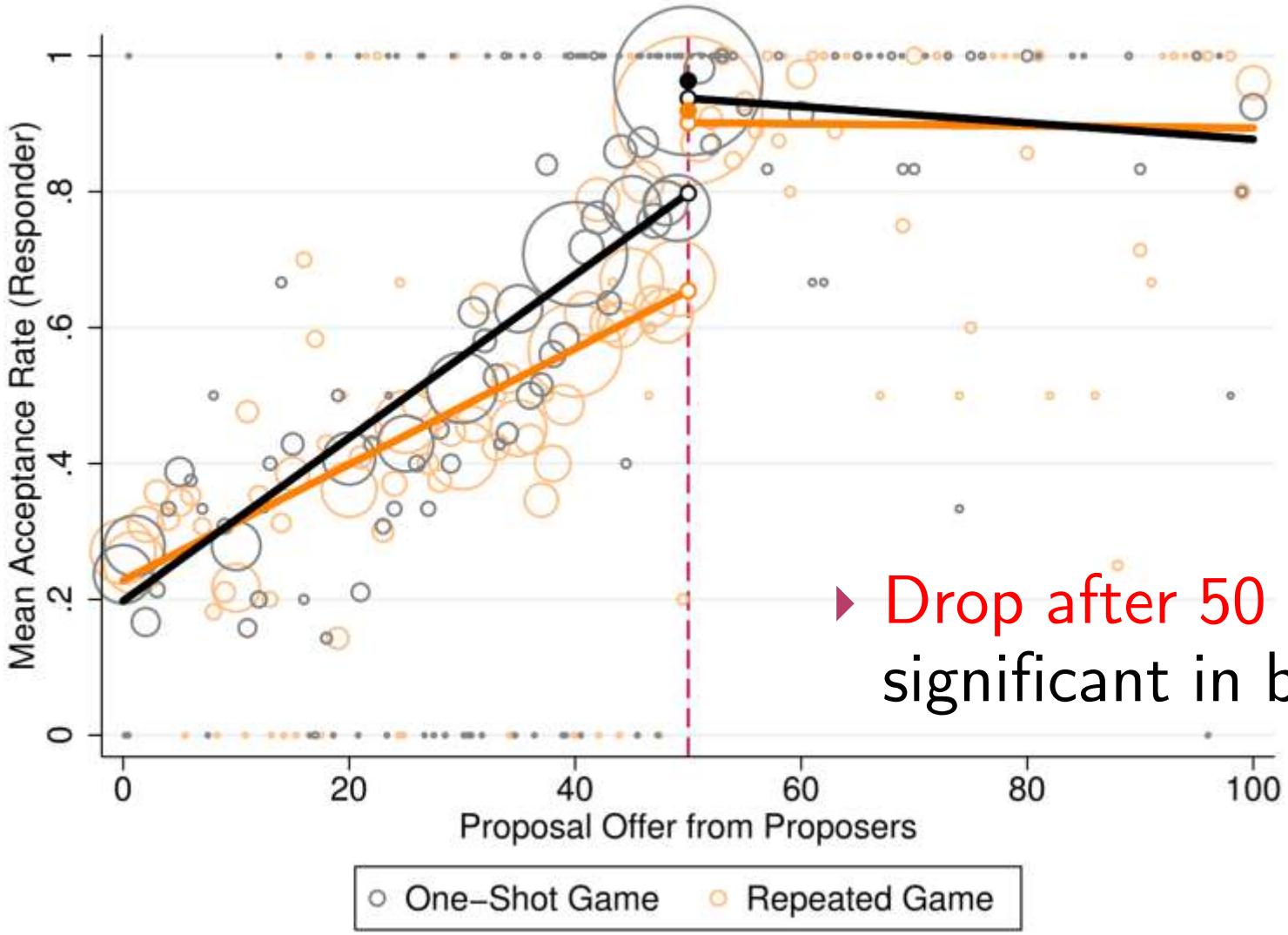
Acceptance Rate of Roth et al. (AER91')



- Roth et al (AER91'): Jump at 50 is Not Significant

Ultimatum:

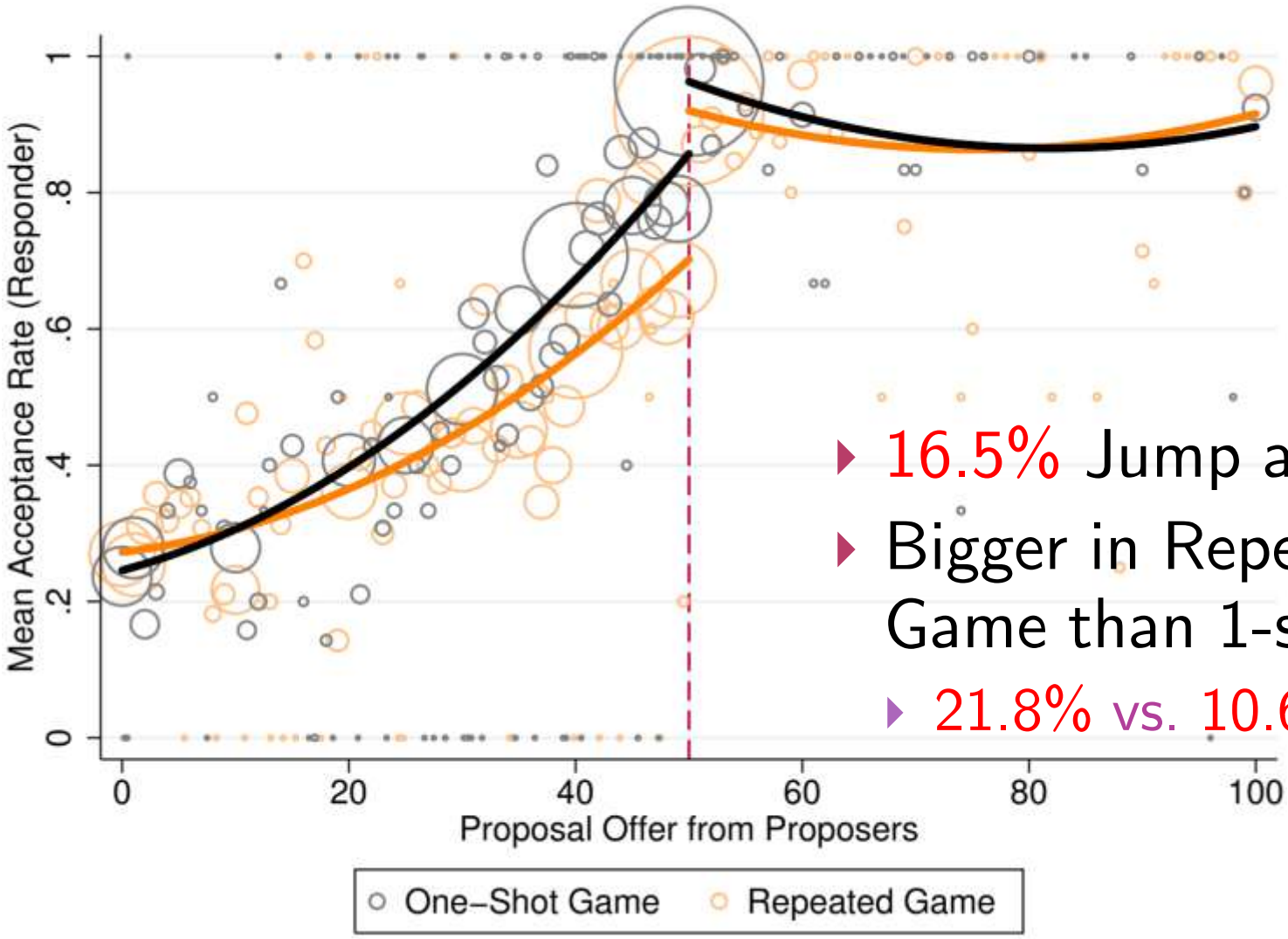
Acceptance Rate (Fit 3-Part Regression)



- Drop after 50 Not significant in both

Ultimatum:

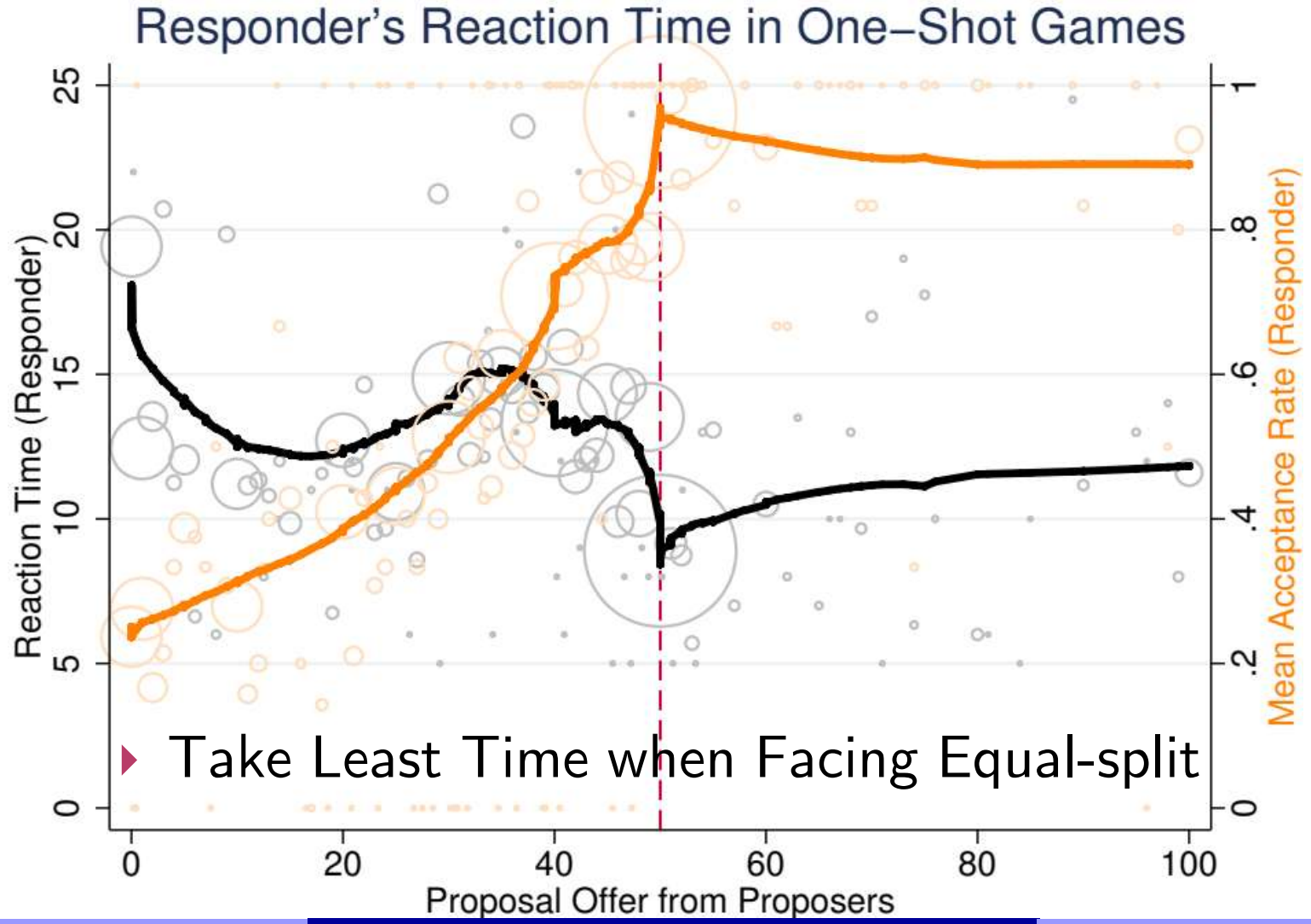
Acceptance Rate (Fit Quadratic)



- 16.5% Jump at 50
- Bigger in Repeated Game than 1-shot
  - 21.8% vs. 10.6%

MobLab Ultimatum Game:

Acceptance Rate and Response Time

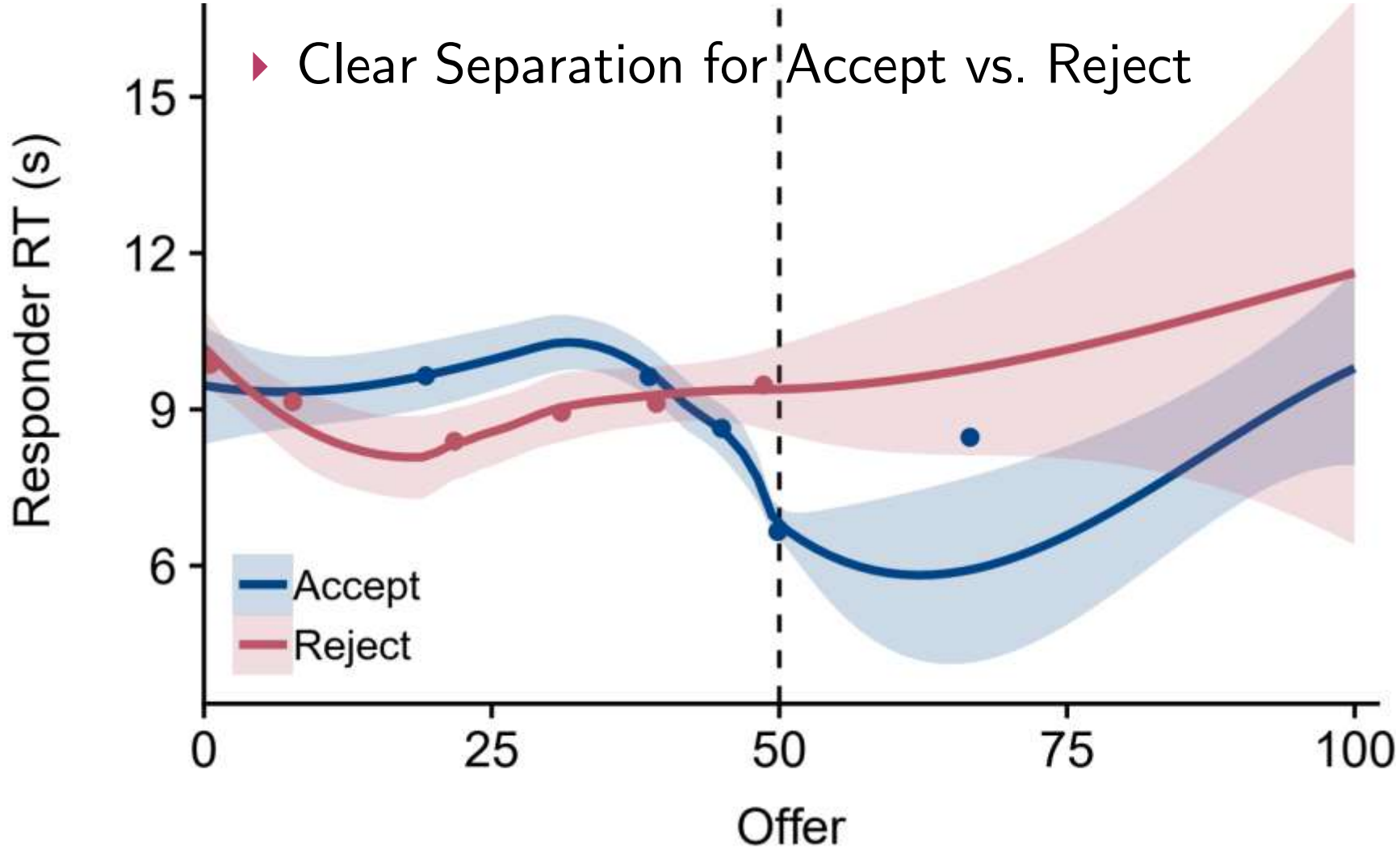


- Take Least Time when Facing Equal-split

MobLab Ultimatum Game:

Response Time

- Clear Separation for Accept vs. Reject



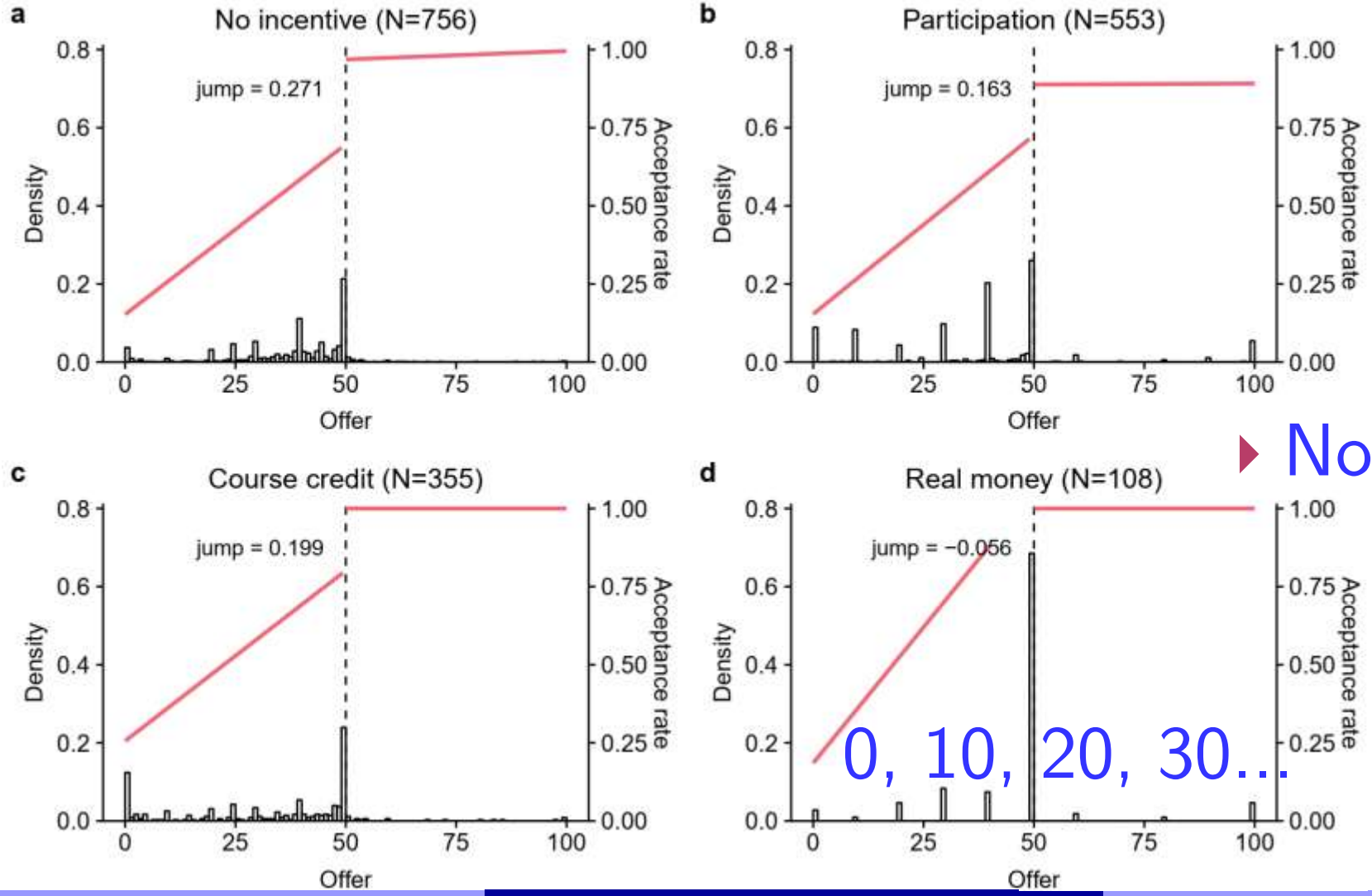
MobLab Ultimatum Game:
Robustness: Do Incentives Matter?

- Find Syllabi online (1,772 out of 10,507 observations)
  - Classify incentives for 58 out of 490 sessions
  1. Course Points (n=355): Performance as grades
  2. Participation (n=553): Participate in enough
  3. No Incentive (n=756): None of the above
- Compared to 1-3, Real Money (n=108) had:
  - Much more 50-50 (More than Double!)
  - Average Proposal 47.22 (>34.00–39.17 of others)
  - Acceptance rate = 91.7% (>61.8–67.3%)

Exp/Beh Econ @US-South SLAC

MobLab Ultimatum Game:

Robustness: Do Incentives Matter?



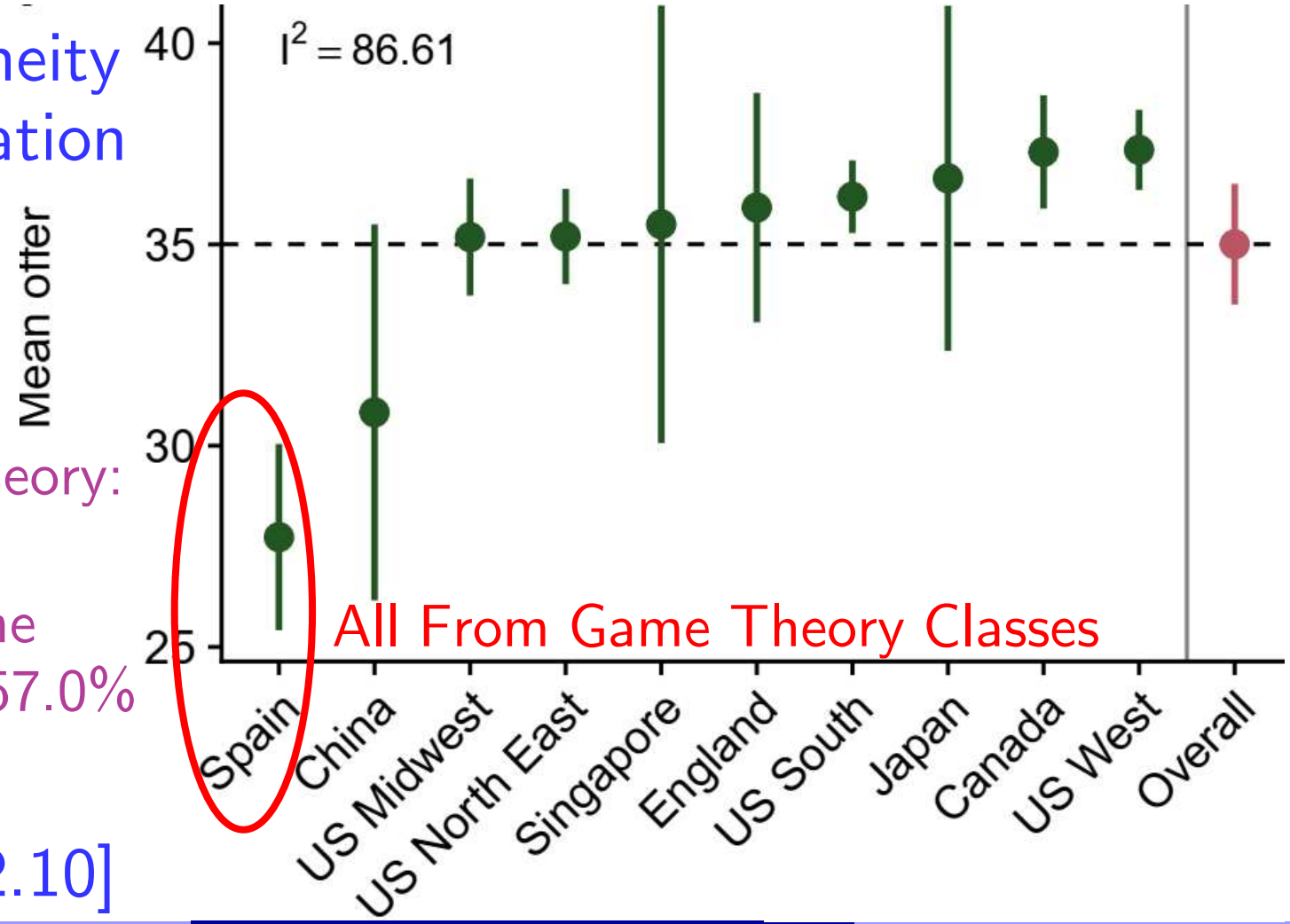
- No 49 vs. 50!

0, 10, 20, 30...

Ultimatum:

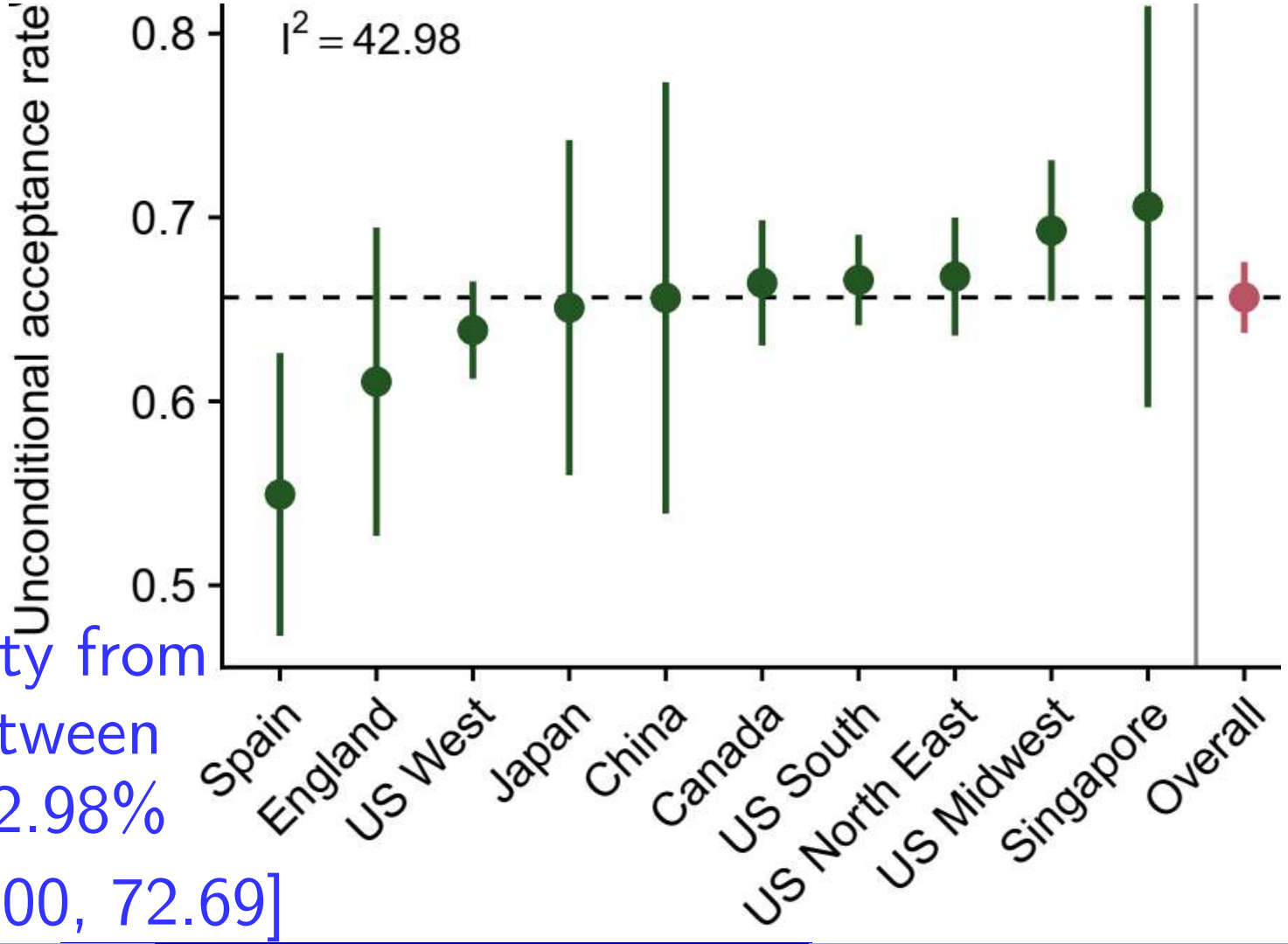
Heterogeneity - Proposal Offers

- Heterogeneity from variation between regions =
- 86.61%
  - Game Theory: 74.0%
  - Non-Game Theory: 57.0%
- 95% CI:
- [77.31, 92.10]



Ultimatum:

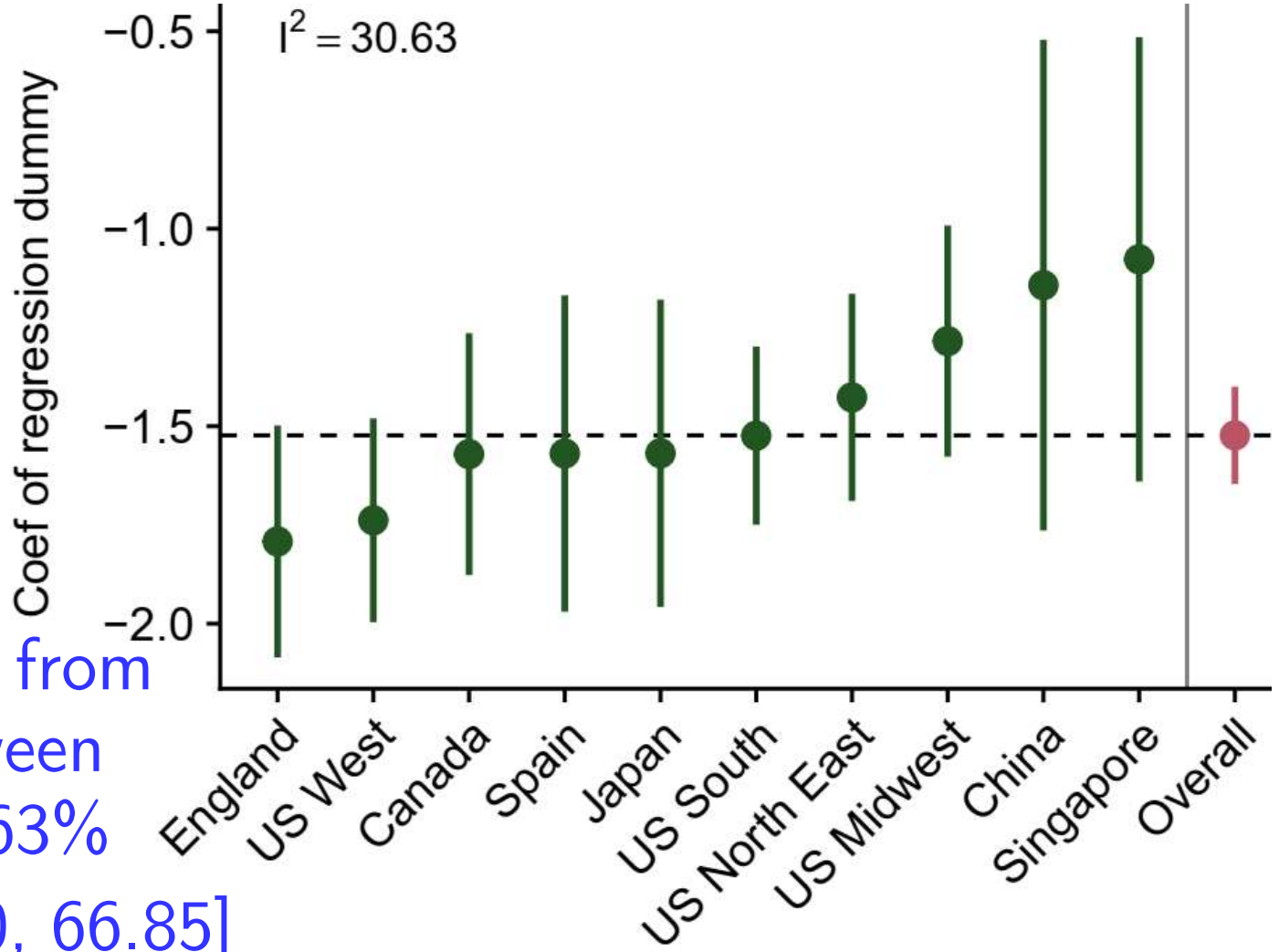
Heterogeneity – Uncond. Acceptance



- Heterogeneity from variation between regions = 42.98%
- 95% CI: [0.00, 72.69]

Ultimatum:

Heterogeneity – Cond. Acceptance

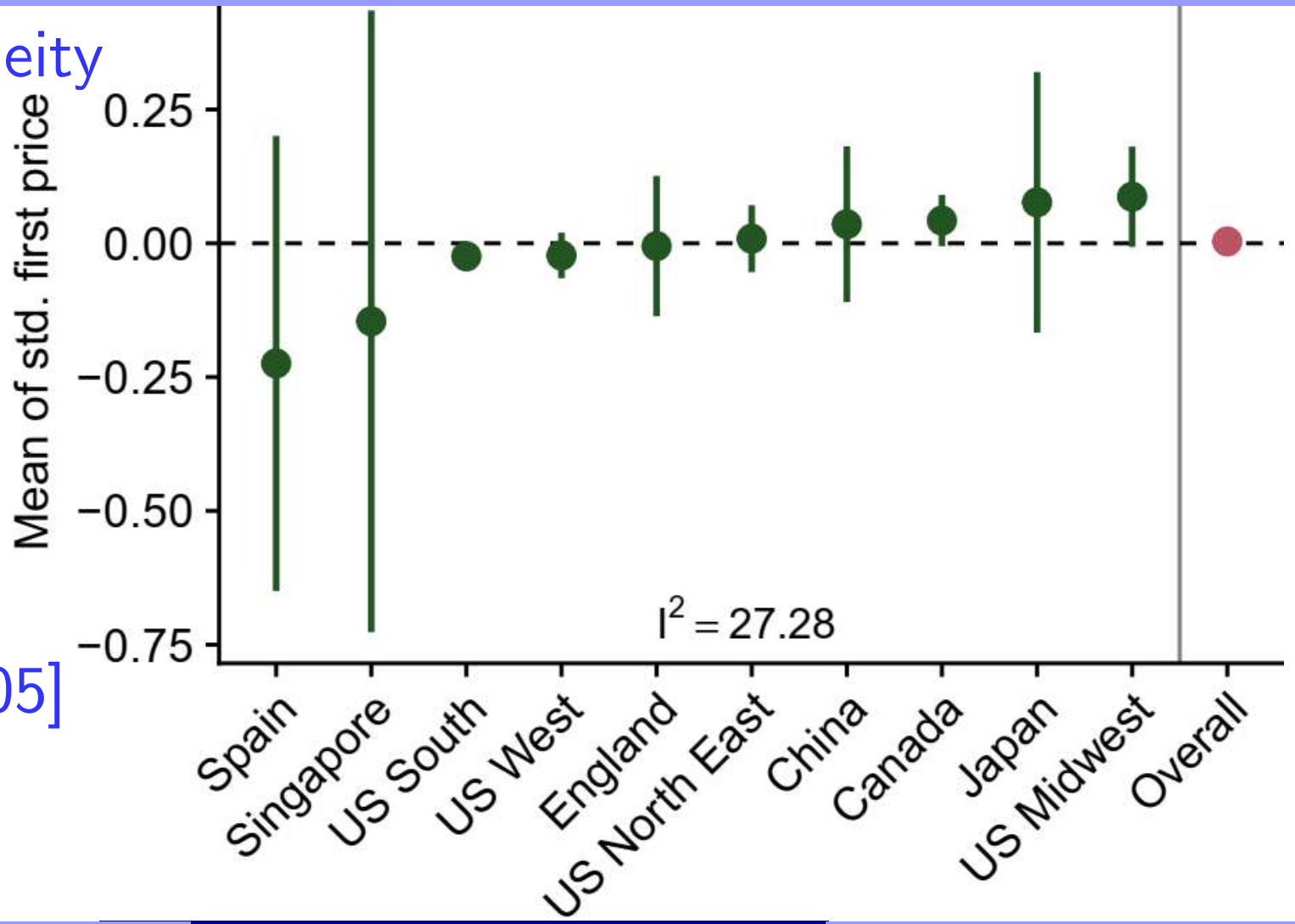


- Heterogeneity from variation between regions = 30.63%
- 95% CI: [0.00, 66.85]

Double Auction:

Heterogeneity — First Price

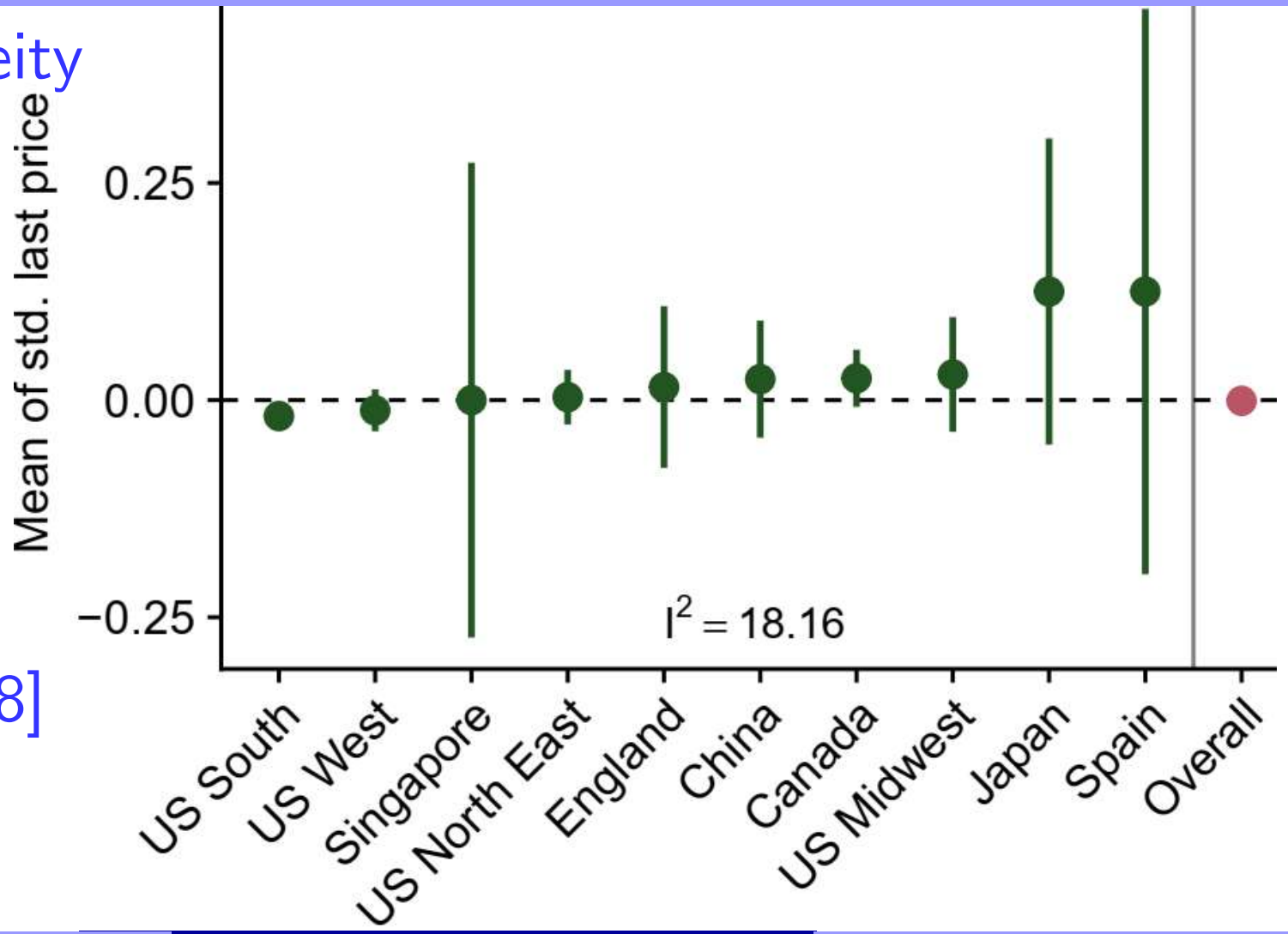
- Heterogeneity from variation between regions = 27.28%
- 95% CI: [0.00, 65.05]



Double Auction:

Heterogeneity — Last Price

- Heterogeneity from variation between regions = 18.16%
- 95% CI: [0.00, 59.08]



Conclusion

- MobLab Data from 2000 Classroom Experiments
  - Standard Design, Many Places, but Obscure Incentives
- Ultimatum Game:
  - Extremely high frequency at Equal-Split
  - Offer at 10s (Natural Focal Points)
- At 50-50:
  - Acceptance jumps 20% (to 94%)
  - Response time on average 8.9 seconds (shortest)

Conclusion

- Double Auction Markets:
  - MED, Smith's alpha and Efficiency all converge to CE
  - Negative price change autocorrelation (-0.457)
  - Transactions happen in order of Values/Costs
- Heterogeneity: Use meta-analysis techniques
- Higher between-region variance
  - In proposal offer than acceptance
  - In ultimatum than double auction

Robustness Checks That Matter!!

- UG: Do Incentives Matter?
  - Found 58 Syllabi online (out of 490 sessions)
  - Separate Real Money from
  - No Incentives / Participation / Course Points
  - More 50-50, High Acceptance, Can't see 49 vs. 50
- DA: Does # of Traders Matter?
  - Similar $\triangle P$ Auto-Corr.
  - Transaction Order closer to ZI in Large markets
  - Transaction Order closer to MA/AN in small ones

Thanks for Your Attention!

$\rho_{\text{buyer}}^{1\ 3}$

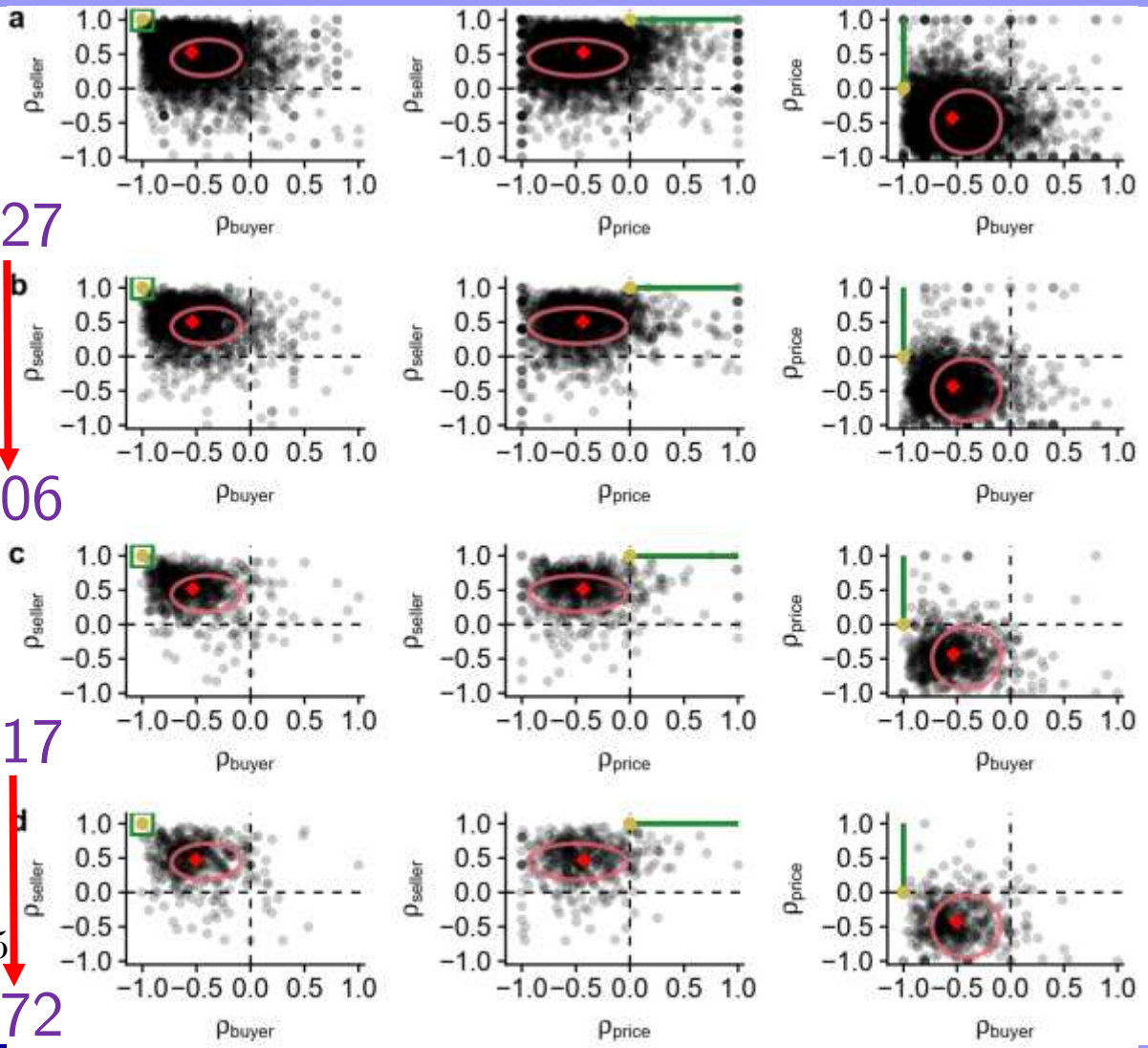
Robustness Checks That Don't Matter

- DA: Experience (Little Learning Across Periods)
  - $\triangle$P Auto-Corr., Transaction Order, # of Trades
- DA: Loss Trades (No Effect)
  - $\triangle$P Auto-Corr., B/S Rank-Order Corr.
- DA: Accepted Bids/Asks (No Effect)
  - $\triangle$P Auto-Corr., B/S Rank-Order Corr.
- UG: Regional Difference?
  - Little Difference in Proposal Offers, Acceptance Rates,...

Robustness: Some Learning?

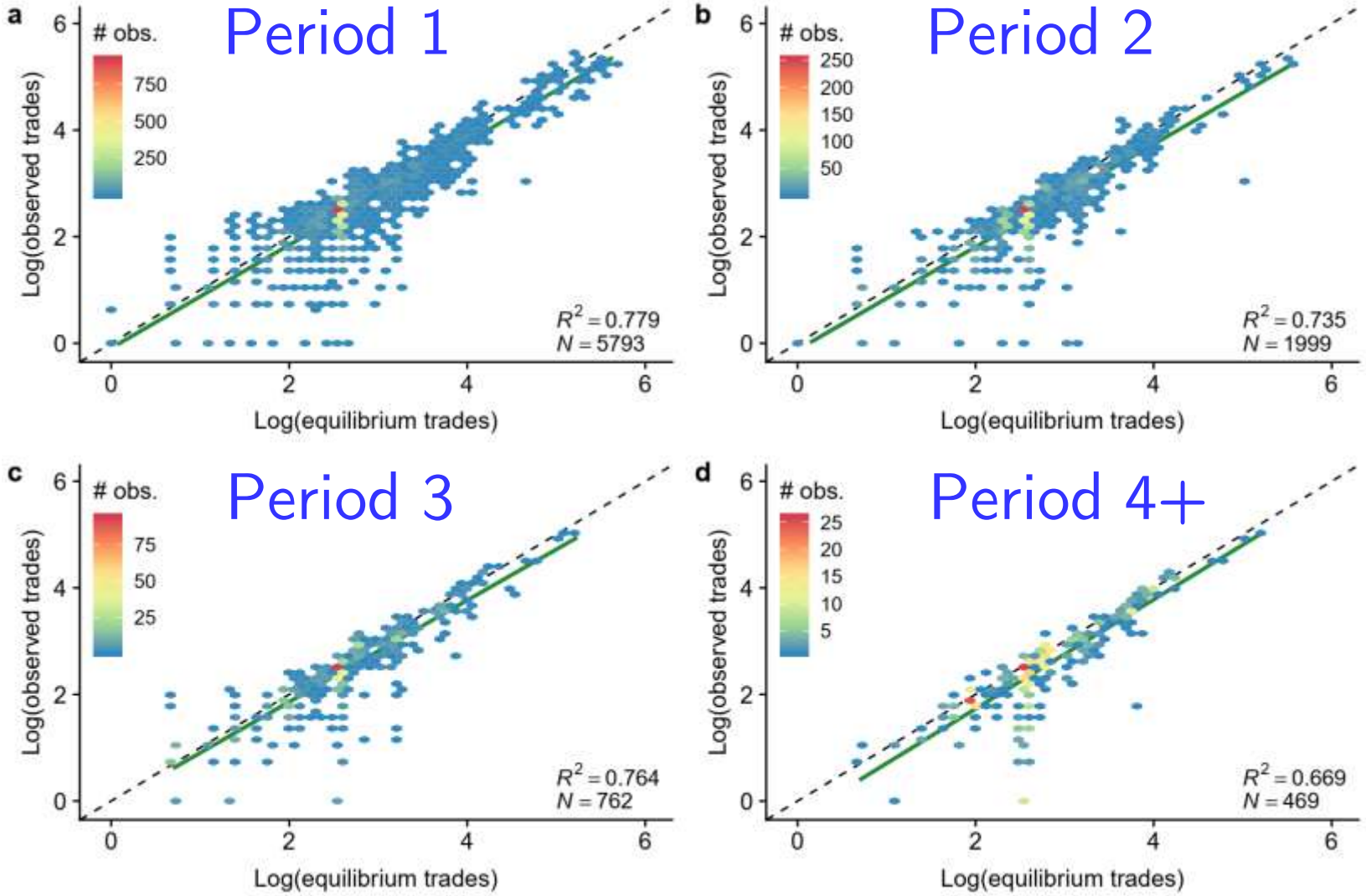
Learning Across Periods? Somewhat?!

- Period 1 (N = 5,498)
  - $\rho_{\text{Buyer}} = -0.543$
    - $\rho_{\text{Seller}} = 0.527$
- Period 2 (N = 1,868)
  - $\rho_{\text{Buyer}} = -0.536$ $p = 0.006$
    - $\rho_{\text{Seller}} = 0.506$
- Period 3 (N = 686)
  - $\rho_{\text{Buyer}} = -0.534$
    - $\rho_{\text{Seller}} = 0.517$
- Period 4+ (N = 440)
  - $\rho_{\text{Buyer}} = -0.506$ $p = 0.049$ $p = 0.015$
    - $\rho_{\text{Seller}} = 0.472$



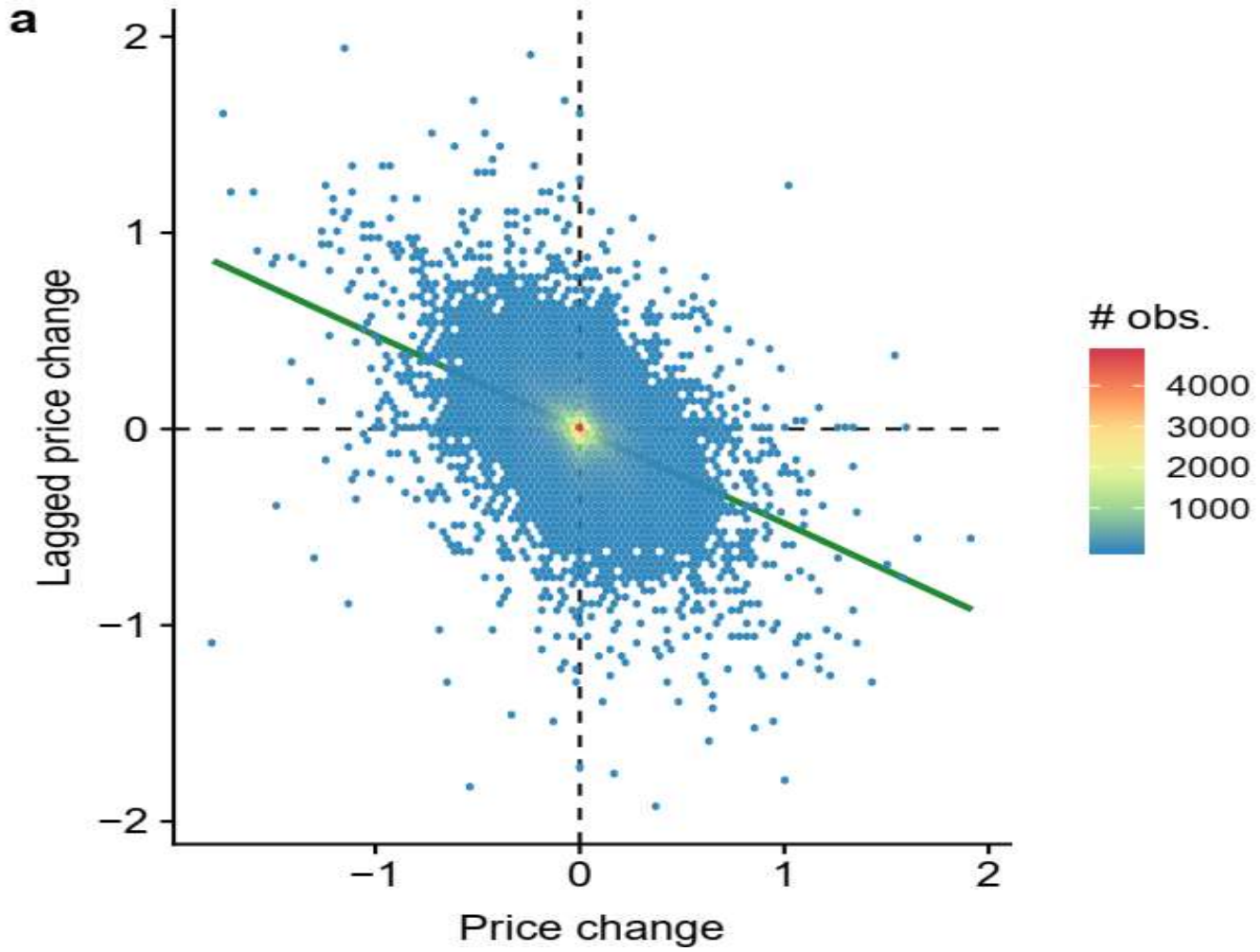
Robustness: Little Learning!

Persistent Under-Trade



Drop All Loss Trades? No Effect!

Price Change Autocorrelation = -0.427

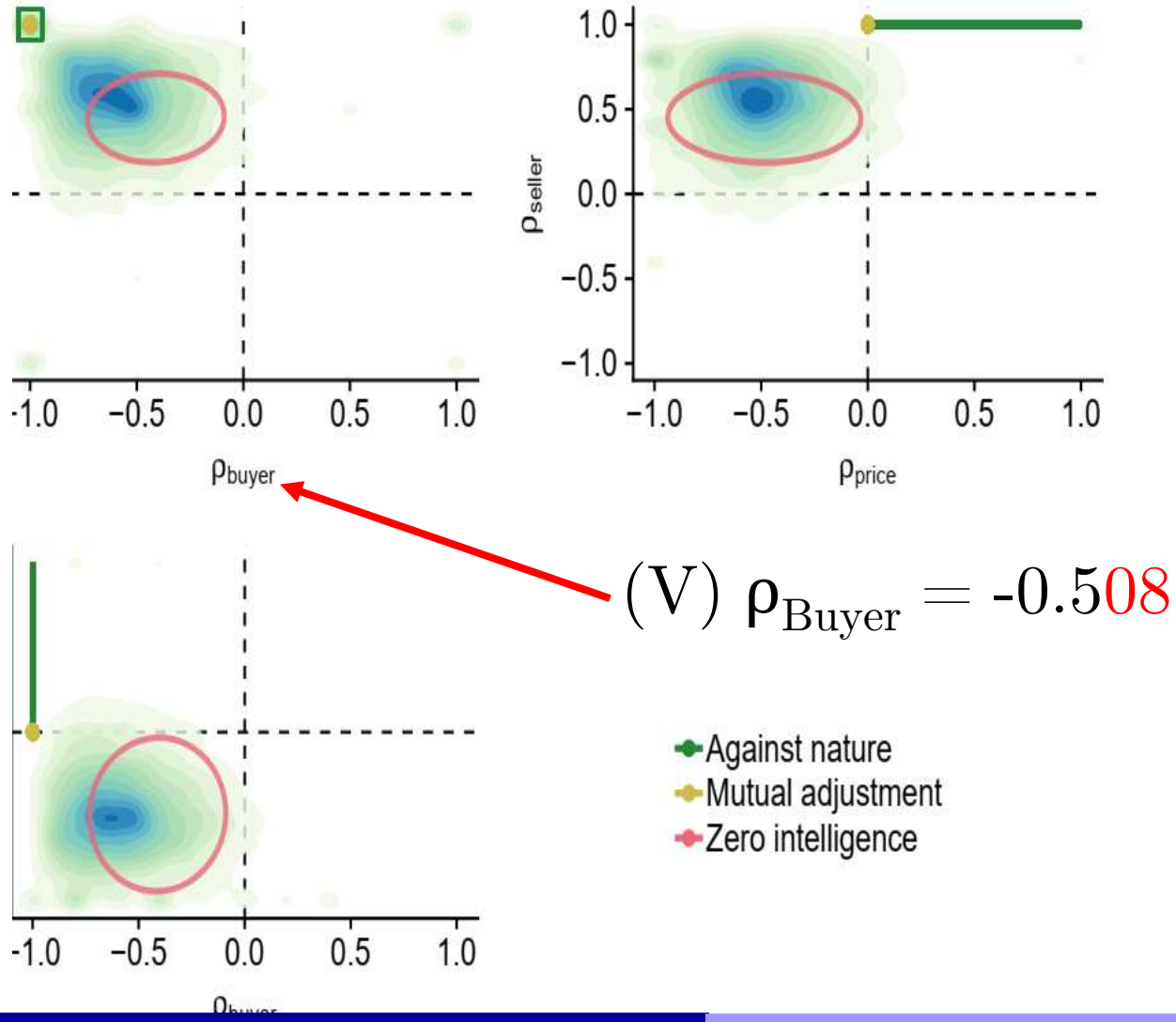


Robustness:

Drop All Loss Trades? No Effect!

$\rho_{\text{Price}} =$ -0.435 (P)

(C) $\rho_{\text{Seller}} =$ 0.570

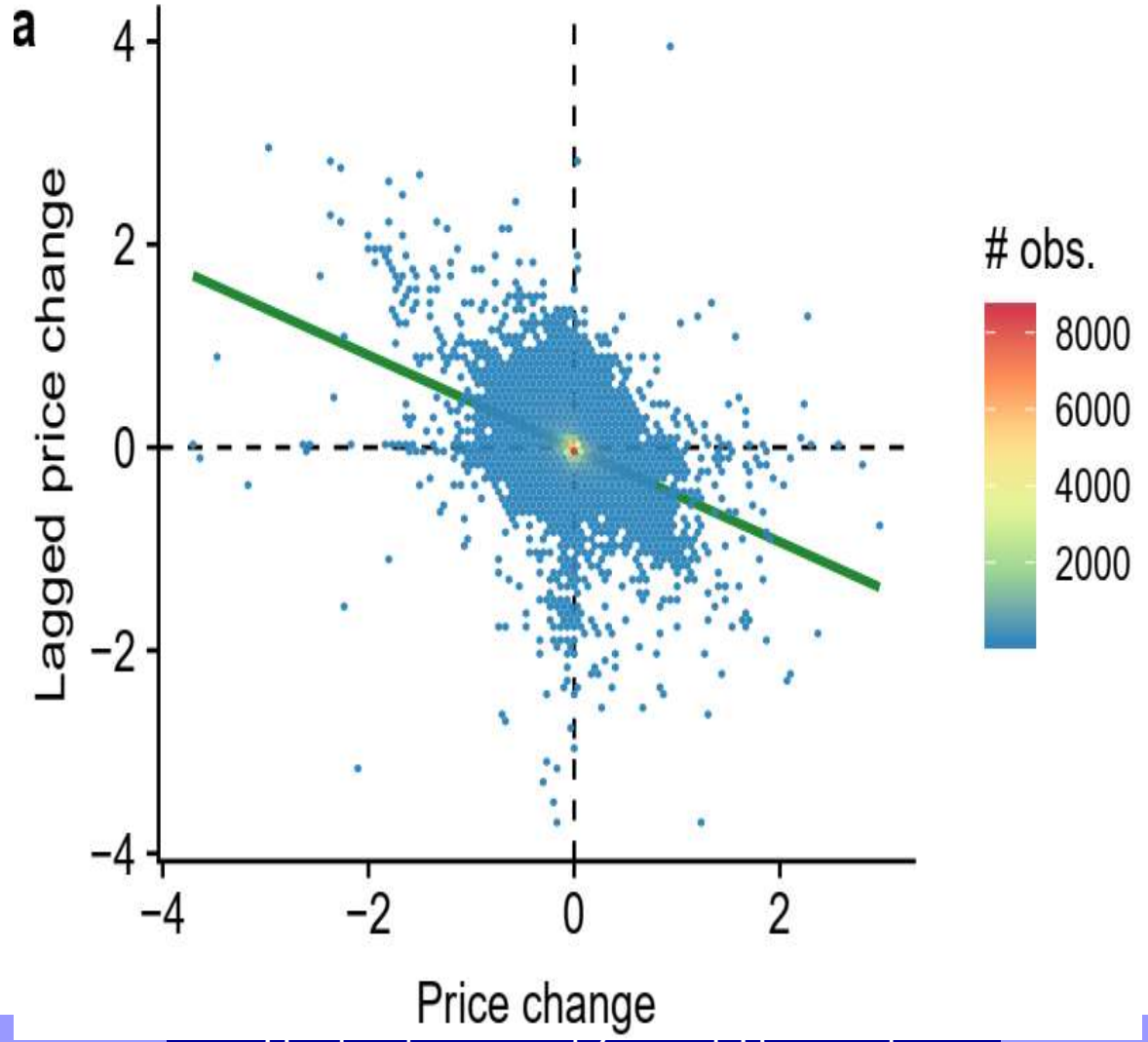


(V) $\rho_{\text{Buyer}} =$ -0.508

Against nature
Mutual adjustment
Zero intelligence

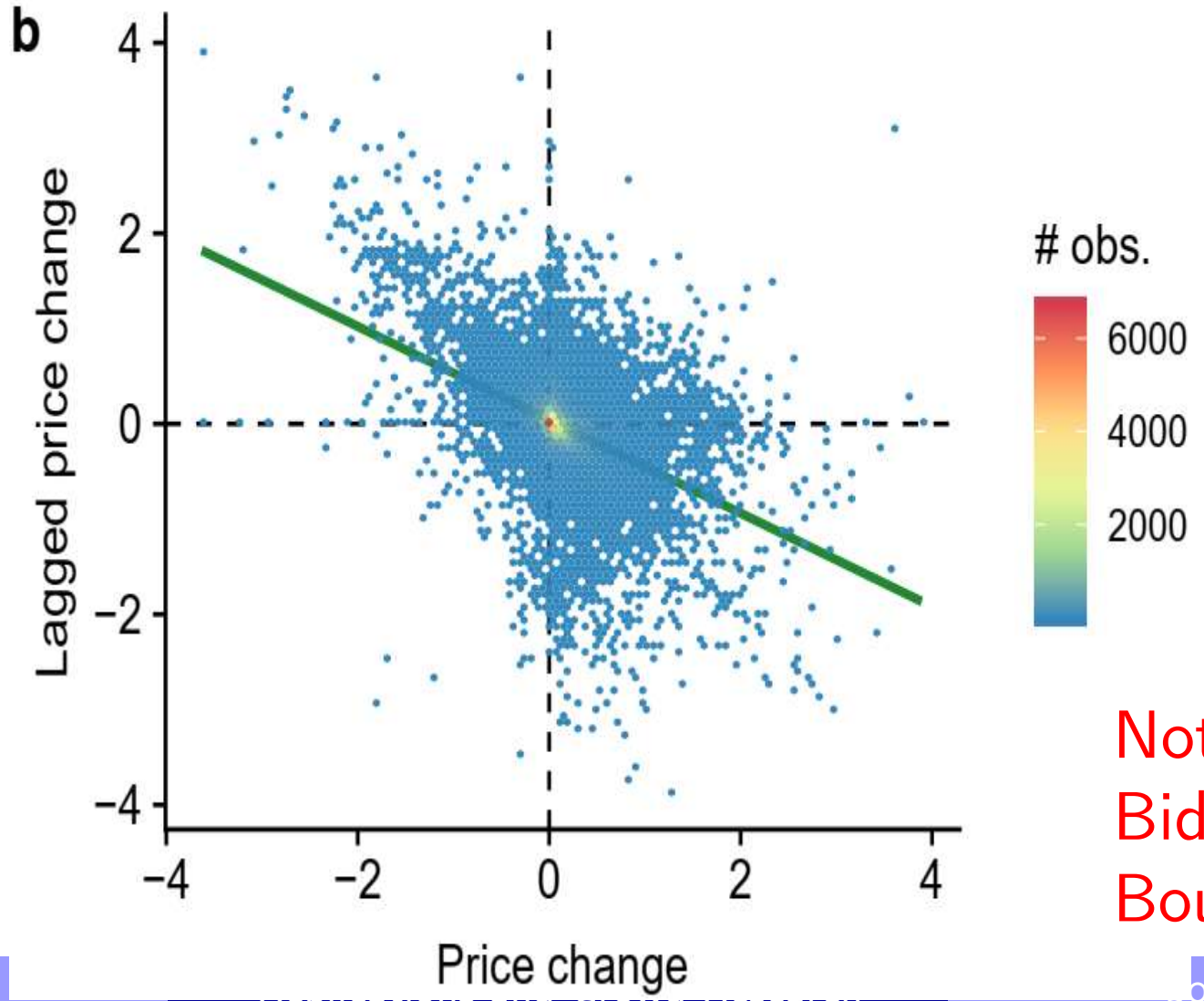
Look at only Accepted Bids? No Effect!

Price Change Autocorrelation = -0.412



Look at only Accepted Asks? No Effect!

Price Change Autocorrelation = -0.451



Not Due to Bids-Ask Bounce!

Regional Differences? Small!