WORKING PAPERS

The robustness of mispricing results in Experimental asset markets

Owen Powell
Natalia Shestakova

Mai 2017

Working Paper No: 1701

DEPARTMENT OF ECONOMICS

UNIVERSITY OF VIENNA

The robustness of mispricing results in experimental asset markets

Owen Powell[1,3] and Natalia Shestakova[2,3]

[1]Institute for Markets and Strategy, Vienna University of Economics and Business

[2]Department of Economics, University of Vienna

[3]Vienna Center for Experimental Economics, University of Vienna

Abstract

Many experiments have been conducted on market mispricing, however there is a distinct lack of guidance over how mispricing should be measured. This raises concerns about the sensitivity of mispricing results to variations in the measurement procedure. In this paper, we investigate the robustness of previous results with respect to four variations: the choice of interval length, the use of the bid-ask spread as a price proxy, the choice of aggregation function, and controlling for observable market characteristics. While a majority of previous results are unaffected, roughly 30% do change significance.

JEL Classification: C43 C90 D49 D84 G14

Keywords: Asset markets; Meta-study; Mispricing

We thank participants at the 2015 VCEE Workshop, the 2016 Experimental Finance Meeting and the 2016 WEHIA Workshop for comments.

1 Introduction

Modern society relies on markets to efficiently allocate resources across different uses. One important property of markets that has received considerable attention is mispricing. Mispricing refers to the extent to which prices in a market might deviate from underlying fundamental values. This is a difficult issue to study in the field since fundamental values are typically not observed. In contrast, experiments are designed in such a way that the fundamental value is known. For this reason, it is common to study mispricing in experiments.

In order to actually measure mispricing, various decisions have to be made. Different studies have used different measurement procedures, and therefore it is not clear to what extent results are sensitive to the choice of procedure. For example, most procedures consist of aggregating a set of price indices over time, yet it is not clear how to choose among the set of possible aggregation functions. Additional issues arise when constructing the price indices themselves. Should they be based on transactions only, or on all available information (i.e. the bid-ask spread)? What is the appropriate length of time that an index should cover? Recent research has highlighted that several variables such as gender and the relative supply of the two assets (the "cash-to-asset" ratio) can influence mispricing, but variations in these factors across markets are typically not controlled for.[1]

[1] For example, Kirchler et al. (2012) find that relative asset supply influences mispricing. Yet even when it is not the treatment variable, this factor may vary substantially across and even within treatments. Under the popular Design 4 of Smith et al. (1988), the relative asset supply can vary by more than 100% (0.81-1.83).

We are not aware of any study that examines the role of such variations, either individually or collectively. For this reason, in this paper we test the robustness of experimental asset market results to four variations: 1) the choice of interval length, 2) the use of the bid-ask spread as a proxy for price during intervals of no transaction activity, 3) the type of mean used to aggregate over indices, and 4) whether or not mispricing is adjusted for observable market characteristics. We estimate the collective and individual effects of each variation, and compare previously reported results to those obtained under a fixed measurement procedure.

First, we find that the choice of interval length and usage of the bid-ask spread have limited impact on mispricing in comparison to the choice of mean and controlling for market characteristics. Second, evaluating all hypotheses under a fixed measurement procedure causes a substantial minority of results to be overturned. As a by-product of our analysis, we also derive estimate of marginal effects of various characteristics on different types of mispricing. We find that relative asset prices are only affected by the relative supply of the assets: the higher the relative supply of an asset, the lower its relative price. On the other hand, absolute mispricing responds negatively to the amount of variation in the fundamental value, positively to variation in the relative asset supply, and negatively to the experience level of subjects.

The remainder of the paper is organized as follows. Section 2 introduces the methodology. We present our dataset in Section 3. Section 4 presents the results. The final section concludes with a discussion of implications for research agendas both past and present.

2 Methodology

Our methodology consists of the following. First, we define the term "experimental asset market". Second, we describe a general procedure for measuring mispricing in experimental asset markets that is commonly used in the literature. We identify four aspects of this procedure that are open to interpretation. By making particular choices for these aspects, we come up with what we term "standard" (which roughly coincides with current practice) and "alternative" procedures. We estimate the collective difference between these two procedures, as well as the effect of individual variations. Finally, we compare published results to results calculated under our alternative measurement procedure.

2.1 Experimental asset markets

We restrict our attention to experimental asset markets which are a generalization of those studied by Smith et al. (1988). To be precise, we define an *experimental asset market* as a market in which:

1. participants trade two assets for one another,
2. participants receive an endowment of the assets, independent of any other previous market activity,
3. assets generate the same returns to all participants,
4. all participants have the same information about the returns, and
5. exchange takes place in a controlled experimental setting.

In cases in which more than two assets are traded simultaneously, the first criterion states that each pair of assets that are traded for one another is treated as a separate market. The second criterion implies that a new market is started every time subjects are given a new set of exogenous endowments (and not, for example, by the payment of dividends). The third and fourth criteria remove any discussion about the appropriate benchmark for measuring mispricing: beliefs about returns are the same for all agents, so the fundamental value is given by the expected returns for a representative agent. Finally, the last criterion insures that market characteristics are observable, while limiting the variation in unobservables.[2]

2.2 Mispricing

Mispricing refers to the relative valuation of two assets: over time, how "close" was their subjective valuation (as given by prices) to their fundamental value (as given by expected returns)? As is standard in the literature, we differentiate between two forms of mispricing (Stöckl et al., 2010):

1. overpricing (OP): both the direction and magnitude of mispricing, and
2. absolute mispricing (AMP): only the magnitude of mispricing.

Irrespective of type, the mispricing measures considered in this paper all take the following form. First, the market is divided into T time intervals of equal length. Within each interval $t \in 1, ..., T$, a price p_t and fundamental value v_t

[2] We make no further assumptions regarding the markets, even though in practice many of the markets we study do share various other characteristics (such as, for instance, the presence of dividend payments).

are constructed. Finally, prices and fundamental values are aggregated and compared to form an overall measure of mispricing m for the market.

2.3 Variations

This description of the procedure for measuring mispricing leaves several implementation details undetermined. For example, the standard practice in the literature is to form intervals based on the timing of so-called dividend payments, and to aggregate observations using the arithmetic mean. To the best of our knowledge, neither of these choices has ever been theoretically justified: they are simply chosen because they are "natural" (Haruvy and Noussair, 2006), "standard" (Cueva and Rustichini, 2015) or because they facilitate "comparability" (Palan, 2010). No formal reason is given for why a different aggregation method and/or interval length could not be used.

This indeterminateness also extends to other issues. Intervals of time may occur during which no transactions take place, especially if shorter interval lengths are used. It is not clear what to do in these cases. A simple solution is to drop these observations, however in principle it is usually possible to interpolate a price index using the order book. Experimental markets are also designed to hold many factors constant across observations within a particular treatment, however often variation arises even within a treatment due to i.e. the realization of random variables. In individual studies, these differences are often ignored, however in principle it is possible to control for such factors.

We test the robustness of mispricing results to each of these four details. Table 1 summarizes our variations. V_0 roughly coincides with established practice, whereas V_1 is our alternative procedure. We also consider procedures that decompose the difference between V_0 and V_1. The differences we study are the following. First, we vary the interval length from the one reported in the original study (i.e. one "period") to the smallest value given the data at hand (V_a). In most cases, this is one second / tick of the market. Second, we vary whether or not the bid-ask spread is used to substitute the transaction price in intervals during which no transactions took place (V_b). Third, we vary the type of mean used to aggregate across intervals (V_c). Our final individual variation varies whether or not mispricing is adjusted for the observable characteristics of the market (V_d). We estimate both individual (V_a-V_d vs. V_1) and collective effects (V_1 vs. V_0).[3]

Additionally, we examine how previous findings (V_{REP}) change when they are re-evaluated using a fixed measurement procedure (V_{SUG}) that roughly coincides with V_1. The characteristics of V_{REP} vary from study to study. V_{SUG} is similar to V_1, with the important difference between the two being that when V_1 is undefined for a particular market, V_{SUG} progressively changes the measurement procedure until the measure is defined.[4]

The remainder of this section discusses in detail each of the variations.

[3]We consider deviations from the alternative (V_1), instead of established practice (V_0), to estimate more precisely the effect of individual variations. Under the alternative, which uses one-tick intervals and the geometric mean, there are many intervals where the bid-ask spread is potentially relevant. Additionally, it is possible to transform mispricing into a linear function of log price deviations (see Section 2.3.4). This is not the case when instead individual variations are taken relative to V_0.

[4]In particular, when V_1 is defined, $V_{SUG} = V_1$. When V_1 is undefined, we in order: 1) stop adjusting for market characteristics, 2) stop using the the bid-ask spread, and 3) change to the originally reported interval length, until V_1 is defined.

Table 1: Variations

Variation	V_0	V_1	V_a	V_b	V_c	V_d	V_{REP}	V_{SUG}
Interval	Period	Tick	Period	-	-	-	Period	Tick*
Bid-ask	No	Yes	-	No	-	-	Varies	Yes**
Aggregation	AM	GM	-	-	AM	-	Varies	GM
Adjusted	No	Yes	-	-	-	No	No	Yes

Notes: - = the alternative procedure (V_1) value; Period = the interval length reported in the original study; AM = arithmetic mean; GM = geometric mean; * = uses 1 tick if available, otherwise the smallest possible value given the data; ** = yes, if data is available, otherwise no.

2.3.1 Interval length

Established practice uses an interval length equal to the length of time between so-called dividend realizations. A dividend realization is any realization of a return by one of the assets that occurs at regular intervals (regardless of whether it is added to a participant's asset holdings immediately or stored in a separate non-trading account). When dividend realizations are not present, the entire market is taken as a single interval. One implication of this definition is that the fundamental value is always constant within an interval.

The alternative we use is the smallest interval length possible given the reporting frequency of the data at hand. In most cases, this is one second. This has the advantage of being the same frequency regardless of the particular return structure of the assets, while also conserving the constant fundamental value property within an interval.

2.3.2 Bid-ask spread

In established practice, intervals with no transaction prices are dropped from the analysis. Alternatively, we use the bid-ask spread to construct a price index for intervals with no transactions.

As noted above, the fundamental value within an interval is constant for all of the interval lengths we consider, therefore the bid-ask spread price is simply the geometric mean of the highest bid and lowest ask prices within an interval. The geometric mean is used to satisfy numeraire independence (Powell, 2016).

2.3.3 Aggregation function

Established practice uses the arithmetic mean to aggregate across intervals (Stöckl et al., 2010), whereas the alternative procedure employs the geometric mean. Both means are members of the set of generalized means, all of which share many of the same properties with respect to mispricing. However, the advantage of the geometric mean is that it is the only generalized mean that is independent of the choice of numeraire (Powell, 2016). Table 2 summarizes the different measures used, depending on the type of mispricing and mean.

2.3.4 Adjusted mispricing

The markets we consider differ from one another in several dimensions, both intentionally and by chance. For example, some markets last longer than others, while others consist of larger quantities of the assets. Established practice does not take into account these differences. Our alternative procedure takes these differences into account by constructing an adjusted measure

Table 2: Mispricing formulae

Mean	Mispricing	Measure	Formula
Arithmetic	OP	RD	$= \frac{1}{T}\sum(p_t - v_t) / \frac{1}{T}\sum v_t$
Arithmetic	AMP	RAD	$= \frac{1}{T}\sum \; \lvert p_t - v_t \rvert / \frac{1}{T}\sum v_t$
Geometric	OP	GD	$= \prod (p_t/v_t)^{1/T} - 1$
Geometric	AMP	GAD	$= \exp\left(\frac{1}{T}\sum \lvert \ln p_t/v_t \rvert\right) - 1$

Notes: p_t and v_t are the price and fundamental in interval $t \in 1, \ldots, T$, respectively. To see the correspondence between RD and GD, note that $RD = \frac{1}{T}\sum p_t / \frac{1}{T}\sum v_t - 1$ and $GD = \prod p_t^{1/T} / \prod v_t^{1/T} - 1$. To see the correspondence between GD and GAD, note that $GD = \exp\left(\frac{1}{T}\sum \; \ln p_t/v_t \; \right) - 1$.

of mispricing m':

$$m' = m - bx \tag{1}$$

where m is the original mispricing given by one of the formulae in Table 2, x is the set of characteristics and b a corresponding set of estimated marginal effects. For the set of market characteristics x, the marginal effects b for a particular variation are estimated using a regression of the form:

$$\begin{aligned} m_{i,j} &= \alpha + x_{i,j}\beta + \gamma_j + \epsilon_{i,j} \\ \epsilon_{i,j} &\sim N(0, \sigma_j^2) \\ i &\in 1, ..., N_j \\ j &\in 1, ..., R \end{aligned}$$

where $m_{i,j}$ is the unadjusted mispricing for market i from study-treatment

j, $x_{i,j}$ its characteristics and $\epsilon_{i,j}$ is a normally-distributed error term with treatment-specific variance σ_j. N_j is the number of markets of treatment j, and R is the number of treatments. Treatment characteristics are captured by the intercept γ_j and variance of the error term σ_j.[5]

We conclude this section by briefly discussing the variables used in the regression.

Mispricing: $\log(GD+1)$, $\log(GAD+1)$

We use transformed levels of mispricing that assign an equal magnitude of mispricing to markets in which deviations of prices from fundamentals are of equal (log) magnitude. This implies, for example, that a market in which prices are always double the fundamental value will have the same magnitude of (transformed) mispricing as a market in which prices are always half of fundamentals.[6]

The regression coefficients β then indicate either the percentage change in or elasticity of mispricing due to changes in the regressors, depending on whether the regressor is measured in logs or levels.

[5] When evaluating hypotheses about treatment effects, we exclude from the set of regressors used to calculate m' any variable associated with the hypothesis in question. For example, if the hypothesis relates to experience level of subjects, then for markets used to test that hypothesis subject experience is omitted from the set of characteristics in (1).

[6] Consider one market where prices $p_t = \alpha \cdot v_t$ are a constant multiple $\alpha > 1$ of fundamentals, and a second where prices $p'_t = v_t/\alpha$ are the corresponding fraction of fundamentals. Transformed GD of the two markets is of equal magnitude but opposite in sign: $\log(GD(p,v)+1) = \alpha = -\log(GD(p',v)+1)$. The same applies to transformed GAD, but not to the original GD, GAD, RD and RAD measures.

Fundamental value: $E(\log FV)$**,** $s.d.(\log FV)$

Recall that our definition of an experimental asset market simply consists of two generic assets that are traded for one another among a set of traders. At any point in time $t \in 1, ..., T$, the fundamental value:

$$v_t = r_t^B / r_t^A \tag{2}$$

is the ratio of the the expected returns r_t^A, r_t^B to holding a single unit of each of the assets A and B from t until the end of the market. It represents a rate of exchange between the two assets that rules out arbitrage possibilities. In the standard design, where A refers to cash and B to shares, $r_t^A = 1$ and r_t^B is a decreasing function of t. FV refers to the entire vector of interval observations, $FV = v_1, ..., v_T$. Logs are taken to give equal weight to proportional deviations from $v = 1$ (see Footnote 6).

It is not surprising that prices rise with the average level of FV. However, Noussair et al. (2012) find that the nominal level of FV has an effect on prices (and hence mispricing) that is more than just proportional. With respect to variation in FV, Stöckl et al. (2014) find that markets with constant fundamentals exhibit much lower absolute mispricing compared to markets with non-constant fundamentals.

Relative asset supply: $E(\log RAS)$**,** $s.d.(\log RAS)$

The relative supply of the two assets in the market, taking into account FV, at interval t of a market is:

$$RAS_t = \frac{A_t}{v_t B_t} \tag{3}$$

where A_t and B_t are the total quantity in the market at interval t of the two assets, respectively. RAS refers to the entire vector of RAS_t interval observations. Logs are taken to give equal weight to proportional deviations from $RAS = 1$ (see Footnote 6). This variable may clearly vary across designs, but even within a particular design it may vary, due to (for example) the realization of stochastic dividend payments.

Kirchler et al. (2012) shows that relative asset supply is an important determinant of prices. In particular, assets that are in relatively high (low) supply tend to be under- (over-) priced. Therefore high RAS creates more overpricing. This supports the downward-sloping demand hypothesis: the larger the supply of an asset, the lower its valuation. Haruvy et al. (2013) find a similar effect after an exogenous intervention that alters the supply of assets in the middle of the market.

Duration (DUR)

The length of the market, measured in hours of trading time.

Number of traders ($NSUBJ$)

The number of human participants in the market, regardless of whether they act independently of one another or participate in groups.

Experience (EXP_{mkts}, EXP_{dur})

Several studies (for example, King et al. (1993)) show that mispricing decreases with repetition of the market environment. The standard way to measure experience in the literature is the average number of markets that a trader had previously participated in within the same study (EXP_{mkts}). However, one issue with this definition of experience is that the meaning of a "market" varies from study to study. Therefore we also consider a second measure of experience that controls to some extent for differences in market design across studies. EXP_{dur} is the average duration of market trading (measured in hours) that traders have previously participated in within the same study. In the same way that mispricing has been shown to decrease in the number of markets that have been experienced by subjects, lower mispricing may also result from a longer time spent in previous markets. Both effects may be interpreted as (distinct) measures of learning.

3 Data

We consider peer-reviewed publications from 2005-2015 that satisfy our inclusion criteria (48 studies). Table 3 shows the list of 27 studies for which we have data. From these studies, we compile a set of 848 market observations from 142 different treatments, and 144 hypotheses related to treatment differences (77 for absolute mispricing, 67 for overpricing) that can be tested using a standard two-sided Mann-Whitney test procedure.

Table 3: Included studies

study	markets	treatments	comparisons
Dufwenberg et al. (2005)	40	6	6
Ackert et al. (2006)	26	6	0
Haruvy and Noussair (2006)	26	8	0
Haruvy et al. (2007)	23	4	0
Hussam et al. (2008)	28	11	0
Ackert et al. (2009)	72	3	0
Noussair and Powell (2010)	40	8	8
Palan (2010)	14	2	0
Fiedler (2011)	13	2	2
Lahav (2011)	6	1	0
Akiyama et al. (2012)	10	1	0
Cheung and Palan (2012)	26	7	4
Kirchler et al. (2012)	42	7	24
Palfrey and Wang (2012)	78	7	0
Schoenberg and Haruvy (2012)	14	2	1
Fischbacher et al. (2013)	58	8	5
Haruvy et al. (2013)	18	3	6
Cheung et al. (2014)	40	4	6
Lugovskyy et al. (2014)	22	3	6
Stöckl et al. (2014)	30	5	20
Breaban and Noussair (2015)	32	4	8
Cason and Samek (2015)	60	10	13
Corgnet et al. (2015)	20	2	1

Cueva and Rustichini (2015)	30	4	8
Cueva et al. (2015)	15	3	0
Eckel and Füllbrunn (2015)	19	3	6
Huber et al. (2016)	46	8	20
Total	848	142	144

In order to calculate mispricing in a market, it is necessary to fix one of the assets as the numeraire. This determines prices, fundamental values and hence mispricing. We use the data as they are originally reported i.e. using the numeraire asset from each study as it is reported by the study. The choice of numeraire affects results based on RD and RAD, but not those based on GD and GAD.

4 Results

We consider two different research questions. First, we estimate the individual and collective effect of variations in the measurement procedure ($V_a - V_d$ vs. V_1). This estimates the relative importance of these different variations in the measurement procedure. Second, we compare mispricing as it is reported in the original study (V_{REP}) to a particular measurement procedure (V_{SUG}) that, given data limitations, is as similar to our V_1 procedure as possible. This tests the robustness of previous results to changes in the measurement procedure.

For each measurement procedure, we first calculate the probability of rejecting the null hypotheses based on a two-sided Mann-Whitney test. Then, comparing any two measurement procedures simply consists of comparing the

Figure 1: Proportion of p-values that switch significance due to different variations

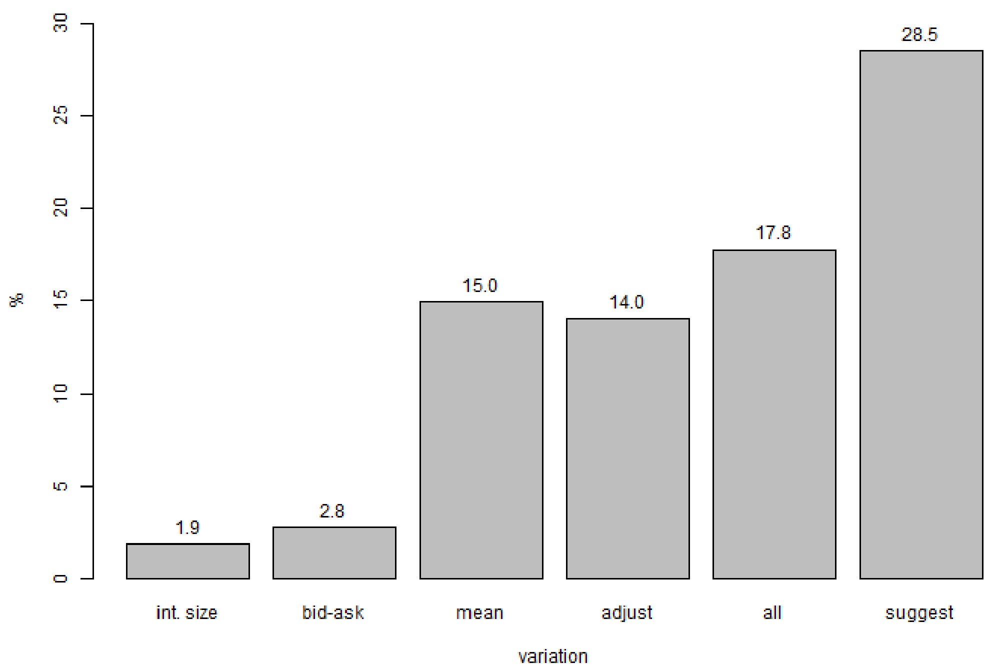

Figure 1 shows the proportion of p-values that switch significance for a threshhold level of 0.05 for a given variation. "int. size": switching from an interval size of 1 tick to one "period"; "bid-ask": switching from using to not using the bid-ask spread as a measure of prices; "mean": switching from a geometric to an arithmetic mean; "adjust": switching from adjusting to not adjusting for market characteristics; "all": the combination of the previous four variations; "suggest": comparing the originally-reported p-values (V_{REP}) with those from V_{SUG}.

associated set of p-values.

4.1 Variation effects

The results for absolute mispricing and overpricing are quite similar, therefore we combine the discussion (see Table 4 for complete results). Figure 1 shows that compared to a baseline of V_1, the interval size and the bid-ask spread only have small effects on test outcomes - only 2-3% of p-values change significance. The choice of mean and adjusting for observable market characteristics switch the significance of 14-15% of results. Including all variations simultaneously is roughly equivalent to the sum of the individual variations (17.8%).

Thus, for the most part mispricing analysis is robust to the variations given by V_a-V_d. Nevertheless, a substantial minority are affected, in particular by changes in the type of aggregation method and whether or not adjustments are made for market characteristics.

The estimates of the marginal effects of different market characteristics on mispricing (based on V_1) is given in Table 6 in the Appendix.[7] The only variable with a significant effect on overpricing is the relative supply of the two assets: the larger the supply, the lower the overpricing. On the other hand, overpricing does *not* depend on the level of trader experience or the properties of the fundamental value.

In contrast, absolute mispricing is affected by more factors. Variation in the fundamental value increases AMP, whereas to a smaller extent variation

[7] The results for the other procedures that control for characteritics, V_a - V_c, are similar and available upon request.

in the relative asset supply decreases AMP. We also replicate the previous finding that absolute mispricing decreases with increased experience, but it is the duration (rather than number of markets) that captures this effect.

4.2 Original vs. suggested mispricing

The previous section examined the impact of controlling for certain variables and market characteristics on mispricing differences across treatments. This section focusses on how the results of actual reported hypothesis tests (V_{REP}) change when analyzed under our suggested conditions (V_{SUG}).

Figure 2 shows how p-values change when moving from the procedures reported in the original study (V_{REP}) to our suggested procedure (V_{SUG}). Overall, more than a quarter (41 out of 144) of hypotheses switch significance. To highlight the implication for previous findings, we discuss two examples: one in which hypothesis results switch to becoming insignificant, and one in which they become significant (complete results are available in the Appendix).

First, Noussair and Powell (2010) consists of treatments that differ in terms of their path for fundamental values (*Peak* vs. *Valley*), and the amount of experience of subjects (between 0 and 3 markets). Mispricing (of both types) in markets with experienced subjects was originally found to differ significantly depending on the path of fundamental value ($P3$ vs. $V3$, and $P4$ vs. $V4$). However, under our suggested mispricing procedure, the treatment difference is no longer significant. This suggests that the path of fundamental value per se is less important than originally thought. Similar conclusions apply to some of the affected hypotheses from Kirchler et al. (2012) and

Figure 2: p-values under original vs. suggested measurement procedures

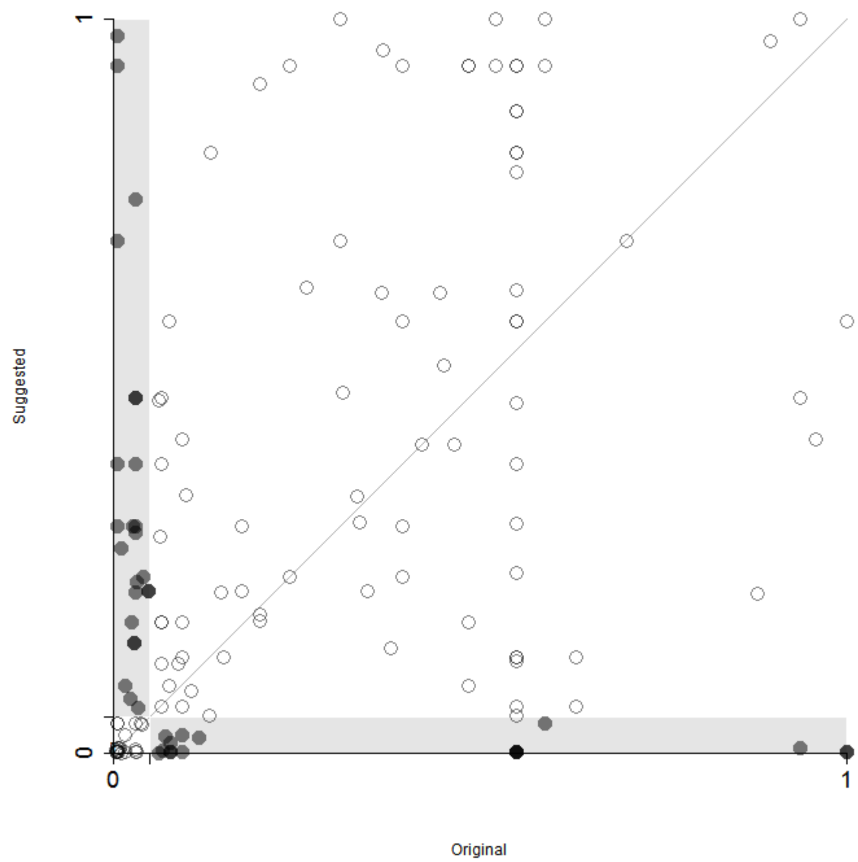

Figure 2 shows the p-value for each hypothesis test calculated under the original vs. suggested procedures. Grey circles are values that switch significance at the 0.05 level. Some studies only report the level at which a value is significant - in these cases, "insignificant" is coded as 0.55, and "significant at the 10% / 5% / 1% level" as 0.075 / 0.03 / 0.005, respectively.

Stöckl et al. (2014).

Second, Cason and Samek (2015) contains an example of results that switch to being significant under the alternative measurement variation. Part of this study compares markets in which information is shown to inexperienced traders as either text ($TextM1$) or graphically ($VisualM1$). Under the originally reported mispricing procedure, absolute mispricing does not differ between the two types of markets ($p = 0.937$). However, under the alternative procedure, the difference between treatments turn out to be strongly significant ($p = 0.008$).

Therefore, although our results show that a majority of results are not affected by the change in measurement procedure, a substantial minority (28.5%) are.

5 Conclusion

This study has examined the sensivity of experimental asset market results to changes in measurement procedure. The results have implications for both design of new market experiments and for previous findings.

First, the choice of interval size and the use of the bid-ask spread for intervals with no transactions have a limited effect compared to the choice of aggregation technique and controlling for the characteristics of the market. Second, we have examined how much actual results change when re-evaluating hypotheses from various studies under a fixed measurement procedure. Our results are of the "*glass half-full, glass half-emtpy*" genre. On the one hand, it is reassuring that a majority of results (71.5%) do not change significance under the new procedure. However, this still leaves a substantial minority (28.5%) that are affected. We think this suggests the need to further discuss and examine the sensitivity of experimental asset market research. For example, two potential areas of discussion are 1) the data requirements (minimum number of observations, recording of market characteristics) and 2) coming up with criteria for selecting among the set of measurement procedures (we have suggested one particular procedure, but others are certainly possible).

Third, we estimate the marginal impact of various market characteristics on mispricing. The characteristics that appear to be important are the level and variation of the fundamental value, and the experience level of subjects. However, it is important to keep in mind that these results may be sensitive to the data and measurement procedure used. It is certainly conceivable that these estimates may be revised as new data becomes available.

We have focussed on what we consider to be four of the most important parts of the mispricing measurement procedure. Nevertheless, we acknowledge that our findings may be sensitive to the set of alternatives under consideration. We think this work is in any case useful by highlighting the robustness of mispricing results, and can help start a discussion about which variations are most important and should be further examined in the future. It also serves as a reminder about the importance of minimizing potential sources of non-treatment variation in the experimental design.

References

Ackert, L. F., Charupat, N., Church, B. K., and Deaves, R. (2006). Margin, short selling, and lotteries in experimental asset markets. *Southern Economic Journal*, 73(2):419–436.

Ackert, L. F., Charupat, N., Deaves, R., and Kluger, B. D. (2009). Probability Judgment Error and Speculation in Laboratory Asset Market Bubbles. *Journal of Financial & Quantitative Analysis*, 44(03):719–744.

Akiyama, E., Hanaki, N., and Ishikawa, R. (2012). Effect of Uncertainty About Others' Rationality in Experimental Asset Markets: An Experimental Analysis. SSRN Scholarly Paper ID 2178291, Social Science Research Network, Rochester, NY.

Breaban, A. and Noussair, C. N. (2015). Trader characteristics and fundamental value trajectories in an asset market experiment. *Journal of Behavioral and Experimental Finance*, 8:1–17.

Cason, T. N. and Samek, A. (2015). Learning through passive participation in asset market bubbles. *Journal of the Economic Science Association*, 1(2):170–181.

Cheung, S. L., Hedegaard, M., and Palan, S. (2014). To see is to believe: Common expectations in experimental asset markets. *European Economic Review*, 66:84–96.

Cheung, S. L. and Palan, S. (2012). Two heads are less bubbly than one:

team decision-making in an experimental asset market. *Experimental Economics*, 15(3):373–397.

Corgnet, B., Hernán-González, R., Kujal, P., and Porter, D. (2015). The effect of earned versus house money on price bubble formation in experimental asset markets. *Review of Finance*, 19(4):1455–1488.

Cueva, C., Roberts, R. E., Spencer, T., Rani, N., Tempest, M., Tobler, P. N., Herbert, J., and Rustichini, A. (2015). Cortisol and testosterone increase financial risk taking and may destabilize markets. *Scientific Reports*, 5(11206):1–16.

Cueva, C. and Rustichini, A. (2015). Is financial instability male-driven? Gender and cognitive skills in experimental asset markets. *Journal of Economic Behavior & Organization*, 119:330–344.

Dufwenberg, M., Lindqvist, T., and Moore, E. (2005). Bubbles and experience: An experiment. *American Economic Review*, 95(5):1731–1737.

Eckel, C. C. and Füllbrunn, S. C. (2015). Thar she blows? Gender, competition, and bubbles in experimental asset markets. *American Economic Review*, 105(2):906–920.

Fiedler, M. (2011). Experience and confidence in an internet-based asset market experiment. *Southern Economic Journal*, 78(1):30–52.

Fischbacher, U., Hens, T., and Zeisberger, S. (2013). The impact of monetary policy on stock market bubbles and trading behavior: Evidence from the lab. *Journal of Economic Dynamics & Control*, 37(10):2104–2122.

Haruvy, E., Lahav, Y., and Noussair, C. N. (2007). Traders' expectations in asset markets: Experimental evidence. *American Economic Review*, 97(5):1901–1920.

Haruvy, E. and Noussair, C. N. (2006). The effect of short selling on bubbles and crashes in experimental spot asset markets. *Journal of Finance*, 61(3):1119–1157.

Haruvy, E., Noussair, C. N., and Powell, O. (2013). The impact of asset repurchases and issues in an experimental market. *Review of Finance*, 18(2):681–713.

Huber, J., Kirchler, M., and Stöckl, T. (2016). The influence of investment experience on market prices: laboratory evidence. *Experimental Economics*, 19(2):394–411.

Hussam, R. N., Porter, D., and Smith, V. L. (2008). Thar she blows: Can bubbles be rekindled with experienced subjects? *American Economic Review*, 98(3):924–937.

King, R., Smith, V., Williams, A., and Van Boening, M. (1993). The robustness of bubbles and crashes in experimental stock markets. In Day, R. and Chen, P., editors, *Nonlinear dynamics and evolutionary economics*, pages 183–200. Oxford: Oxford University Press.

Kirchler, M., Huber, J., and Stöckl, T. (2012). Thar she bursts: Reducing confusion reduces bubbles. *American Economic Review*, 102(2):865–883.

Lahav, Y. (2011). Price patterns in experimental asset markets with long horizon. *Journal of Behavioral Finance*, 12(1):20–28.

Lugovskyy, V., Puzzello, D., Tucker, S., and Williams, A. (2014). Asset-holdings caps and bubbles in experimental asset markets. *Journal of Economic Behavior & Organization*, 107:781–797.

Noussair, C. N. and Powell, O. (2010). Peaks and valleys: Price discovery in experimental asset markets with non-monotonic fundamentals. *Journal of Economic Studies*, 37(2):152–180.

Noussair, C. N., Richter, G., and Tyran, J.-R. (2012). Money illusion and nominal inertia in experimental asset markets. *Journal of Behavioral Finance*, 13(1):27–37.

Palan, S. (2010). Digital options and efficiency in experimental asset markets. *Journal of Economic Behavior & Organization*, 75(3):506–522.

Palfrey, T. R. and Wang, S. W. (2012). Speculative overpricing in asset markets with information flows. *Econometrica*, 80(5):1937–1976.

Powell, O. (2016). Numeraire independence and the measurement of mispricing in experimental asset markets. *Journal of Behavioral and Experimental Finance*, 9:56 – 62.

Schoenberg, E. J. and Haruvy, E. (2012). Relative performance information in asset markets: An experimental approach. *Journal of Economic Psychology*, 33(6):1143–1155.

Smith, V. L., Suchanek, G. L., and Williams, A. W. (1988). Bubbles, crashes, and endogenous expectations in experimental spot asset markets. *Econometrica*, 56(5):1119–1151.

Stöckl, T., Huber, J., and Kirchler, M. (2010). Bubble measures in experimental asset markets. *Experimental Economics*, 13(3):284–298.

Stöckl, T., Huber, J., and Kirchler, M. (2014). Multi-period experimental asset markets with distinct fundamental value regimes. *Experimental Economics*, pages 1–21.

Appendix

Table 4: Effects of variations on mispricing

comparison	$V_1 - V_a$	$V_1 - V_b$	$V_1 - V_c$	$V_1 - V_d$	$V_1 - V_0$	$V_{SUG} - V_{REP}$
description	int. size	bid-ask	mean	adjust	all	suggest
	(1)	(2)	(3)	(4)	(5)	(6)
Absolute mispricing						
avg. ch.	-2.35	0.07	1.67	2.43	5.94	6.22
insig. $\rightarrow$ sig.	1.88	1.88	7.54	0.00	0.00	7.79
sig. $\rightarrow$ insig.	0.00	1.88	7.54	13.20	15.09	19.48
total ch. sig.	1.88	3.77	15.09	13.20	15.09	27.27
N	53	53	53	53	53	77
Overpricing						
avg. ch.	0.20	2.27	3.83	-1.64	3.53	4.46
insig. $\rightarrow$ sig.	0.00	1.85	5.55	9.25	11.11	16.41
sig. $\rightarrow$ insig.	1.85	0.00	9.25	5.55	9.25	13.43
total ch. sig.	1.85	1.85	14.81	14.81	20.37	29.85
N	54	54	54	54	54	67

"avg. ch" refers to the average change in p-values. The values for significance show the proportion of hypotheses that switch in significance from insignificant to significant ("insig. $\rightarrow$ sig."), from significant to insignificant ("sig. $\rightarrow$ insig."), or the sum of both types of switches ("total ch. sig."). All switches are for the given threshhold significance level of $\alpha = 0.05$. Total number of hypotheses for each type of mispricing are 77 for absolute mispricing and 67 for overpricing. It is not possible to calculate all variations for all studies. For example, some studies do not use or report bids and asks. Each comparison of variations only includes data from those studies which are present in both variations.

Table 5: Hypotheses

study	category	comparison	V_{REP}	V_{SUG}	change
Dufwenberg et al. (2005)	AMP	$R1 - R4$	0.032	0.002	|*
		$R3 - R4$	0.061	0.481	*|
		$R4_{23} - R4_{13}$	0.421	0.420	|
	OP	$R1 - R4$	0.011	0.279	**|
		$R3 - R4$	0.897	0.970	|
		$R4_{23} - R4_{13}$	0.310	1.000	|
Noussair and Powell (2010)	AMP	$P1 - V1$	0.347	0.222	|
		$P2 - V2$	0.175	0.309	|
		$P3 - V3$	0.047	0.222	**|
		$P4 - V4$	0.028	0.150	**|
	OP	$P1 - V1$	0.465	0.420	|
		$P2 - V2$	0.175	0.222	|
		$P3 - V3$	0.047	0.222	**|
		$P4 - V4$	0.028	0.150	**|
Fiedler (2011)	OP	$1AtLrg - 2TrdGrp$	0.445	0.628	|
	AMP	$1AtLrg - 2TrdGrp$	0.365	0.628	|
Cheung and Palan (2012)	OP	$DA2H - DAIND$	1.000	0.002	|***
		$DA2H - DAIND$	0.078	0.002	|**
	AMP	$DA2H - DAIND$	1.000	0.002	|***
		$DA2H - DAIND$	0.078	0.002	|**

Schoenberg and Haruvy (2012)	OP	$DOWN-UP$	0.010	0.000		
Kirchler et al. (2012)	AMP	$T1T2-T3T4$	0.550	0.477		
		$T1-T3$	0.550	0.132		
		$T2-T4$	0.550	0.064		*
		$T1T3-T2T4$	0.005	0.977	***	
		$T1-T2$	0.030	0.309	**	
		$T3-T4$	0.005	0.309	***	
		$T1-T5$	0.025	0.179	**	
		$T1-T6R1$	0.037	0.041		
		$T1-T6R2$	0.007	0.002		
		$T5-T6R1$	0.522	0.937		
		$T5-T6R2$	0.631	0.132		
		$T6R1-T6R2$	0.550	0.132		
	OP	$T1T2-T3T4$	0.030	0.218	**	
		$T1-T3$	0.550	0.588		
		$T2-T4$	0.550	0.002		***
		$T1T3-T2T4$	0.030	0.755	**	
		$T1-T2$	0.005	0.008		
		$T3-T4$	0.550	0.393		
		$T1-T5$	0.004	0.002		
		$T1-T6R1$	0.078	0.015		*
		$T1-T6R2$	0.004	0.002		
		$T5-T6R1$	0.150	0.132		

		$T5 - T6R2$	0.631	0.064	|*
		$T6R1 - T6R2$	0.550	0.002	|***
Fischbacher et al. (2013)	OP	$E1P0 - E1P1$	0.106	0.085	|*
		$E123P0 - E123P1$	0.062	0.000	|**
	AMP	$E123P0 - E123P1$	0.067	0.004	|**
	OP	$E2P0 - E4P1$	0.093	0.025	|*
		$E2P1 - E4P2$	0.093	0.002	|**
Haruvy et al. (2013)	OP	$1B - 2R$	0.485	0.179	|
		$1B - 3SI$	0.310	0.699	|
		$2R - 3SI$	0.015	0.002	|*
	AMP	$1B - 2R$	0.394	0.937	|
		$1B - 3SI$	0.015	0.025	|
		$2R - 3SI$	0.009	0.008	|
Cheung et al. (2014)	AMP	$1PK - 4BASE$	0.003	0.005	|
		$2NPK - 4BASE$	0.088	0.123	*|
		$1PK - 2NPK$	0.033	0.063	*|
	OP	$1PK - 4BASE$	0.335	0.314	|
		$2NPK - 4BASE$	0.066	0.123	*|
		$1PK - 2NPK$	0.099	0.352	*|
Lugovskyy et al. (2014)	OP	$G1 - G2$	0.030	0.005	|*
		$G2 - G3$	0.550	0.246	|
		$G1 - G3$	0.030	0.003	|*
	AMP	$G1 - G2$	0.550	0.051	|*
		$G2 - G3$	0.550	0.125	|

		$G1-G3$	0.030	0.301	**|
Stöckl et al. (2014)	OP	$R1-R2$	0.005	0.004	|
		$R1-R3$	0.005	0.699	***|
		$R1-R4$	0.030	0.393	**|
		$R1-R5$	0.075	0.588	*|
		$R2-R3$	0.005	0.002	|
		$R2-R4$	0.550	0.002	|***
		$R2-R5$	0.550	0.002	|***
		$R3-R4$	0.030	0.041	|
		$R3-R5$	0.075	0.093	|
		$R4-R5$	0.550	0.588	|
	AMP	$R1-R2$	0.005	0.393	***|
		$R1-R3$	0.030	0.484	**|
		$R1-R4$	0.030	0.484	**|
		$R1-R5$	0.550	0.818	|
		$R2-R3$	0.005	0.041	*|
		$R2-R4$	0.005	0.002	|
		$R2-R5$	0.005	0.041	*|
		$R3-R4$	0.550	0.937	|
		$R3-R5$	0.550	0.937	|
		$R4-R5$	0.005	0.937	***|
Breaban and Noussair (2015)	AMP	$DECR1-INCR1$	0.550	0.874	|
		$DECR2-INCR2$	0.550	0.792	|

<table>
<tr><td></td><td></td><td>$DECR1-DECR2$</td><td>0.378</td><td>0.143</td><td>|</td></tr>
<tr><td></td><td></td><td>$INCR1-INCR2$</td><td>0.065</td><td>0.393</td><td>*|</td></tr>
<tr><td></td><td>OP</td><td>$DECR1-INCR1$</td><td>0.550</td><td>0.313</td><td>|</td></tr>
<tr><td></td><td></td><td>$DECR2-INCR2$</td><td>0.550</td><td>0.874</td><td>|</td></tr>
<tr><td></td><td></td><td>$DECR1-DECR2$</td><td>0.550</td><td>0.630</td><td>|</td></tr>
<tr><td></td><td></td><td>$INCR1-INCR2$</td><td>0.550</td><td>0.818</td><td>|</td></tr>
<tr><td>Cason and Samek (2015)</td><td>AMP</td><td>$TextM1-VisualM1$</td><td>0.937</td><td>0.008</td><td>|***</td></tr>
<tr><td></td><td></td><td>$TextM2-VisualM2$</td><td>0.484</td><td>0.093</td><td>|*</td></tr>
<tr><td></td><td></td><td>$TextM3-VisualM3$</td><td>0.588</td><td>1.000</td><td>|</td></tr>
<tr><td></td><td></td><td>$PreTM2-PreVM2$</td><td>0.093</td><td>0.179</td><td>*|</td></tr>
<tr><td></td><td></td><td>$PreTM3-PreVM3$</td><td>0.015</td><td>0.093</td><td>*|</td></tr>
<tr><td></td><td></td><td>$PreTM2-TextM1$</td><td>0.588</td><td>0.041</td><td>|**</td></tr>
<tr><td></td><td></td><td>$PreTM3-TextM2$</td><td>0.393</td><td>0.240</td><td>|</td></tr>
<tr><td></td><td></td><td>$PreTM2-TextM2$</td><td>0.093</td><td>0.132</td><td>*|</td></tr>
<tr><td></td><td></td><td>$PreTM3-TextM3$</td><td>0.025</td><td>0.309</td><td>**|</td></tr>
<tr><td></td><td></td><td>$PreVM2-VisualM1$</td><td>0.064</td><td>0.484</td><td>*|</td></tr>
<tr><td></td><td></td><td>$PreVM3-TextM2$</td><td>0.484</td><td>0.937</td><td>|</td></tr>
<tr><td></td><td></td><td>$PreVM2-TextM2$</td><td>0.937</td><td>1.000</td><td>|</td></tr>
<tr><td></td><td></td><td>$PreVM3-VisualM3$</td><td>0.240</td><td>0.937</td><td>|</td></tr>
<tr><td>Corgnet et al. (2015)</td><td>AMP</td><td>$EM-HM$</td><td>0.130</td><td>0.052</td><td>|*</td></tr>
<tr><td>Cueva and Rustichini (2015)</td><td>OP</td><td>$T1FEM-T2MALE$</td><td>0.199</td><td>0.190</td><td>|</td></tr>
<tr><td></td><td></td><td>$T1FEM-T2HET$</td><td>0.199</td><td>0.911</td><td>|</td></tr>
<tr><td></td><td></td><td>$T2MALE-T2HET$</td><td>0.023</td><td>0.075</td><td>*|</td></tr>
<tr><td></td><td></td><td>$T1HOM-T2HET$</td><td>0.333</td><td>0.350</td><td>|</td></tr>
</table>

	AMP	$T1FEM - T2MALE$	0.450	0.528	|
		$T1FEM - T2HET$	0.879	0.217	|
		$T2MALE - T2HET$	0.070	0.023	|*
		$T1HOM - T2HET$	0.039	0.039	|
Eckel and Füllbrunn (2015)	OP	$M - F$	0.007	0.008	|
		$M - Mix$	0.032	0.234	**|
		$Mix - F$	0.116	0.022	|**
	AMP	$M - F$	0.522	1.000	|
		$M - Mix$	0.063	0.294	*|
		$Mix - F$	0.199	0.180	|
Huber et al. (2016)	AMP	$T1 - T2$	0.093	0.064	|
		$T1 - T3$	0.093	0.427	*|
		$T1 - T4$	0.041	0.240	**|
		$T2 - T5$	0.064	0.179	*|
		$T2 - T6$	0.240	0.240	|
		$T3 - T4$	0.263	0.635	|
		$T3 - T5$	0.957	0.427	|
		$T4 - T5$	0.588	0.937	|
		$T4 - T6$	0.937	0.484	|
		$T5 - T6$	0.393	0.309	|
	OP	$T1 - T2$	0.064	0.064	|
		$T1 - T3$	0.147	0.219	|
		$T1 - T4$	0.064	0.179	*|
		$T2 - T5$	0.132	0.818	|

$T2 - T6$	0.393	0.588	|
$T3 - T4$	0.313	0.492	|
$T3 - T5$	0.367	0.957	|
$T4 - T5$	1.000	0.588	|
$T4 - T6$	0.699	0.699	|
$T5 - T6$	0.484	0.937	|

Notes: Table shows how originally reported p-values of hypotheses (column V_{REP}) differ from those calculated under our suggested procedure (column V_{SUG}). The original procedure varies from study to study, whereas the suggested procedure is as close to V_1 as possible, given the data at hand. Some studies only report the level at which a value is significant - in these cases, "insignificant" is coded as 0.55, and "significant at the 10% / 5% / 1% level" as 0.075 / 0.03 / 0.005, respectively. The last column, *change*, reports how the number of conventionally reported stars (* = 0.1, ** = 0.05, *** = 0.01) are affected. The vertical bar indicates no change, stars to the left (right) of the bar represent fewer (more) stars under the suggested variation.

Table 6: Characteristics and mispricing regressions

	Overpricing	Absolute mispricing
dep. variable	$\log(GD+1)$	$\log(GAD+1)$
$E(\log FV)$	-0.269	
	(0.287)	
$E(\log RAS)$	0.280**	
	(0.123)	
$\lvert E(\log FV)\rvert$		0.290
		(0.183)
$\lvert E(\log RAS)\rvert$		0.055
		(0.085)
$s.d.(\log FV)$	-0.113	1.042***
	(0.566)	(0.315)
$s.d.(\log RAS)$	-0.077	-0.262**
	(0.260)	(0.131)
$NSUBJ$	-0.020	0.005
	(0.025)	(0.037)
DUR	-0.212	0.715*
	(0.223)	(0.425)
EXP_{mkts}	-0.013	-0.014
	(0.035)	(0.033)
EXP_{dur}	-0.097	-0.136**

	(0.073)	(0.055)
N	394	394

OLS regression of mispricing explained by market characteristics, based on procedure V_1. Study-treatment dummies included in all regressions. Errors are clustered at the study-treatment level.