The robustness of mispricing results in experimental asset markets

Owen Powell
Department of Economics
University of Vienna
opowell@gmail.com

Natalia Shestakova
Department of Economics
University of Vienna
natalia.shestakova@univie.ac.at

December 31, 2016

Abstract

Many experimental studies study 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: 1) the choice of interval length, 2) the use of the bid-ask spread as a price proxy, 3) the choice of aggregation function, and 4) controlling for observable market characteristics. While a majority of previous results are unaffected, we still find that roughly 30% of previous hypothesis results 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 price efficiency. Efficiency refers to the degree to which prices in a market reflect 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 measure mispricing, various decisions have to be made. Different papers in the literature have used different approaches, 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. A popular way to do this is with with the arithmetic mean (see the *RD* and *RAD* measures proposed by Stöckl and Kirchler (2014)) because it satisfies certain criteria. However the arithmetic mean is but one member of the set of generalized means, all of which have these same properties. 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 relative asset supply ("cash-to-asset" ratio) can influence mispricing, but variations in these factors across experimental markets is typically not controlled for. Consider the case of relative asset supply. Kirchler et al. (2012) find that this has a clear influence on market prices and hence mispricing. Yet even in designs where it is not the treatment variable, this factor may vary substantially across and even within treatments.

For example, in many designs realized asset returns are stochastic. Therefore, while ex-ante markets of the same design have the same expected asset supplies, ex-post the supplies in the market may differ because of different return realizations. Under the popular Design 4 of Smith et al. (1988), the average asset supply can vary by more than 100% (0.81-1.83). Of course, with large numbers of markets, on average these differences should cancel out across treatments. Unfortunately, experimental asset market studies typically consist of relatively few independent observations, thus making it difficult to ignore this

issue.

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 both the effect of each variation, and compare previously reported results to results obtained under our preferred method of measuring mispricing.

First, we find that the choice of interval length, usage of the bid-ask spread and the choice of mean have limited impact on mispricing in comparison to the role of controlling for market characteristics. Second, evaluating all hypotheses under a fixed measurement specification causes a substantial majority of results to be overturned. Finally, we derive estimates of the marginal impact of various characteristics on market mispricing.

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 baseline approach for measuring mispricing in experimental asset markets, which is more or less the standard approach used in the literature. Third, we describe four variations to the baseline approach. Fourth, we compare the results from using the variations versus the baseline approach. We consider the effect of variations both individually and collectively. Finally, we compare actual reported results to our suggested variation.

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 criteria states that each traded pair of assets is treated as a separate market. The second criteria 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 single representative agent. Finally, the last criteria insures that market characteristics are observable, while limiting the variation in unobservables. We make no further assumptions regarding the assets, even though in practice many of the assets in the markets we study do share various other characteristics (such as, for instance, the presence of dividend payments).

2.2 Mispricing

Price efficiency 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: both the direction and magnitude of mispricing, and
2. absolute mispricing: only the magnitude of mispricing.

One can imagine many ways in which both types of mispricing may be measured. We restrict our attention to an approach that captures many of the most popular measures such as *Relative Deviation (RD)* and *Relative Absolute Deviation (RAD)* (Stöckl and Kirchler, 2014), and *Average Bias (AB)* and *Total Dispersion (TD)* (Haruvy and Noussair, 2006)). First, the market is divided

into T time intervals of equal length. During each interval of time, indices of prices and fundamental value are constructed. Finally, the interval observations are aggregated and compared to form an overall measure of mispricing for the market.

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 the final value of the measure. In general, the choice of numeraire will affect the value of mispricing, hence the results themselves are sensitive to this choice[1]. 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.

2.3 Variations

The previous description of the approach to measuring mispricing leaves open several implementation details. 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 to facilitate comparison to previous work (Palan, 2010). There does not appear to be any formal reason 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 would be to drop these observations, however in principle it is usually possible to interpolate a price index using unfulfilled bid and ask offers.

Experimental markets are 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.

We test the robustness of mispricing results to each of these four details. Table 1 summarizes our variations and a baseline design which roughly coincides with established practice. First, we vary the interval length from one "period"

[1] *Geometric Deviation (GD)* and *Geometric Absolute Deviation (GAD)* (Powell, 2016) are not affected by the choice of numeraire.

Table 1: Variations

Variation Item	V_0 Baseline	V_1	V_2	V_3	V_4	V_5 All	V_{REP}	V_{SUG}
Interval length	Period	1 tick	-	-	-	1 tick	Period	As small as possible
Bid-ask spread	No	-	Yes	-	-	Yes	Varies	Yes if possible
Aggregation function	AM	-	-	GM	-	GM	Varies	GM
Adjust for observables	No	-	-	-	Yes	Yes	No	Yes

Notes: - = the baseline value; Period = the interval length reported in the original study; AM = arithmetic mean; GM = geometric mean.

to the smallest value given the data at hand (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. Third, we vary the type of mean used to aggregate across intervals (arithmetic vs. geometric). Since the geometric mean is equivalent to taking the arithmetic mean of log values, it shares many of the same properties as the arithmetic mean. It has the advantage, however, of being insensitive to the choice of numeraire (Powell, 2016). Our final variation varies whether or not mispricing is adjusted for the observable characteristics of the market.

We are not aware of any study that comprehensively examines the role of these changes, either individually or collectively. We estimate both individual (V_1-V_4 vs. V_0) and collective effects (V_5 vs. V_0). Additionally, we examine how previous findings change when they are re-evaluated using a fixed measurement technique (V_{REP} vs. V_{SUG})[2].

[2] V_{REP} refers to the measurement technique used in the original studies. It varies from

The remainder of this section discusses in detail each of the variations from our baseline variation.

2.3.1 Interval length

The baseline variation 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 participants 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 sharing the constant fundamental value property of the first definition.

2.3.2 Bid-ask spread

In the baseline variation, intervals with no transaction prices are dropped from the analysis. Alternatively, this item uses 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 set as the geometric mean of the highest bid and lowest ask prices within an interval.

2.3.3 Aggregation function

As described above, we consider both *Overpricing* and *Absolute mispricing*. For each type of mispricing, we use two different aggregation functions. Table 2 summarizes the relevant mispricing formulae. The baseline measures use the arithmetic mean to aggregate across intervals, whereas the alternative measures emply the geometric mean. The actual formula for a given variation depends

study to study, but usually tends to be quite close to our baseline V_0 variation. The alternative we suggest, V_{SUG}, is closer to V_5, and hence in some sense we maximize the difference between the original V_{REP} and suggested V_{SUG} variations.

Table 2: Mispricing formulae

Mispricing	Mean	Name	Formula
Overpricing	Arithmetic	RD	$= \frac{1}{T}\sum \frac{p_t - v_t}{v_t}$
	Geometric	GD	$= \prod \left(\frac{p_t - v_t}{v_t}\right)^{1/T}$
Absolute	Arithmetic	RAD	$= \frac{1}{T}\sum \left\lvert\frac{p_t - v_t}{v_t}\right\rvert$
	Geometric	GAD	$= \prod \left\lvert\frac{p_t - v_t}{v_t}\right\rvert^{1/T}$

$p_t > 0$ and $v_t > 0$ are the price and fundamental value in interval $t \in 1, \ldots, T$, respectively.

on the type of aggregation function and mispricing being measured (absolute vs. overpricing).

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. The baseline variation does not take into account any of these differences. As an alternative, we construct an adjusted measure of mispricing m':

$$m' = m - bx$$

where m is the original mispricing given by one of the formulae in Table 2, x is the set of characteristics and b the corresponding coefficient estimates[3].

For the set of market characteristics x, the marginal effects b are estimated using a regression of the form:

[3]When testing treatment effects, we exclude from the set of regressors used to calculate m' any variable associated with the treatment. For example, if the treatment relates to experience level of subjects, then this variable is omitted from the set of characteristics that are controlled for (although it is still included in the regression).

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

where $m_{i,j}$ is the unadjusted mispricing for market i from study-treatment j, $x_{i,j}$ its characteristics and $e_{i,j}$ 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.

The characteristics we use are summarized in Table 3. For each characteristic, we briefly describe its construction and previous research (if any) regarding its effect on mispricing.

Mispricing: $\log RD$**,** $\log RAD$**,** $\log GD$**,** $\log GAD$

We take logs to correct for the non-linear nature of these variables (the same applies to the variables FV and RAS below).

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

Recall that our definition of an experimental asset market simply consists of two generic assets A and B that are traded for one another among a set of traders. Let both expected asset returns, which are the same for all traders, be expressed in units of A per unit of B. At any point in time $t \in 1, ..., T$, the fundamental value v_t:

$$v_t = r_t^B / r_t^A$$

is the rate of exchange that equalizes the expected returns r_t^A, r_t^B to holding an equivalent investment position in each of the assets A and B from t until the end of the market. 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, $FV = v_1, ..., v_T$.

Table 3: Variables used in regressions

Variable	Effect (from prev. res.) on		Mean	Std. Dev.
	OP	AMP		
Dependent variables				
V_4: $\log RD$	n/a	n/a	0.3	0.7
V_4: $\log RAD$	n/a	n/a	0.5	0.6
V_5: $\log GD$	n/a	n/a	0.3	0.6
V_5: $\log GAD$	n/a	n/a	0.7	0.7
V_{REP}: OP	n/a	n/a	5.3	29.2
V_{REP}: AMP	n/a	n/a	11.7	36.8
V_{SUG}: $\log GD$	n/a	n/a	0.3	0.7
V_{SUG}: $\log GAD$	n/a	n/a	0.8	1.2
Independent variables				
$E(\log FV)$	$-$	n/a	4.6	0.8
$E(\log RAS)$	$+$	n/a	2.0	4.3
$\lvert E(\log FV)\rvert$	n/a	$-$	4.6	0.8
$\lvert E(\log RAS)\rvert$	n/a	$+$	2.1	4.3
$s.d.(\log FV)$		$+$	6.3	22.2
$s.d.(\log RAS)$		$+$	1.0	2.8
$NSUBJ$	0	0	9.2	1.5
DUR	?	?	0.6	0.3
EXP_{mkts}	$-$	$-$	0.5	1.1
EXP_{dur}	$-$	$-$	0.4	0.4

$s.d.(.)$ and $E(.)$ refer to the standard deviation and expected value over intervals, respectively. The Mean and Std. Dev. headings refer to means and standard deviations *over markets*.

The measures of mispricing we study explicitly control for the relative level of FV, however it is still possible for the nominal FV level to have an effect on mispricing (Noussair et al., 2012). 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:

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

where A_t and B_t are the total quantity in the market at interval t of the two assets, respectively.

Kirchler et al. (2012) shows that an important determinant of relative prices is the relative supply of assets in the market (in the standard environment where "shares" are traded for "cash", this is referred to as the "cash-to-asset ratio"). In particular, assets that are in relatively high (low) supply tend to be under- (over-) priced. Therefore high RAS creates more overpricing. RAS may clearly vary across designs, but even within a particular design it may vary, due to (for example) the realization of stochastic dividend payments.

Duration (DUR**)**

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

Number of traders ($NSUBJ$**)**

$NSUBJ$ is the number of human participants in the market, regardless of whether they act independently of one another or not.

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 variable EXP_{mkts}, which is the average number

of markets that a trader had previously participated in within the same study. 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 markets (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

For practical reasons we consider only peer-reviewed studies published from 2005-2015 inclusive. Table 4 shows the list of 28 studies that satisfy our inclusion criteria and for which we have data (Table 9 in the Appendix lists the 21 studies that satisfy our inclusion criteria but for which we do not have data). From these studies, we compile a set of 878 market observations (from 142 different treatments) and 142 hypotheses related to treatment differences (76 for absolute mispricing, 66 for overpricing) that are tested using a standard two-sided Mann-Whitney test procedure.

It is not possible to calculate all variations for all studies. For example, some studies do not use or report bids and asks. For this reason the number of observations for each variation differs. Each comparison of variations only includes data from those studies which are present in both variations.

4 Results

We present two different types of results. First we compare the individual and collective effect of the variations to a baseline procedure. This answers the question of how sensitive mispricing results are to the dimensions of these variations. Second, we compare mispricing as it is reported in the original study to a particular mispricing procedure that, given data limitations, is as similar as possible to our V_5 variation. We choose V_5 as the comparison variation because most studies use procedures close to (but not always equivalent to) V_0.

Table 4: 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
Corgnet et al. (2010)	30	10	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	2
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. (2015)	46	8	20
Total	878	142	142

Comparing original results to those under V_5 allows us to study the robustness of previous results to as broad a change in measurement technique (given the alternatives we consider in this paper) as possible.

4.1 Variation effects

Under each mispricing variation, we calculate the probability of rejecting each of the null hypotheses based on a two-sided Mann-Whitney test. Then we compare the set of p-values under each of our variations (V_1-V_5) to those from a benchmark (see second row of Table 5). For each comparison, we report:

1. the average change in p-value,
2. the percentage of p-values that switch from being significant to insignificant, where a value is significant if it is below a threshhold value of 0.05, and
3. the percentage of p-values that switch from being insignificant to significant, again at the threshhold value of 0.05.

Our results are presented in Table 5. The columns present each pair of variations that are being compared. The table is divided into two sections, with absolute mispricing on top and overpricing on the bottom. Each section lists by row 1) the average change in p-values, 2) the proportion of tests that switch from being insignificant to significant at the 5% level, 3) again for significant to insignificant, 4) the sum of both types of reversals, and 5) the number of observations.

Compared to a baseline of V_0 (columns 1-5), the results suggest that a small interval size and the geometric mean have small effects on test outcomes - between 3.9-10.6% of p-values change significance. The use of the bid-ask spread has no effect, which is not surprising since under V_0 almost all intervals contain transactions. Adjusting for observable market characteristics (V_4) switches the significance of by far the largest proportion of hypotheses results (16.1-23.0%). Including all variations (V_5) has a slightly larger effect, but the magnitudes (26.3-33.3%) are comparable to those from V_4[4].

[4]These findings are robust to a number of alterations. Results are robust to 1) changing the threshhold significance level to 10% or 1%, 2) excluding call markets, and 3) dropping one of the experience variables. Detailed results available upon request.

Table 5: Effects of variations on mispricing

variation	V_1	V_2	V_3	V_4	V_5	V_{SUG}
baseline	V_0	V_0	V_0	V_0	V_0	V_{REP}
description	int. size	bid-ask	geometric	adjust	all	suggest
	(1)	(2)	(3)	(4)	(5)	(6)
Absolute mispricing						
avg. ch.	0.79	0.13	4.41	3.25	3.47	5.38
insig. → sig.	1.96	0.00	2.63	6.55	2.04	7.89
sig. → insig.	1.96	0.00	5.26	16.39	14.28	18.42
total ch. sig.	3.92	0.00	7.89	22.95	16.32	26.31
N	51	51	76	61	49	76
Overpricing						
avg. ch.	1.46	0.00	2.83	8.43	6.68	7.45
insig. → sig.	5.55	0.00	6.06	0.00	14.00	16.66
sig. → insig.	3.70	0.00	4.54	16.12	14.00	16.66
total ch. sig.	9.25	0.00	10.60	16.12	28.00	33.33
N	54	54	66	62	50	66

"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. → sig."), from significant to insignificant ("sig. → insig."), or the sum of both types of switches ("total ch. sig."). All switches are for the given threshhold significance level of 0.05. Total number of hypotheses for each type of mispricing are 76 for absolute mispricing and 66 for overpricing.

Thus, for the most part the hypotheses we study exhibit are found to be robust to variations $V1 - V5$. Nevertheless, up to 33% are found to be affected by variations in the measurement of mispricing, mostly due to adjustments for the characteristics of individual markets.

4.2 Market characteristics

Next we consider the regression results from the V_4 and V_5 variations which control for market characteristics.

The results are presented in Table 6. Each type of mispricing (overpricing and absolute mispricing) consists of two columns, depending on which types of interval size, bid-ask spread and aggregation procedure are used. The first column (V_4) uses the baseline values from V_0: 1) original interval size, 2) no bid-ask spread and 3) arithmetic mean; the second variation (V_5) uses a combination of V_1, V_2 and V_3: 1) smallest interval size possible, 2) bid-ask spread when no transactions, and 3) geometric mean.

Under V_4, both forms of mispricing are found to be decreasing in the level of FV, and weakly increasing in the duration of the market. Additionally, absolute mispricing decreases with both variation in FV and duration experience.

In contrast, the results for V_5 differ substantially from those for V_4. The level of FV is no longer signficant, while the level of the relative asset supply is found to have a small positive effect on overpricing. Variation in FV now *increases* absolute mispricing, whereas variation in RAS tends to decrease it. The coefficient on market duration for absolute mispricing remains positive, but is no longer significant. The most consistent finding across both variations is that experience tends to decrease both types of mispricing: duration experience for overpricing, and both types of experience for absolute mispricing.

4.3 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 change when analyzed under what we term *suggested* conditions.

Table 7 shows the hypotheses whose significance changes (at the 5% level) when moving from the specification reported in the original study (V_{REP}) to our

Table 6: Market characteristics and mispricing regressions

mispricing	Overpricing		Absolute mispricing	
variation	V_4	V_5	V_4	V_5
regressor				
$E(\log FV)$	-0.404***	-0.161		
	(0.145)	(0.342)		
$E(\log RAS)$	0.042	0.328*		
	(0.029)	(0.179)		
$\lvert E(\log FV)\rvert$			-0.480**	0.491
			(0.214)	(0.347)
$\lvert E(\log RAS)\rvert$			-0.003	-0.028
			(0.017)	(0.101)
$s.d.(\log FV)$	-0.007	0.241	-0.015*	1.745**
	(0.005)	(0.636)	(0.009)	(0.715)
$s.d.(\log RAS)$	-0.041	-0.199	0.000	-0.439*
	(0.026)	(0.285)	(0.014)	(0.249)
$NSUBJ$	0.033	-0.027	0.022	0.016
	(0.040)	(0.033)	(0.037)	(0.080)
DUR	0.915*	0.056	1.448*	1.266
	(0.500)	(0.192)	(0.783)	(0.879)
EXP_{mkts}	-0.150	-0.013	-0.009	-0.054**
	(0.160)	(0.031)	(0.052)	(0.026)
EXP_{dur}	0.025	-0.223***	-0.236***	-0.312***
	(0.168)	(0.084)	(0.068)	(0.071)
Study-treatment dummies	Yes	Yes	Yes	Yes
N	656	354	656	354

OLS regression of mispricing explained by study-treatment dummies and market characteristics. Errors are clustered at the study-treatment level.

suggested specification (V_{SUG}). Results for all hypotheses, whether affected or not, are reported in the Appendix. The original specification varies from study to study, whereas our suggested specification V_{SUG} is as close to V_5 as possible, given the available data. In particular, V_{SUG} uses a two-sided Mann-Whitney test to compare mispricing calculated 1) using as small an interval as possible, 2) when possible, using the bid-ask spread as a proxy for price in intervals with no transactions, 3) taking the geometric mean across intervals, and 4) adjusting for market characteristics based on the results of the V_5 regressions.

Table 7: Affected hypotheses

study	category	comparison
Originally significant, now insignificant		
Dufwenberg et al. (2005)	OP	$R1 - R4$
Noussair and Powell (2010)	AMP	$P3 - V3$
		$P4 - V4$
	OP	$P3 - V3$
		$P4 - V4$
Kirchler et al. (2012)	AMP	$T1T3 - T2T4$
		$T1 - T2$
		$T3 - T4$
		$T1 - T5$
	OP	$T1T2 - T3T4$
		$T1T3 - T2T4$
		$T1 - T2$
Cheung et al. (2014)	AMP	$1PK - 2NPK$
Lugovskyy et al. (2014)	AMP	$G1 - G3$
Stöckl et al. (2014)	OP	$R1 - R2$
		$R1 - R3$
		$R1 - R4$

	AMP	$R1 - R3$
		$R1 - R4$
		$R4 - R5$
Cueva and Rustichini (2015)	OP	$T2MALE - T2HETEROG$
Cason and Samek (2015)	AMP	$PreTextM3 - PreVisualM3$
		$PreTextM3 - TextM3$
Eckel and Füllbrunn (2015)	OP	$M - Mix$
Huber et al. (2015)	AMP	$T1 - T4$
Originally insignificant, now significant		
Cheung and Palan (2012)	OP	$DA2H - DAIND$
	AMP	$DA2H - DAIND$
Kirchler et al. (2012)	AMP	$T1T2 - T3T4$
	OP	$T2 - T4$
		$T1 - T6R1$
		$T5 - T6R2$
		$T6R1 - T6R2$
Fischbacher et al. (2013)	OP	$E123P0 - E123P1$
	AMP	$E123P0 - E123P1$
	OP	$E2P0 - E4P1$
		$E2P1 - E4P2$
Stöckl et al. (2014)	OP	$R2 - R4$
		$R2 - R5$
Cueva and Rustichini (2015)	AMP	$T2MALE - T2HETEROG$
Cason and Samek (2015)	AMP	$TextM1 - VisualM1$
		$PreTextM2 - TextM1$
Eckel and Füllbrunn (2015)	OP	$Mix - F$

Table shows the hypotheses whose significance changes (at the 5% level) when moving from the specification reported in the original study V_{REP} to our suggested specification (V_{SUG}). The original specification varies from study to study, whereas our suggested specification is as close to V_5 as possible, given the data at hand.

Overall, almost a third (42 out of 142) 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.

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, when measuring mispricing under V_{SUG}, including controling for the level and variation in the fundamental value and relative asset supplies, the treatment difference is no longer significant. This suggests that the path of fundamental value is less important than originally thought. Similar conclusions apply to some of the affected hypotheses from Kirchler et al. (2012) and Stöckl et al. (2014).

Second, Cheung and Palan (2012) is an example of results that switch to being significant under the alternative measurement variation. This study compares markets populated by individual traders ("IND") to markets composed of teams of two traders ("2H"). The two types of double auction ("DA") markets appear similar apart from the main treatment difference, however upon further inspection several other differences emerge (level and variation in relative asset supply, and the duration and hence experience of the market). After controlling for these differences, and using the alternative interval length and aggregation techniques, the main treatment effect turns out to be strongly significant for both absolute mispricing and overpricing.

Therefore, although our results show that a majority of results are not affected by the change in measurement specification, a substantial minority (30%) 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, we looked at the individual and aggregate effect of variations in mea-

surement procedure. Our results show that the choice of interval size and aggregation procedure have a limited effect compared to controlling for the characteristics of the market. The usage of the bid-ask spread has no effect when using conventional interval lengths.

Second, we have examined how much actual results change when re-evaluating hypotheses from various studies under a fixed measurement specification. Our results are of the "*glass half-full, glass half-emtpy*" genre. On the one hand, it is reassuring that a majority of results (70%) do not change significance under the new specification. However, this still leaves a substantial minority (30%) 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 specifications (we have suggested one particular specification, 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 are themselves sensitive to the data and specification used. We hope to update these results as newly published data becomes available.

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 and Quantitative Analysis*, 44(03):719–744.

Ackert, L. F., Mazzotta, S., and Qi, L. (2011). An Experimental Investigation of Asset Pricing in Segmented Markets. *Southern Economic Journal*, 77(3):585–598.

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.

Baghestanian, S. and Walker, T. B. (2015). Anchoring in experimental asset markets. *Journal of Economic Behavior & Organization*, 116:15–25.

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.

Caginalp, G. and Ilieva, V. (2008). The dynamics of trader motivations in asset bubbles. *Journal of Economic Behavior & Organization*, 66(3):641–656.

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.

Chan, K. S., Lei, V., and Vesely, F. (2013). Differentiated assets: An experimental study on bubbles. *Economic Inquiry*, 51(3):1731–1749.

Cheung, S. L. and Coleman, A. (2014). Relative performance incentives and price bubbles in experimental asset markets. *Southern Economic Journal*, 81(2):345–363.

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.

Childs, J. (2009). Rate of return parity and currency crises in experimental asset markets. *Journal of International Financial Markets, Institutions and Money*, 19(1):157–170.

Childs, J. and Mestelman, S. (2006). Rate-of-return parity in experimental asset markets. *Review of International Economics*, 14(3):331–347.

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.

Corgnet, B., Kujal, P., and Porter, D. (2010). The effect of reliability, content and timing of public announcements on asset trading behavior. *Journal of Economic Behavior & Organization*, 76(2):254–266.

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.

Deck, C., Porter, D., and Smith, V. (2014). Double bubbles in assets markets with multiple generations. *Journal of Behavioral Finance*, 15(2):79–88.

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. *The 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 and Control*, 37(10):2104–2122.

Ghosh, S., Radhakrishnan, S., Srinidhi, B., and Su, L. N. (2015). Recognition of future news in earnings and price bubbles in experimental asset markets. *Journal of Accounting, Auditing & Finance*, 30(4):558–575.

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. *The 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.

Hauser, F. and Huber, J. (2012). Short-selling constraints as cause for price distortions: An experimental study. *Journal of International Money and Finance*, 31(5):1279–1298.

Huber, J. and Kirchler, M. (2012). The impact of instructions and procedure on reducing confusion and bubbles in experimental asset markets. *Experimental Economics*, 15(1):89–105.

Huber, J., Kirchler, M., and Stöckl, T. (2015). The influence of investment experience on market prices: Laboratory evidence. *Experimental Economics*, pages 1–18.

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. R., Smith, V. L., Williams, A. W., and Boening, M. V. (1993). The Robustness of Bubbles and Crashes in Experimental Stock Markets. pages 183–2999.

Kirchler, M. and Huber, J. (2009). An exploration of commonly observed stylized facts with data from experimental asset markets. *Physica A: Statistical Mechanics and its Applications*, 388(8):1631–1658.

Kirchler, M., Huber, J., and Kleinlercher, D. (2011). Market microstructure matters when imposing a tobin taxâĂŤevidence from the lab. *Journal of economic behavior & organization*, 80(3):586–602.

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

Kleinlercher, D., Huber, J., and Kirchler, M. (2014). The impact of different incentive schemes on asset prices. *European Economic Review*, 68:137–150.

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

Lei, V. and Vesely, F. (2009). Market efficiency: Evidence from a no-bubble asset market experiment. *Pacific Economic Review*, 14(2):246–258.

Levine, S. S., Apfelbaum, E. P., Bernard, M., Bartelt, V. L., Zajac, E. J., and Stark, D. (2014). Ethnic diversity deflates price bubbles. *Proceedings of the National Academy of Sciences*, 111(52):18524–18529.

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. and Tucker, S. (2006). Futures markets and bubble formation in experimental asset markets. *Pacific Economic Review*, 11(2):167–184.

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, A., Lohrenz, T., King, J., Montague, P. R., and Camerer, C. F. (2014). Irrational exuberance and neural crash warning signals during endogenous experimental market bubbles. *Proceedings of the National Academy of Sciences*, 111(29):10503–10508.

Smith, V. L. (2014). New insights into old discoveries: Two kinds of markets. *International Journal of the Economics of Business*, 21(1):33–35.

Smith, V. L., Suchanek, G. L., and Williams, A. W. (1988). Bubbles, crashes, and endogenous expectations in experimental spot asset markets. *Econometrica: Journal of the Econometric Society*, pages 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.

Stöckl, T. and Kirchler, M. (2014). Trading behavior and profits in experimental asset markets with asymmetric information. *Journal of Behavioral and Experimental Finance*, 2:18–30.

Sutter, M., Huber, J., and Kirchler, M. (2012). Bubbles and information: An experiment. *Management Science*, 58(2):384–393.

Appendix

Table 8: Adjusted mispricing formulae

Measure	Correction	Original formula	Corrected formula
GD	$\lvert\log(GD+1)\rvert$	$-1+\exp\frac{1}{T}\sum\log\frac{p_t}{v_t}$	$\frac{1}{T}\left\lvert\sum\log\frac{p_t}{v_t}\right\rvert$
GAD	$\log(GAD+1)$	$-1+\exp\frac{1}{T}\sum\left\lvert\log\left(\frac{p_t}{v_t}\right)\right\rvert$	$\frac{1}{T}\sum\log\left\lvert\frac{p_t}{v_t}\right\rvert$

Table 9: Studies which satisfy our selection criteria but for which data is either not available or incomplete

study
Childs and Mestelman (2006)
Noussair and Tucker (2006)
Caginalp and Ilieva (2008)
Childs (2009)
Kirchler and Huber (2009)
Lei and Vesely (2009)
Ackert et al. (2011)
Kirchler et al. (2011)
Sutter et al. (2012)
Hauser and Huber (2012)
Huber and Kirchler (2012)
Noussair et al. (2012)
Chan et al. (2013)
Deck et al. (2014)
Kleinlercher et al. (2014)
Levine et al. (2014)
Cheung and Coleman (2014)
Smith (2014)
Smith et al. (2014)
Baghestanian and Walker (2015)
Ghosh et al. (2015)

Table 10: Hypotheses

study	category	comp.	V_{REP}	V_{SUG}	change
Dufwenberg et al. (2005)	AMP	$R1 - R4$	0.032	0.002	|*
		$R3 - R4$	0.061	0.481	*|
		$R4_23 - R413$	0.421	0.420	|
	OP	$R1 - R4$	0.011	0.247	**|
		$R3 - R4$	0.897	0.970	|
		$R4_23 - R413$	0.310	1.000	|
Noussair and Powell (2010)	AMP	$P1 - V1$	0.347	0.309	|
		$P2 - V2$	0.175	0.309	|
		$P3 - V3$	0.047	0.547	**|
		$P4 - V4$	0.028	0.222	**|
	OP	$P1 - V1$	0.465	0.841	|
		$P2 - V2$	0.175	0.309	|
		$P3 - V3$	0.047	0.690	**|
		$P4 - V4$	0.028	0.547	**|
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	0.008	|***
	AMP	$DA2H - DAIND$	0.078	0.002	|**
Schoenberg and Haruvy (2012)	OP	$DOWN - UP$	0.010	0.001	|
Kirchler et al. (2012)	AMP	$T1T2 - T3T4$	0.550	0.033	|**
		$T1 - T3$	0.550	0.093	|*
		$T2 - T4$	0.550	0.240	|
		$T1T3 - T2T4$	0.005	0.377	***|
		$T1 - T2$	0.030	0.818	**|
		$T3 - T4$	0.005	0.064	**|

		$T1-T5$	0.025	0.179	**|
		$T1-T6R1$	0.037	0.025	|
		$T1-T6R2$	0.007	0.002	|
		$T5-T6R1$	0.522	0.937	|
		$T5-T6R2$	0.631	0.484	|
		$T6R1-T6R2$	0.550	0.240	|
	OP	$T1T2-T3T4$	0.030	0.265	**|
		$T1-T3$	0.550	0.393	|
		$T2-T4$	0.550	0.002	***
		$T1T3-T2T4$	0.030	0.932	**|
		$T1-T2$	0.005	0.132	***|
		$T3-T4$	0.550	0.240	|
		$T1-T5$	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	0.041	**
		$T6R1-T6R2$	0.550	0.002	***
Fischbacher et al. (2013)	OP	$E1P0-E1P1$	0.106	0.210	|
		$E123P0-E123P1$	0.062	0.001	**
	AMP	$E123P0-E123P1$	0.067	0.003	**
	OP	$E2P0-E4P1$	0.093	0.025	*
		$E2P1-E4P2$	0.093	0.002	**
Corgnet et al. (2015)	AMP	$EM-HM$	0.130	0.089	*
Cheung et al. (2014)	AMP	$1PK-4BASE$	0.003	0.006	|
		$2NPK-4BASE$	0.088	0.105	*|
		$1PK-2NPK$	0.033	0.063	*|
	OP	$1PK-4BASE$	0.335	0.314	|
		$2NPK-4BASE$	0.066	0.123	*|

		$1PK - 2NPK$	0.099	0.314	*|
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	|
Lugovskyy et al. (2014)	OP	$G1 - G2$	0.030	0.003	|*
		$G2 - G3$	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	0.148	**|
Stöckl et al. (2014)	OP	$R1 - R2$	0.005	0.393	***|
		$R1 - R3$	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	0.015	|
		$R3 - R5$	0.075	0.064	|
		$R4 - R5$	0.550	0.588	|
	AMP	$R1 - R2$	0.005	0.002	|
		$R1 - R3$	0.030	0.393	**|
		$R1 - R4$	0.030	0.937	**|
		$R1 - R5$	0.550	1.000	|
		$R2 - R3$	0.005	0.002	|
		$R2 - R4$	0.005	0.002	|

		$R2 - R5$	0.005	0.002	|
		$R3 - R4$	0.550	0.064	|*
		$R3 - R5$	0.550	0.132	|
		$R4 - R5$	0.005	0.937	***|
Breaban and Noussair (2015)	AMP	$DECR_M1 - INCR_M1$	0.550	0.427	|
		$DECR_M2 - INCR_M2$	0.550	0.792	|
		$DECR_M1 - DECR_M2$	0.378	0.143	|
		$INCR_M1 - INCR_M2$	0.065	0.588	*|
	OP	$DECR_M1 - INCR_M1$	0.550	0.367	|
		$DECR_M2 - INCR_M2$	0.550	0.874	|
		$DECR_M1 - DECR_M2$	0.550	0.684	|
		$INCR_M1 - INCR_M2$	0.550	0.818	|
Cueva and Rustichini (2015)	OP	$T1FEMALE - T2MALE$	0.199	0.190	|
		$T1FEMALE - T2HETEROG$	0.199	0.911	|
		$T2MALE - T2HETEROG$	0.023	0.075	*|
		$T1HOMOG - T2HETEROG$	0.333	0.350	|
	AMP	$T1FEMALE - T2MALE$	0.450	0.528	|
		$T1FEMALE - T2HETEROG$	0.879	0.217	|

		$T2MALE - T2HETEROG$	0.070	0.023	|*
		$T1HOMOG - T2HETEROG$	0.039	0.039	|
Cason and Samek (2015)	AMP	$TextM1 - VisualM1$	0.937	0.008	|***
		$TextM2 - VisualM2$	0.484	0.093	|*
		$TextM3 - VisualM3$	0.588	1.000	|
		$PreTextM2 - PreVisualM2$	0.093	0.179	*|
		$PreTextM3 - PreVisualM3$	0.015	0.093	*|
		$PreTextM2 - TextM1$	0.588	0.041	|**
		$PreTextM3 - TextM2$	0.393	0.240	|
		$PreTextM2 - TextM2$	0.093	0.132	*|
		$PreTextM3 - TextM3$	0.025	0.309	**|
		$PreVisualM2 - VisualM1$	0.064	0.484	*|
		$PreVisualM3 - TextM2$	0.484	0.937	|
		$PreVisualM2 - TextM2$	0.937	1.000	|
		$PreVisualM3 - VisualM3$	0.240	0.937	|
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. (2015)	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	|

Table shows how originally reported p-values (column V_{REP}) of hypotheses differ those calculated under our suggested specification (column V_{SUG}). The original specification varies from study to study, whereas our suggested specification is as close to V_5 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 our suggested variation.