Anchoring in Economics: A Meta-Analysis of Studies on Willingness-To-Pay and Willingness-To-Accept*

Lunzheng Li[†], Zacharias Maniadis[‡], Constantine Sedikides[§]

Abstract

Anchoring is considered one of the most robust psychological phenomena in judgment and decision-making. Earlier studies produced strong and consistent evidence that anchoring is relevant for the elicitation of economic preferences, but subsequent studies found weaker and less consistent effects. We examined the economic significance of numerical anchoring by conducting a meta-analysis of 53 studies. We used the Pearson correlation coefficient between the anchor number and target response (in our case, Willingness-to-Pay and Willingness-to-Accept) as the primary effect size. Both fixed-effects and random-effects models pointed to a moderate overall effect, smaller than the effects reported in early studies. Given some well-known limitations of our meta-analytic methodology, these results should be viewed with caution and the effect size as an upper bound. Also, meta-regression analysis indicates that non-random anchors and non-laboratory experiments were associated with higher anchoring effects, whereas selling tasks and anchors incompatible with the evaluated item were associated with lower (but often non-significant) anchoring effects. The use of financial incentives did not have a discernible effect.

Keywords: Anchoring, Willingness to Pay, Meta-Analysis

JEL Classification: A12, C83, D03

*We thank authors who send their data.

[†]Shenzhen University, Shenzhen Audencia Business School, email: lunzheng.li@szu.edu.cn

[‡]University of Southampton, Department of Economics, email: Z.Maniadis@soton.ac.uk, corresponding author.

[§]University of Southampton, Department of Psychology, email: cs2@soton.ac.uk

1 Introduction

Anchoring is generally defined as the influence of a normatively irrelevant cue on a subsequent expression of judgment. The cue and the judgment can be numerical. Since the influential work of Tversky and Kahneman (1974), numerical anchoring has been considered one of the most robust psychological phenomena in judgment and decision-making. In Tversky and Kahneman's pioneering study, a roulette wheel delivered a random number, based on which participants were asked a binary question about some unknown quantity. For instance, 'does the average temperature in Antarctica exceed the random number drawn from the roulette wheel'? Subsequently, participants made judgments about the actual magnitude of the given variable (in this case, the average temperature in Antarctica). The elicited numerical judgments were greatly affected by the initial binary question. In particular, if a participant drew a large random number, they also expressed higher magnitudes in their numerical judgment tasks.

Anchoring belongs to the domain of behavioral research termed 'heuristics and biases' by Tversky and Kahneman (1974), in which consumers deviate systematically from the benchmark of rational economic behavior. A key question is the economic importance of anchoring. A fundamental postulate of economic theory is the concept of consumer preferences, which shape economic behavior and constitute the foundation of demand and thus market prices. Demand is the expression of Willingness-to-Pay (WTP) for economic goods. Anchoring matters in the elicitation of economic preferences, as demonstrated early on by Northcraft and Neale (1987), Green et al. (1998), and Ariely et al. (2003). Ariely et al. employed the prototypical design of Tversky and Kahneman (1974) to elicit the WTP for a series of consumer goods as well as Willingness-to-Accept (WTA) in regards to simple negative hedonic experiences.[1]

The prototypical design of anchoring in economic evaluation consists of three stages. The first stage is what we call the *anchoring manipulation*. Here, the experimenter shows participants a good and asks: "Would you purchase (sell) this object for 'X' Dollars?" where 'X' is the numerical anchor. The second stage is the *elicitation* of the participant's economic valuation. Here, to elicit Willingness-to-Pay (WTP), the experimenter might ask a question such as "What is the maximum dollar amount would you be willing to pay for this object?". The last stage typically provides incentives for truthfully revealing the valuation, such as the Becker-

[1]Such experiences may entail hearing annoying but harmless sounds through headphones, or drinking bad-tasting but harmless liquids.

Degroot-Marschak Mechanism that emulates a real market.[2] A large number of subsequent studies followed this prototypical design, and most of them included the first and second stage. Few studies included the third, or incentivization, stage.

As a consequence of these early results, anchoring has been considered not only relevant for economic preferences, but also highly robust across several dimensions, with the magnitude of the effect believed to be large (Strack and Mussweiler, 1997). However - perhaps because of the robustness of the phenomenon in the psychology literature - the exact economic magnitude of anchoring has not been paid sufficient attention. Most subsequent experiments reported weaker and less robust anchoring than earlier ones, but the literature has not been synthesized into a statement about the quantitative economic significance of numerical anchoring.

Assessing the evidence and quantifying the effect of anchors on consumers' economic valuation entails several benefits. To begin, an assessment of whether the effect is large enough to be a concern for the contingent valuation methodology would be useful for methodological appraisals of this methodology as well as for public policy aiming to protect consumers. Marketers would also like to know how malleable their customers' WTP is. Of course, information about factors that moderate the anchoring effect would be equally important. Some key questions from an economic point of view are: What types of anchors amplify anchoring? Do market forces ameliorate anchoring? From a methodological point of view, economists might ask whether certain methodologies (such as monetary incentives) are associated with higher or lower anchoring effects.

Examining factors that affect the magnitude of anchoring may illuminate its causes. The most well-known theory of anchoring is 'anchoring and adjustment' (Tversky and Kahneman, 1974), according to which people consider the arbitrary cue as a possible answer to the evaluation question. They treat the anchor as a starting point and then adjust, but the adjustment is always insufficient. Another well-known approach is the selective accessibility model (Mussweiler et al., 2000), according to which the anchoring manipulation increases the accessibility of anchor-consistent information that is used for the later evaluation. Although not our main focus, we will comment on both theories in our moderator analysis.

[2]Let us illustrate how this mechanism works in WTP tasks. A random price is drawn and applies for all participants. If a participant's stated numerical evaluation is less than this price, no transaction takes place. If it exceeds this price, the participant purchases the given good and pays the randomly drawn price. This provides incentives for participants to reveal honestly their evaluation.

We addressed the above issues by conducting a systematic synthesis of studies that examine the effects of numerical anchors on statements of economic valuation. We aimed to systematize the measurement of anchoring quantitatively and to examine the determinants of variability in the magnitude of its effect size. We included 53 studies from 24 articles, and chose the Pearson correlation coefficient between the anchor number and target response (in our case, WTP/WTA) as the primary effect size. Both fixed-effects and random-effects models point to a moderate overall effect. Regarding the important problem of publication bias, whereas a mere visual inspection of the funnel plots indicates potential asymmetry problems, formal statistical tests do not confirm this.[3] Further meta-regression analysis shows that participants in WTP tasks are more likely to be influenced by anchoring, compared to participants in WTA tasks. Incentives do not attenuate the effects. Also, the relevance and compatibility of the anchor to the target response matter for the magnitude of the effect size. Interestingly, studies whose data became available to us as well as studies published in recent years seem to be associated with smaller effect sizes.

We provide an overview of the rest of the article. In Section 2, we discuss our design and methodological choices, in particular our literature search as well as the choice of effect size measure and of moderator variables. In Section 3, we report the standard meta-analytic results that include the overall effect size and meta-regression results. In Section 4, we examine the robustness of the results. We conclude in Section 5.

2 Methods

2.1 Effect Sizes

A key methodological choice involves the main effect size measure. The principal effect size that we analyzed is the Pearson correlation coefficient (r) between anchor number and target response (i.e., WTP/WTA). We chose this effect size for its intuitive interpretation, because it is reported in several studies in the included literature, and because it is a standard

[3]Notice that the practice of making inferences on publication bias based on visual inspection of funnel plots has been criticized as arbitrary and non-robust to measurement choices (Ioannidis and Trikalinos, 2005; Stanley, 2008).

meta-analytic measure (Cooper et al., 2009).[4]

For studies where the raw data were unavailable, we extracted the effect sizes using information reported in the articles. We used standard meta-analytic methodology for translating reported effect size measures into alternative ones (Borenstein et al., 2009), following the guidelines of Cooper et al. (2009) (pp. 224-234) for transforming the reported measures to r. We also used the Campbell Collaboration online effect size calculator (Wilson, 2001) as a complementary tool.[5]

Let us introduce an important term: 'study' represents the 'unit' included in the meta-analysis. The unit can be an experiment or a condition within an experiment. We also wish to make a methodological aside. In many studies, authors reported the elicitation of valuations for multiple goods. Also, several designs were within-subjects, implying that often a given participant was faced with multiple anchors. Given that the psychological processing of multiple anchors is complex and distinct from the processing of a single anchor (Whyte and Sebenius, 1997), we opted to focus on the effect of single anchors. This practice necessitates pinpointing the data that originated from exposure to a single anchor, and including only those in the analysis. Such practice, however, was not always possible due to limitations in the information contained in each article. We will return to this issue.

2.2 Literature Search and Inclusion Criteria

We retrieved the studies via the following four channels. (1) We searched the Web of Science, Google Scholar, and EconLit databases.[6] (2) We used Web of Science to focus on the references and citations of three early studies (Northcraft and Neale, 1987; Green et al., 1998;

[4]We considered two other measures: Spearman's rank correlation coefficient, and Jacowitz and Kahneman (1995)'s anchoring index. The latter can be defined as follows for a binary between-subjects treatment, where one sample of participants has been exposed to a low anchor value and another sample to a high anchor value: AI = [Median Elicited Valuation (High Anchor) - Median Elicited Valuation (Low Anchor)] / [High Anchor - Low Anchor]. Both of these measures have desirable properties in terms of assessing the magnitude of the anchoring effect. Spearman's rank correlation coefficient captures monotonicity in cases where the effect is not linear, and the anchoring index offers an intuitive descriptive metric. However, we needed raw data to calculate any of these measures, and such data were unavailable for more than half of the included studies (32/53).

[5]For most studies for which raw data were unavailable, we derived the effect size using the aforementioned methods. The only exceptions were studies that only reported the coefficient of a multiple regression. For those, we used the formula provided in Peterson and Brown (2005). The formula is $r = 0.98\beta + 0.5\lambda$, where $\lambda = 1$ if β is non-negative, otherwise $\lambda = 0$.

[6]We used the following search strings in Web of Science: TS = (anchoring AND willingness to pay) OR TS = (anchoring AND willingness to accept) OR TS = (anchoring AND valuation) OR TS = (anchoring AND "WTA") OR TS = (anchoring AND "WTP")) AND LANGUAGE: (English). In EconLit, we searched for similar keywords.

Ariely et al., 2003). (3) We engaged in personal communication with researchers that helped us to trace unpublished articles. (4) Finally, we posted a literature searching advertisement on the ESA Experimental Methods Discussion group.[7]

Next, we screened the articles and included studies on the basis of the following two a priori inclusion/exclusion criteria. First, we included *only articles published in English-speaking journals*. Second, we included studies that *elicited a numerical statement of WTP/WTA for economic goods after an unambiguous and unique numeric anchor was presented*. We provide next some examples of applying the rich second criterion. A numerical statement of WTP/WTA excludes studies like (1) Wansink et al. (1998), who only measured quantities purchased, (2) Jung et al. (2016) who elicited 'pay what you want' that in our view is not a representation of WTP, and (3) Mussweiler et al. (2000), who asked participants to carry out a neutral pricing task.[8] The second criterion also requires an unambiguous and unique numeric anchor. Therefore, we excluded studies with explicit multiple anchors, such as Sugden et al. (2013), where a within-subjects design was used in which a given participants was exposed to different goods and different anchors in a random order. For the same reason (multiple anchors) we excluded studies that used the list method to elicit WTA/WTP, such as Araña and León (2008) and Tufano (2010).[9] However, we included part of the data in studies with multiple anchors, if it was possible to identify the first anchor to which a given participant was exposed. In these cases, we calculated the effect size using the first anchor and the corresponding elicited WTA or WTP, and did not include the data derived using the second and subsequent anchors. Therefore, we included studies such as (1) Bavolár et al. (2017), where participants were presented with different anchors and goods, and it was easy to identify which the first anchor and corresponding good was, and (2) Green et al. (1998), where participants were presented with five numerical evaluation questions with the first one being a WTP evaluation. Our included studies pertained not only to market goods, but also to lotteries (Fudenberg et al., 2012), environmental goods (Green et al., 1998; Schläpfer and Schmitt, 2007), and simple hedonic experiences (Maniadis

[7]The ESA Experimental Methods Discussion group is a Google group for economists to discuss experimental methods in economics, and it is sponsored by the Economic Science Association.

[8]In their experiment, the following question was asked in the elicitation phase: "Could you tell me, what do you think is the approximate price for the car as you see it?" In our judgement, this question does not elicit WTP or WTA, given that the answer could depend on perceived market conditions and other factors that do not directly impact economic evaluation.

[9]The list method asks participants repeatedly (in the elicitation phase) whether they would buy or sell an object for different prices. These prices are all salient at the time of final choice and therefore can serve as anchors. Hence, we decided to exclude these studies.

et al., 2014).

2.3 Moderators

Our methodological objective was to code for theoretically relevant aspects of the design that likely influenced the estimated effect sizes. Apart from the standard moderator variables in meta-analysis, such as sample size, time period of publication, availability of the raw data (unavailable vs. available), country where the study was conducted,[10] participant pool (students vs. general population), and experiment type (laboratory vs. classroom vs. field) we coded the following six moderators: anchor type, task type, incentive type, experiment type, compatibility, and manipulation type. All moderators were categorical variables. We present in Table 1 a summary of the moderators and respective categories for each moderator.

Table 1: Summary of Moderators

Moderators	Categories
Anchor Type	Explicitly random Fixed and provided without explanation Having some relevance with the target
Task Type	WTP WTA
Manipulation Type	Canonical design Non-Canonical design
Subject Pool	Students General population
Incentives Type	Not incentivized Probabilistically incentivized Fully incentivized
Experiment Type	Lab experiment Class experiment Field experiment
Compatibility	Compatible Incompatible
Raw Data Availability	Available Unavailable
Country of Study	USA Non-USA

[10]More than 75% of included experiments were conducted in the USA, while others were conducted in countries such as Sweden, Poland, Italy, etc.

Anchor Type

Experiment 1 of Ariely et al. (2003), and many subsequent replications, used explicitly random anchors, such as the last two digits of a participant's social security number. Other studies provided a fixed anchor number without an explanation of its origin.[11] In several studies, participants were given anchors that have potential relevance for the target. For example, in Bavolár et al. (2017), the anchor was provided as a price paid by a hypothetical person. To summarize, a non-random anchor may convey useful information about the underlying properties of the good, and in the case of a fixed anchor, participants might assume that the anchor number is provided for a reason. Hence, we hypothesized that more informative anchors would be associated with a stronger anchoring effect, and random anchors with a weaker one.

Task Type (WTA vs WTP)

There are key differences in the way people express their WTP and WTA, and there is a noted gap between the two (Kahneman et al., 1991). In pure WTP and WTA tasks, WTA is usually larger than WTP. However, the literature on the possible disparity of the anchoring effect across these two tasks is limited. Simonson and Drolet (2004) suggested that WTP is more susceptible to the anchoring effect compared to WTA. They argued that in WTA tasks sellers set prices based on the market price, which is objective. However, buyers' subjective valuations of the goods play a key role in WTP tasks. Using this theoretical argument, we hypothesized that the effect size for WTA would be smaller than for WTP.

Incentive Type

Another relevant variable is the magnitude of monetary incentives. Using financial incentives is considered a methodological norm in economic experiments. Accordingly, we should expect that provision of incentives facilitates accurate elicitations of economic valuation. There is heterogeneity in our included studies regarding the use of incentives. The majority of studies were not incentivized - especially those that involved expensive goods. Many studies only picked a random participant in a given session for whom one of the choices was consequential.

[11] For instance, in Experiment 2 of Ariely et al. (2003), participants were randomly exposed either to a high anchor of 50 cents or a low anchor of 10 cents, and were asked whether they would be willing to accept the anchor amount in order to listen to annoying sounds of certain duration.

Only a few studies presented at least one incentivized decision for each participant. From the perspective of experimental economics, it has often been argued that behavioral anomalies – such as anchoring effects – will be reduced when financial incentives are higher (Caplan, 2000). Accordingly, we hypothesized that the higher the financial incentives, the lower the anchoring effects would be.

Compatibility

We also coded for a variable that, as per the psychology literature, might be relevant to anchoring. The variable is referred to as compatibility between the anchor and the target, denoting the degree to which the anchor and evaluation stimuli are based on the same dimension and scale (Strack and Mussweiler, 1997). The majority of the included studies express both the anchor and the evaluation in monetary units of the same currency, and thus achieve compatibility.[12] According to the selective accessibility model (Strack and Mussweiler, 1997), studies with designs that satisfy compatibility will result in stronger anchoring, because the information activated in the process of answering the comparative question (*anchoring manipulation*) will be relevant for the value *elicitation* task. Similarly, in the anchoring-and-adjustment theory compatible anchors will be more likely to be used as candidate answers for the elicitation task, so again stronger anchoring is predicted.

Manipulation Type

The manner in which the anchoring manipulation is operationalized is also relevant. In particular, we coded experiments that do not use the *anchoring manipulation* stage followed by the *elicitation* stage (the aforementioned prototypical design described in Section 1) as having a non-canonical design.[13] We hypothesized that, because the *anchoring manipulation* stage

[12]Only a few studies, such as Dogerlioglu-Demir and Koçaş (2015), Schläpfer and Schmitt (2007), and Tanford et al. (2019) involved incompatible anchors. Dogerlioglu-Demir and Koçaş (2015) used as anchors numbers that appear in the name of a given good; for example, participants evaluated an average meal at "Studio 17 versus Studio 97". Schläpfer and Schmitt (2007) used tax rates presented in the form of percentages as anchors. Finally, Tanford et al. (2019) claimed that in one of their treatments the metric for the anchor was incompatible with the one of the stimulus good, given that the value of the good (hotel room) was measured in price per night, while anchors were presented in price per week.

[13]For instance, Yu et al. (2017) simply put a label with the anchor number on the goods and did not ask the comparative question; Northcraft and Neale (1987) and Tanford et al. (2019) gave to participants the listing price for the goods; and Bavolár et al. (2017) introduced a stage where they presented the anchor number, but not in the form of a question.

promotes the possibility that anchor value is an answer at the elicitation stage, it should enhance anchoring effects, in line with both the anchoring-and-adjustment and selective accessibility theories.

We also considered other potentially relevant variables. These were: whether the anchor was plausible (Mussweiler and Strack, 2001), elicitation method for WTP/WTA (an "open-ended" question vs. a form of auction), intensity of emotions - e.g., happiness, sadness -(Araña and León, 2008),[14] and forewarnings about the role of anchoring (Epley and Gilovich, 2005; LeBoeuf and Shafir, 2009). We decided, though, to exclude these variables, because in the process of coding we became aware of lack of heterogeneity.

In regards to this decision, Thompson and Higgins (2002) emphasized that the results of meta-regression analysis will necessarily be correlational, not causal, and warned against 'data dredging', namely, examining multiple models and post-hoc theorizing. This is particularly problematic in meta-analysis, because it uses the totality of the evidence and thus it is not possible to validate a model with out-of-sample predictions. Hence, we proceeded to drop variables that did not show sufficient heterogeneity. We kept a small number of key moderators with their hypothesized effects before embarking in our meta-regression.

2.4 Summary of Hypotheses

We proceed to review our hypotheses for the meta-regression as follows:

1. Informative anchors will be associated with larger anchoring effects relative to purely random anchors.
2. Financial incentives will be associated with smaller anchoring effects.
3. WTP tasks will be associated with larger anchoring effects relative to WTA tasks.
4. The use of the canonical design will be associated with larger anchoring effects.
5. Compatible anchors will be associated with larger anchoring effects.

[14]Emotion intensity is defined as a relatively stable individual characteristic pertaining to the strength with which individuals experience their emotions (Larsen and Diener, 1987).

3 Meta-Analytic Results

3.1 Description of Studies

We included 24 articles, comprising 53 studies. We incorporated articles in which the author(s) reported several studies. In some articles, certain studies were presented as a single one with multiple conditions (moderators). In these cases, we treated the article as containing several studies, depending on the number of moderators. We report detailed information in Table 2. We treated studies contained in a single article as independent. In the practice of meta-analysis, this methodological choice is reasonable, as long as the studies use different participants, and participants are not counted multiple times. We assumed that article-level effects or author-level effects (biases introduced by the fact that some experiments are conducted by the same group or by overlapping groups of researchers) are small enough to be safely neglected. However, we will report relevant robustness checks following the main analysis.

We were also able to obtain raw data for 13 of the 24 articles (21 of 53 studies).[15] In Figure 1, we depict the numbers of studies falling into each category of the various moderators. All moderators had a degree of heterogeneity. In Figure 2a, we illustrate the number of studies per year, observing how the topic became popular after the seminal study of Ariely et al. (2003). In Figure 2b, we display the distribution of sample sizes in the literature and show that most studies had a sample size smaller than 200.

[15]We requested the data by sending emails to researchers. We did not follow-up with reminders, but included 'raw data availability' as a variable in our meta-regression.

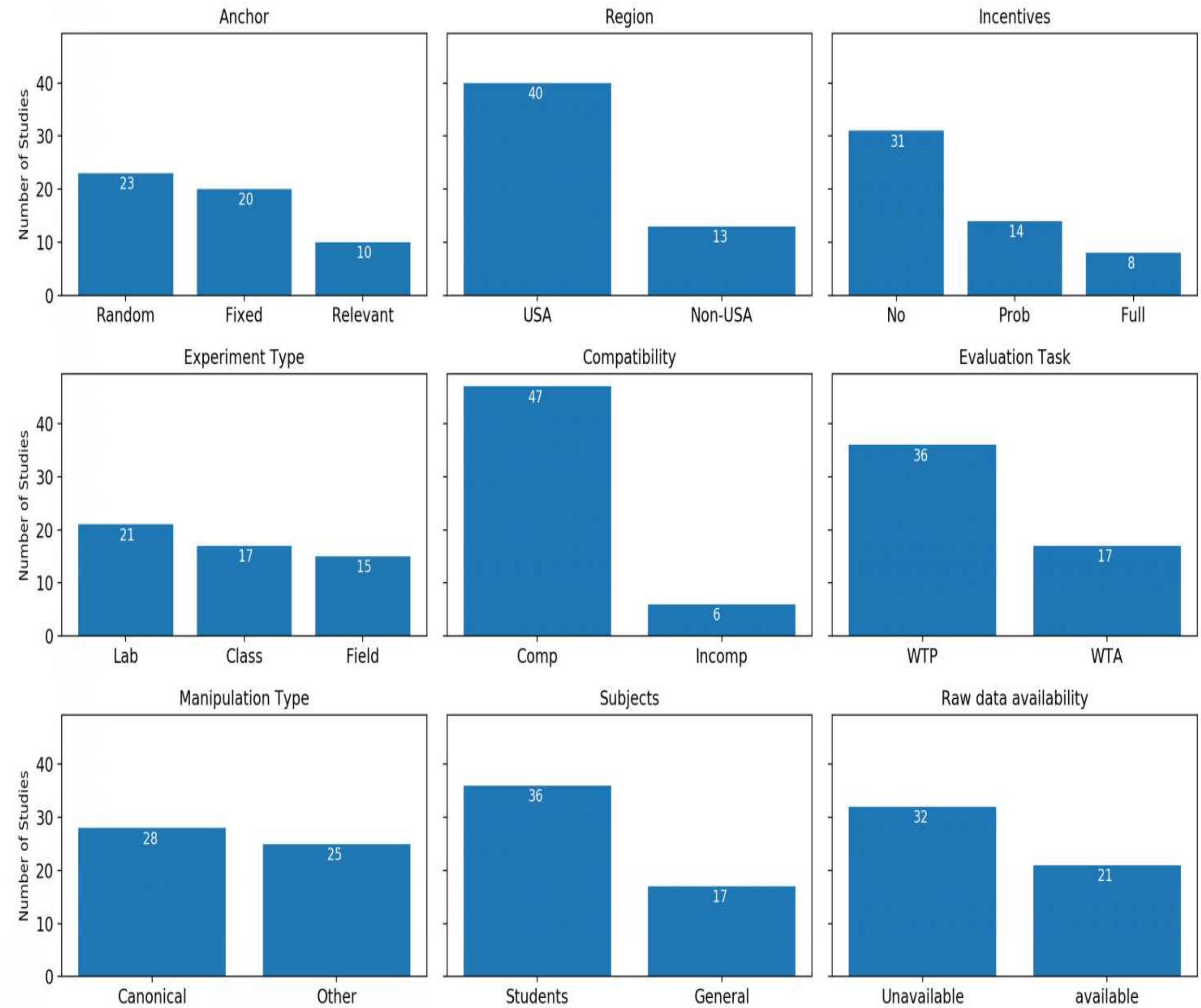
Table 2: Summary of included articles

Article	# of Studies	Method	Raw data
Adaval and Wyer (2011)	2	2	Yes
Alevy et al. (2015)	2	2	Yes
Ariely et al. (2003)	4	1	Partial
Andrersson and Wisaeus (2013)	1	NA	No
Bavolár et al. (2017)	1	NA	Yes
Bergman et al. (2010)	1	NA	Yes
Brzozowicz et al. (2017)	2	2	Yes
Brzozowicz and Krawczyk (2019)	2	2	Yes
Dogerlioglu-Demir and Koçaş (2015)	4	2	No
Fudenberg et al. (2012)	4	1	Yes
Green et al. (1998)	1	NA	No
Koçaş and Dogerlioglu-Demir (2014)	1	1	No
Li et al. (2019)	1	NA	No
Maniadis et al. (2014)	1	NA	Yes
Northcraft and Neale (1987)	4	1 and 2	No
Nunes and Boatwright (2004)	1	NA	No
Schläpfer and Schmitt (2007)	1	NA	Yes
Simonson and Drolet (2004)	8	1, 2	Partial
Tanford et al. (2019)	2	1 and 2	Yes
Wu et al. (2008)	2	1	No
Wu and Cheng (2011)	2	2	No
Yoon et al. (2013)	1	NA	Yes
Yu et al. (2017)	1	NA	No
Yoon and Fong (2019)	4	1	No

1. The Method column - 1: one article breaks into several studies because of multiple experiments conducted; 2: one article breaks into several studies based on moderators.

2. For two studies we have partial data. For Ariely et al. (2003), we have data pertaining to "EXPERIMENT 1: COHERENTLY ARBITRARY VALUATION OF ORDINARY PRODUCTS" (pp. 75). For Simonson and Drolet (2004), we have data for 'Study 1' (pp. 683).

Figure 1: Summary of number of studies in each category, complete dataset (53 studies)



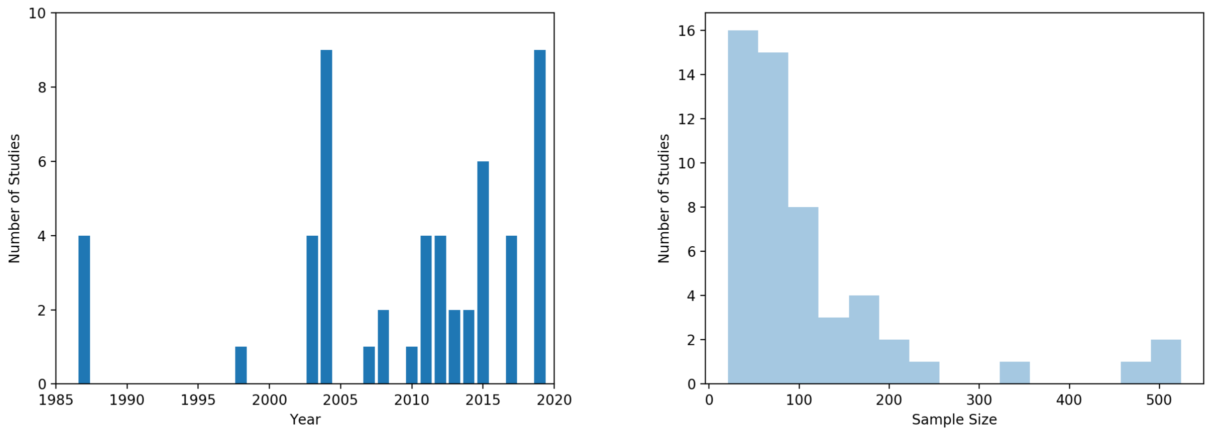
3.2 Average Effect Size

In Figure 3, we present an overview of the extracted effect sizes. Panel 1 illustrates that we obtained a negative relationship between publication year and effect size magnitude. In the second panel, we plot effect size against study sample size, illustrating a weak positive relationship.[16]

In practice, meta-researchers do not typically conduct meta-analysis directly with r (Pear-

[16] One study (Green et al., 1998) had an unusually large effect size (larger than 0.8). These authors conducted a large field experiment to elicit the WTP of an environmental good. Given that this is an extreme outlier, we consider the robustness of our results when excluding it, especially in the meta-regression part. In particular, in Appendix C, we provide meta-analytic results where this study is excluded. Compared to the results of the whole dataset, the overall effect size is smaller but the significance levels of the coefficients in the regression are very similar.

Figure 2: Number of studies against publication year and sample size, complete dataset (53 studies)



(a) (b)

son correlation coefficient), but they transform r to Fisher's z using this formula:

$$z = 0.5 \times ln(\frac{1+r}{1-r}) \tag{1}$$

This is so, because the variance of r (v_r) highly depends on the correlation itself.[17] The z transformation avoids this problem, as the variance of z is:

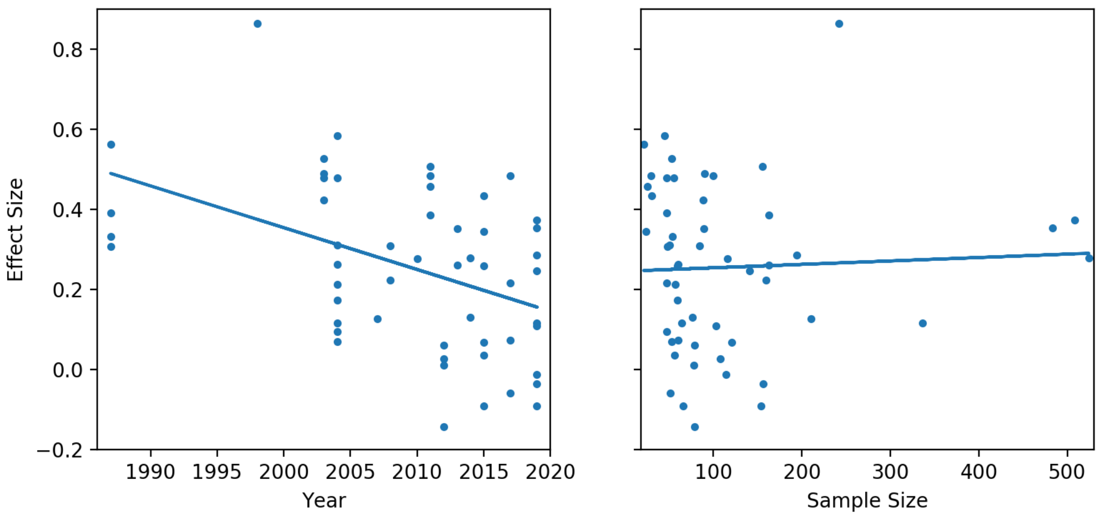
$$v_z = \frac{1}{n-3}, \tag{2}$$

which is a simple and 'excellent approximation' (Cooper et al., 2009, p. 231). We followed the convention of using z. To facilitate the interpretation of the results, we used equation $r = \frac{e^{2z}-1}{e^{2z}+1}$ to convert the results expressed in terms of z back to r.

To combine estimates of effect size from different studies, we employed two standard sets of models. The fixed-effects model assumes a single true (population) effect size for all studies, whereas the random-effects model assumes variation in the true effect size between studies. The fixed-effect estimate of the overall correlation coefficient between anchor number and elicited valuation is 0.286, with 95% confidence interval [0.263, 0.309]. The results did not change much when we applied random effects analysis. The overall average effect size is 0.267, with 95% confidence interval [0.194, 0.338], and the estimate of between-study variance is 0.068. The results point to a moderate overall effect, smaller than the effects reported in early studies. We

[17]The formula for v_r is $v_r = \frac{(1-r^2)^2}{n-1}$, where r is the sample correlation and n is the sample size.

Figure 3: Effect size as a function of publication year and sample size, complete dataset (53 studies)



report the forest plots for both fixed and random effects meta-analysis in Figures A.1 and A.2 in the Appendix A.

We found substantial heterogeneity among studies. We carried out a test of heterogeneity of the effect sizes and obtained a very large I^2 statistic equal to 88.2% (highly significant heterogeneity).[18] This indicates that differences across studies play a major role, and hence necessitate a deeper examination into how such differences matter for determining the effect size. Indeed, our moderator analysis shows that measurable differences across studies can explain a substantial fraction of this heterogeneity.

3.3 Moderator Analysis

As we explained, we coded a series of moderators that can be used as explanatory variables for the observed effect sizes. We present in Table 3 sub-group results based on the moderators. In addition, we did a meta-regression based on those moderators. The model we estimated is as follows:

$$z_i = \alpha + X_i\beta + e_i + u_i, \tag{3}$$

[18] I^2 measures the percentage of variation in effect sizes that is attributable to heterogeneity rather than pure chance (Higgins and Thompson, 2002; Higgins et al., 2003).

Table 3: Sub-group random-effect estimates of the overall ES, complete dataset (53 studies)

Category	Overall ES	95% CI	# of Studies
Random	0.205	[0.136, 0.272]	23
Fixed	0.275	[0.101, 0.433]	20
Related	0.415	[0.328, 0.495]	10
WTP	0.273	[0.180, 0.360]	36
WTA	0.251	[0.144, 0.352]	17
Canonical	0.276	[0.167, 0.379]	28
Non-canonical	0.249	[0.166, 0.329]	25
Students	0.229	[0.162, 0.294]	36
General population	0.339	[0.185, 0.477]	17
Not incentivized	0.307	[0.200, 0.407]	31
Prob. incentivized	0.240	[0.136, 0.338]	14
Fully incentivized	0.163	[0.040, 0.281]	8
Lab	0.204	[0.127, 0.279]	21
Class	0.305	[0.183, 0.417]	17
Field	0.312	[0.144, 0.463]	15
Compatible	0.279	[0.201, 0.353]	47
Incompatible	0.141	[-0.011, 0.288]	6
No raw data	0.351	[0.261, 0.435]	32
Raw data	0.126	[0.044, 0.206]	21
USA	0.275	[0.183, 0.363]	40
Non-USA	0.249	[0.150, 0.344]	13

where z_i is z transformation of r for study i,[19] X_i is a vector of coded moderators, the parameter $e_i \sim N(0, \sigma_i^2)$ captures within-study variation, and the parameter $u_i \sim N(0, \tau^2)$ captures between-study variation.[20]

We report the regression results in Table 4. All variables are binary,[21] and for all moderators we treated the most common category (see Figure 1) as the baseline, zero variable. For example, the baseline for the moderator *type of anchor* is *random anchor* (category 1), and the baseline for the moderator *type of experiment* is *lab experiment* (category 1).

In general, the estimated coefficients of our meta-regression had the expected sign. In particular, the presence of non-random anchors (either directly related to the goods or not)

[19]We have mentioned in Section 3.2 that r is not suitable for performing syntheses, thus here we use z in our regressions. However, for comparison purposes, and since the interpretation of r is more intuitive, we also report the results using r in Appendix B. These results are similar in terms of the significance levels of the coefficients.

[20]We used the 'metareg' command in Stata. It is essentially variance-weighted least squares regression with between-study variation u_i. Please note that the variance of e_i is known, since we have calculated it using formula 2 (and the formula in footnote 17 when we regress r).

[21]It is worth providing some clarification regarding the variable 'time period of publication'. We chose the year 2012 as a cutoff, because there is an almost equal number of studies before and after this year (26 studies before 2012 and 27 studies in and after 2012). Additional analysis using year of publication as a continuous variable produced similar results (see Appendix E).

significantly increased the anchoring effect. This conforms to our hypothesis, and is consistent with earlier studies that experimentally tested this hypothesis (Bavolár et al., 2017; Sugden et al., 2013). Moreover, selling tasks were associated with a lower anchoring effect (although the coefficient was non-significant for some model specifications), consistent with Simonson and Drolet (2004). Anchors that are incompatible with the elicited valuation object were also associated with lower elicited evaluations, as predicted by Strack and Mussweiler (1997) and consistent with early experimental tests of this question, such as by Chapman and Johnson (1994). Non-laboratory experiments yielded stronger, but often non-significant, anchoring effects, which could be attributed to the weaker experimental control and higher variance. On the other hand, incentives did not influence the anchoring effect. Studies conducted in more recent years have generally smaller (but often non-significant) effect sizes, as do studies where we use the available raw data. The use of raw data likely leads to higher precision, whereas the non-response decision may be correlated with the effect size in several ways.

Table 4: Meta-regression on z, complete dataset (53 studies)

VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
Fixed anchor	0.198**	0.181**	0.181**	0.196**	0.181**
	(0.0842)	(0.0800)	(0.0792)	(0.0839)	(0.0883)
Related anchor	0.284**	0.256**	0.252**	0.292***	0.328***
	(0.121)	(0.113)	(0.0989)	(0.103)	(0.107)
Omitted var.: random anchor					
Prob. incentive	0.117	0.133	0.135	0.101	0.105
	(0.110)	(0.106)	(0.102)	(0.107)	(0.113)
Full incentive	0.0545	0.0355	0.0355	-0.0546	-0.132
	(0.107)	(0.102)	(0.101)	(0.100)	(0.0998)
Omitted var.: no incentive					
Class experiment	0.0815	0.113	0.111	0.162*	0.213**
	(0.102)	(0.0899)	(0.0863)	(0.0889)	(0.0910)
Field experiment	0.183	0.185	0.169*	0.177*	0.267**
	(0.232)	(0.230)	(0.0987)	(0.105)	(0.103)
Omitted var.: lab experiment					
WTA	-0.142*	-0.127*	-0.127*	-0.0954	-0.126
	(0.0785)	(0.0748)	(0.0740)	(0.0772)	(0.0802)
Incompatible	-0.157	-0.169	-0.170	-0.250*	-0.307**
	(0.131)	(0.129)	(0.127)	(0.130)	(0.135)
Non-canonical	-0.146	-0.127	-0.123	-0.109	-0.0713
	(0.0974)	(0.0921)	(0.0796)	(0.0840)	(0.0870)
Raw data	-0.145*	-0.152**	-0.152**	-0.181**	
	(0.0751)	(0.0739)	(0.0728)	(0.0761)	
2012 or later	-0.203**	-0.185**	-0.186**		
	(0.0803)	(0.0753)	(0.0745)		
General population	-0.0568	-0.0180			
	(0.231)	(0.222)			
Non-USA	0.0683				
	(0.103)				
Constant	0.288**	0.307**	0.305**	0.203*	0.0988
	(0.123)	(0.119)	(0.115)	(0.114)	(0.112)
Observations	53	53	53	53	53

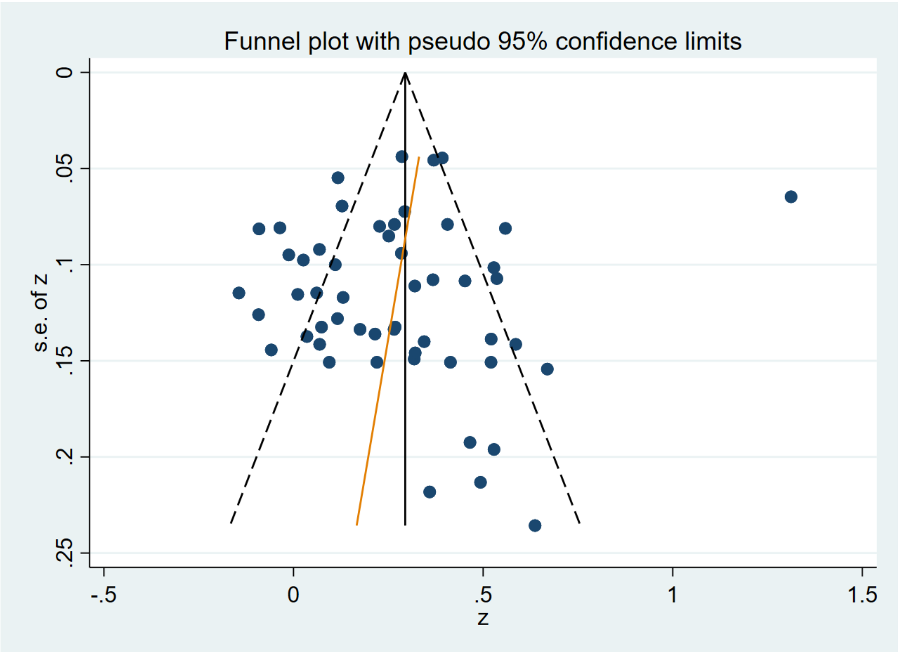
Standard errors in parentheses. $***p < 0.01$, $**p < 0.05$, $*p < 0.1$.

4 Robustness Checks

4.1 Publication Bias

The published literature is more likely to contain studies that report statistically significant effects, as researchers have incentives to place studies with non-significant results in the "file drawer", and journal articles are more likely to publish studies reporting significant results. Also, it is possible that studies with interesting findings (significant effect sizes) are more likely to be accepted in prestigious journals that are more accessible to the meta-analyst. The retrieved studies for a particular meta-analysis could be a fraction of all relevant studies, and there is evidence that hidden studies may be systematically different from available ones (Dickersin, 2005; Song et al., 2000). The publication bias problem, then, may distort the information available to the meta-analyst.[22]

Figure 4: Funnel plot with z, complete dataset (53 studies)



To address this problem, we present the 'funnel plot' of our meta-analysis in Figure 4, noting that in the x-axis we have the z-transformation of r, and in the y-axis the standard

[22]Perhaps dissemination bias is a more accurate term to describe all the aforementioned types of bias. We follow convention and use publication bias instead (Song et al., 2000). The bias against null results is not the only dissemination mechanism that is problematic and the meta-analyst needs to address. As we have noted elsewhere (Maniadis et al., 2017), there can also be biases against replicating an initial effect. These may ensue from research incentives associated with the 'Proteus phenomenon', whereby contradicting famous studies may be preferable to successfully replicating it (Ioannidis and Trikalinos, 2005). Finally, there may also be reluctance for publishing positive findings that are in the opposite direction of established results, especially if the sample size is small.

error of z (a measure of effect size precision). The graph is asymmetric and reveals that small studies reporting small effect sizes are missing (studies with large s.e. of z). This suggests the possibility of publication bias, but we need to be circumspect when interpreting the funnel plot. Publication bias is not the only possible cause of funnel plot asymmetry.[23]Also, the choice of precision measurement could significantly change the appearance of the plot. For instance, the asymmetry in the effect sizes against sample sizes graph in Figure 3 is unnoticeable. Therefore, we used the linear regression test of Egger et al. (1997) to examine statistically the asymmetry. We regressed the standardized effect size against a measure of result precision:

$$z_i/\sqrt{v_i} = \beta_0 + \beta_1(1/\sqrt{v_i}) + e_i, \tag{4}$$

where $e_i \sim N(0, s^2)$ and v_i is the sampling variance of study i. We were interested in the significance level of the intercept, given that it is a measure of bias. The result shows that the intercept is not significant ($p = 0.398$). Hence, we cannot conclude that there is significant asymmetry.

We carried out a comprehensive literature search. We retrieved some of the grey literature, such as working papers and unpublished manuscripts, which probably helped in reducing the publication bias. Therefore, we tentatively conclude that publication bias is not a major problem in our analysis. However, it is possible that there are some studies with low sample sizes and small effect sizes left in the file drawer, in which case we may have overestimated the overall effect size.[24]

4.2 Other Potential Biases

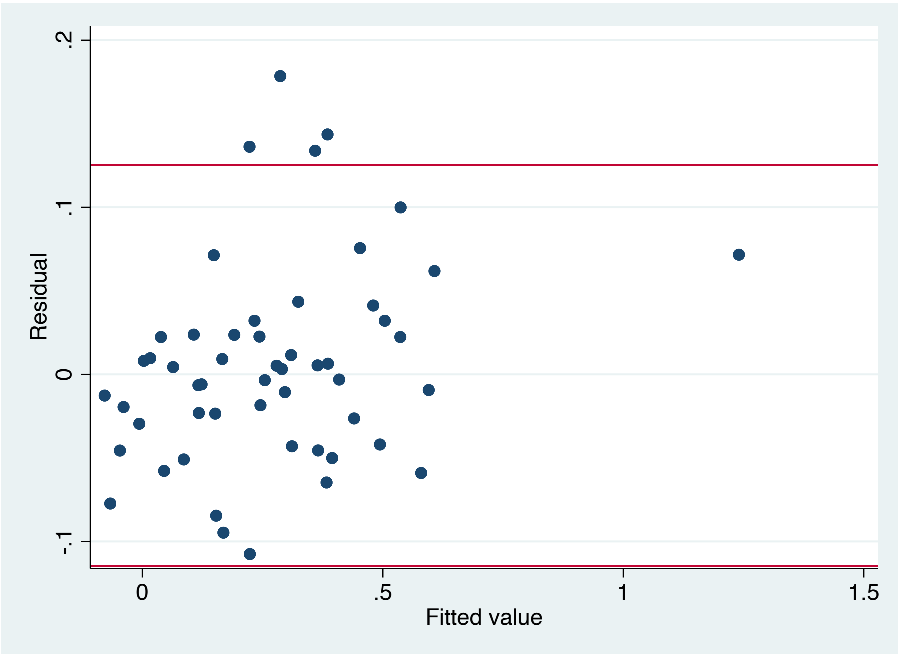
Several of our included articles contain multiple studies. We have assumed so far that the studies are independent. In this section, we conducted robustness analysis to examine the effect of possible research team-level bias (Ioannidis, 2005). First, we followed a standard recommendation (Higgins et al., 2019) to address this concern. For each article with multiple

[23]According to Sterne and Egger (2001), other causes of asymmetry are inadequate methodological quality of smaller studies and true heterogeneity (effect size is correlated to study size).

[24]There is evidence that existing tools for accounting for publication biases are not sufficient, and there is an ongoing dialogue that considers the validity of meta-analysis as a tool. Kvarven et al. (2020) compared meta-analyses to pre-registered multiple-laboratory replications, and found that the replication effect sizes are significantly smaller than the average meta-analytic effect sizes. This is another reason why our effect size may be viewed as an upper bound of the true effect size.

studies, we selected one study at random and excluded the others. This left us with a *reduced dataset* that contained 24 studies.[25] The fixed-effect estimate of the overall effect size is 0.311 with 95% confidence interval [0.281, 0.342], and the random-effect estimate of the overall effect size is 0.300 with 95% confidence interval [0.173, 0.416]. The overall effect sizes of this *reduced dataset* are thus slightly larger than of the complete dataset. Unfortunately, after selecting randomly one study per article the sample size becomes too small to meaningfully detect the effect or moderators. The regression results can be found in Table D.4 in the Appendix.

Figure 5: Residual plot with z, complete dataset (53 studies)



To detect potential outliers, we also visually inspected the residual-fitted value plot in Figure 5. Note that the fitted value is $\alpha + \widehat{X\beta} + e$ and the red lines correspond to ± 2 standard deviations from the mean. The residuals between the two red lines show no clear pattern (probably a weak positive correlation). However, there were four studies with large residuals (roughly between 0.1 and 0.2), which all belong to two articles: Adaval and Wyer (2011) and Dogerlioglu-Demir and Koçaş (2015). This might imply some particular bias, or some effect that we failed to capture with our moderators. We examined the robustness of our results if we exclude all studies from these two articles. That is, we excluded six studies from Adaval and Wyer (2011) and Dogerlioglu-Demir and Koçaş (2015). We re-conducted the regression with the remaining 47 studies, and present the results in Table 5. This analysis produced

[25]The selection is random, and Appendix D provides a general description of the selected studies. Figure D.5 shows that some heterogeneity remains in the reduced dataset. Figure D.7 plots effect sizes against time and sample size, and it reveals a pattern similar to that of the complete dataset.

roughly similar results to those of Table 4.[26] Overall, although some particular team-level bias or missing moderator cannot be ruled out, it did not jeopardize the main regression analysis.

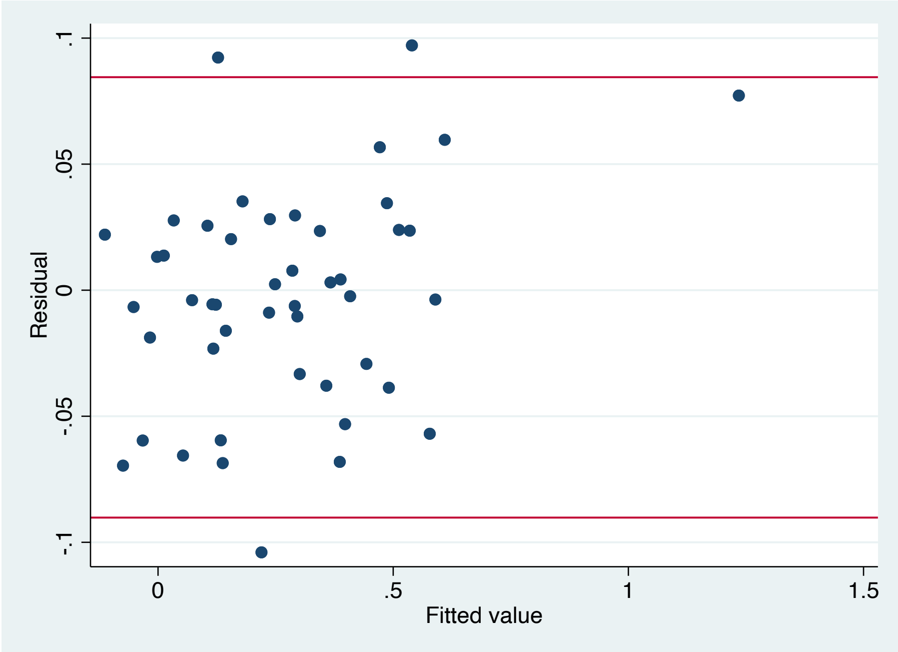
[26] For this dataset, the negative effect of non-canonical designs now appears significant, whereas the analogous effect of the availability of the raw data is now insignificant. There are no sign changes or significant differences in the value of the respective coefficients.

Table 5: Meta-regression on z, reduced dataset (47 studies)

VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
Fixed anchor	0.150*	0.161**	0.162**	0.173**	0.161*
	(0.0870)	(0.0793)	(0.0785)	(0.0837)	(0.0885)
Related anchor	0.290**	0.302**	0.284***	0.324***	0.361***
	(0.121)	(0.115)	(0.0986)	(0.103)	(0.107)
Omitted var.: random anchor					
Prob. incentive	0.152	0.141	0.150	0.129	0.119
	(0.111)	(0.106)	(0.102)	(0.108)	(0.114)
Full incentive	0.0399	0.0493	0.0483	-0.0225	-0.113
	(0.106)	(0.101)	(0.100)	(0.102)	(0.100)
Omitted var.: no incentive					
Class experiment	0.158	0.135	0.126	0.182*	0.229**
	(0.113)	(0.0920)	(0.0871)	(0.0897)	(0.0925)
Field experiment	0.287	0.279	0.213**	0.217**	0.306***
	(0.237)	(0.232)	(0.0994)	(0.106)	(0.105)
Omitted var.: lab experiment					
WTA	-0.128	-0.138*	-0.137*	-0.0974	-0.132
	(0.0803)	(0.0745)	(0.0738)	(0.0767)	(0.0795)
Incompatible	-0.294*	-0.290*	-0.292*	-0.262	-0.435**
	(0.172)	(0.169)	(0.167)	(0.179)	(0.175)
Non-canonical	-0.175*	-0.182*	-0.166*	-0.156*	-0.122
	(0.101)	(0.0977)	(0.0822)	(0.0873)	(0.0909)
Raw data	-0.144	-0.136	-0.133	-0.199**	
	(0.0881)	(0.0836)	(0.0824)	(0.0828)	
2012 or later	-0.178*	-0.193**	-0.194**		
	(0.0921)	(0.0811)	(0.0803)		
General population	-0.0547	-0.0719			
	(0.234)	(0.226)			
Non-USA	-0.0414				
	(0.117)				
Constant	0.306**	0.297**	0.290**	0.186	0.0866
	(0.123)	(0.118)	(0.115)	(0.114)	(0.112)
Observations	47	47	47	47	47

Standard errors in parentheses. $***p < 0.01$, $**p < 0.05$, $*p < 0.1$.

Figure 6: Residual plot with z, reduced dataset (47 studies)



5 Conclusion

The anchoring bias in consumers' assessments of WTP and WTA is thought to yield a highly robust and large effect. We conducted a research synthesis pertaining to the importance and determinants of anchoring effects on statements of economic valuation. We retrieved 53 studies from 24 articles using an exhausting search of the literature. We obtained an effect (correlation coefficient between anchor and target item) of moderate size. This is generally smaller for the effects of earlier studies of the phenomenon. Overall, anchoring might not be as strong and robust as it has been considered.

Yet, our analysis also uncovered substantial heterogeneity of the effect size, and the overall effect size might be slightly problematic in the presence of significant heterogeneity.[27] To address this issue, we carried out a meta-regression to examine parameters that could drive this heterogeneity. Accordingly, we embarked in a moderator analysis, where we relied on published articles to gauge the existence of differences in theoretically relevant parameters. In this analysis, several features of the design were associated with higher anchoring effects, such as non-random and compatible anchors, and buying (rather than selling) tasks. On the other hand, financial incentives did not matter for anchoring effects. This affirms the relative

[27] Heterogeneity is also a problem for the use of funnel plot analysis pertaining to the publication bias issue, because this analysis assumes that studies do not differ fundamentally. We have already noted that our related analysis should be viewed with caution.

robustness of anchoring, and that it is likely to be relevant for important economic decisions outside the laboratory.

References

Adaval, R. and Wyer, R. S. (2011). Conscious and nonconscious comparisons with price anchors: Effects on willingness to pay for related and unrelated products. *Journal of Marketing Research*, 48(2):355–365.

Alevy, J. E., Landry, C. E., and List, J. A. (2015). Field experiments on the anchoring of economic valuations. *Economic Inquiry*, 53(3):1522–1538.

Andrersson, P. and Wisaeus, B. (2013). Age and anchoring. http://arc.hhs.se/download.aspx?MediumId=1927.

Araña, J. E. and León, C. J. (2008). Do emotions matter? coherent preferences under anchoring and emotional effects. *Ecological Economics*, 66(4):700–711.

Ariely, D., Loewenstein, G., and Prelec, D. (2003). "coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, 118(1):73–106.

Bavolár, J. et al. (2017). Experience with the product does not affect the anchoring effect, but the relevance of the anchor increases it. *Ekonomickỳ časopis*, 95(03):282–293.

Bergman, O., Ellingsen, T., Johannesson, M., and Svensson, C. (2010). Anchoring and cognitive ability. *Economics Letters*, 107(1):66–68.

Blankenship, K. L., Wegener, D. T., Petty, R. E., Detweiler-Bedell, B., and Macy, C. L. (2008). Elaboration and consequences of anchored estimates: An attitudinal perspective on numerical anchoring. *Journal of Experimental Social Psychology*, 44(6):1465–1476.

Borenstein, M., Cooper, H., Hedges, L., and Valentine, J. (2009). Effect sizes for continuous data. *The handbook of research synthesis and meta-analysis*, 2:221–235.

Brzozowicz, M. and Krawczyk, M. (2019). Anchors don't hold for real? the anchoring effect and hypothetical bias in declared willingness to pay. *Unpublished, in review.*

Brzozowicz, M., Krawczyk, M., Kusztelak, P., et al. (2017). Do anchors hold for real? anchoring effect and hypothetical bias in declared wtp. *University of Warsaw, Faculty of Economic Sciences.*

Caplan, B. (2000). Rational irrationality: a framework for the neoclassical-behavioral debate. *Eastern Economic Journal*, 26(2):191–211.

Chapman, G. B. and Johnson, E. J. (1994). The limits of anchoring. *Journal of Behavioral Decision Making*, 7(4):223–242.

Cooper, H., Hedges, L. V., and Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis.* Russell Sage Foundation.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. *Publication bias in meta-analysis: Prevention, assessment and adjustments*, pages 11–33.

Dogerlioglu-Demir, K. and Koçaş, C. (2015). Seemingly incidental anchoring: the effect of incidental environmental anchors on consumers' willingness to pay. *Marketing Letters*, 26(4):607–618.

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, 315(7109):629–634.

Epley, N. and Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making*, 18(3):199–212.

Fudenberg, D., Levine, D. K., and Maniadis, Z. (2012). On the robustness of anchoring effects in wtp and wta experiments. *American Economic Journal: Microeconomics*, 4(2):131–145.

Green, D., Jacowitz, K. E., Kahneman, D., and McFadden, D. (1998). Referendum contingent valuation, anchoring, and willingness to pay for public goods. *Resource and Energy Economics*, 20(2):85–116.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M., and Welch, V. (2019). Cochrane handbook for systematic reviews of interventions. https://training.cochrane.org/handbook/current.

Higgins, J. P. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11):1539–1558.

Higgins, J. P., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *Bmj*, 327(7414):557–560.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLos med*, 2(8):e124.

Ioannidis, J. P. and Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: the proteus phenomenon in molecular genetics research and randomized trials. *Journal of clinical epidemiology*, 58(6):543–549.

Jacowitz, K. E. and Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, 21(11):1161–1166.

Jung, M. H., Perfecto, H., and Nelson, L. D. (2016). Anchoring in payment: Evaluating a judgmental heuristic in field experimental settings. *Journal of Marketing Research*, 53(3):354–368.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives*, 5(1):193–206.

Koçaş, C. and Dogerlioglu-Demir, K. (2014). An empirical investigation of consumers' willingness-to-pay and the demand function: The cumulative effect of individual differences in anchored willingness-to-pay responses. *Marketing Letters*, 25(2):139–152.

Kvarven, A., Strømland, E., and Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, 4(4):423–434.

Larsen, R. J. and Diener, E. (1987). Affect intensity as an individual difference characteristic: A review. *Journal of Research in personality*, 21(1):1–39.

Lau, J., Ioannidis, J. P., Terrin, N., Schmid, C. H., and Olkin, I. (2006). The case of the misleading funnel plot. *Bmj*, 333(7568):597–600.

Le, H., Schmidt, F., and Oh, I. (2009). Correcting for the distorting effects of study artifacts in meta-analysis. *The handbook of research synthesis and meta-analysis*, page 317.

LeBoeuf, R. A. and Shafir, E. (2009). Anchoring on the" here" and" now" in time and distance judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(1):81.

Li, T., Fooks, J. R., Messer, K. D., and Ferraro, P. J. (2019). A field experiment to estimate the effects of anchoring and framing on residents' willingness to purchase water runoff management technologies. *Resource and Energy Economics*. In press.

Maniadis, Z., Tufano, F., and List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *The American Economic Review*, 104(1):277–290.

Maniadis, Z., Tufano, F., and List, J. A. (2017). To replicate or not to replicate? exploring reproducibility in economics through the lens of a model and a pilot study. *Economic Journal*, 127(605).

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, 151(4):264–269.

Mussweiler, T. and Strack, F. (2001). Considering the impossible: Explaining the effects of implausible anchors. *Social Cognition*, 19(2):145–160.

Mussweiler, T., Strack, F., and Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin*, 26(9):1142–1150.

Northcraft, G. B. and Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational behavior and human decision processes*, 39(1):84–97.

Nunes, J. C. and Boatwright, P. (2004). Incidental prices and their effect on willingness to pay. *Journal of Marketing Research*, 41(4):457–466.

Olkin, I. and Gleser, L. (2009). Stochastically dependent effect sizes. *The handbook of research synthesis and meta-analysis*, pages 357–376.

Orwin, R. G. and Vevea, J. L. (2009). Evaluating coding decisions. *The handbook of research synthesis and meta-analysis*, 2:177–203.

Peterson, R. A. and Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology*, 90(1):175.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638.

Schkade, D. A. and Johnson, E. J. (1989). Cognitive processes in preference reversals. *Organizational Behavior and Human Decision Processes*, 44(2):203–231.

Schläpfer, F. and Schmitt, M. (2007). Anchors, endorsements, and preferences: a field experiment. *Resource and Energy Economics*, 29(3):229–243.

Simonson, I. and Drolet, A. (2004). Anchoring effects on consumers' willingness-to-pay and willingness-to-accept. *Journal of consumer research*, 31(3):681–690.

Slovic, P. (1967). The relative influence of probabilities and payoffs upon perceived risk of a gamble. *Psychonomic Science*, 9(4):223–224.

Song, F., Easterwood, A., Guilbody, S., Duley, L., and Sutton, A. (2000). Publication and other selection bias in systematic reviews. *Health Technology Assessment*, 4(10):1–115.

Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and statistics*, 70(1):103–127.

Sterne, J. A. and Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of clinical epidemiology*, 54(10):1046–1055.

Strack, F. and Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of personality and social psychology*, 73(3):437.

Sugden, R., Zheng, J., and Zizzo, D. J. (2013). Not all anchors are created equal. *Journal of Economic Psychology*, 39:21–31.

Tanford, S., Choi, C., and Joe, S. J. (2019). The influence of pricing strategies on willingness to pay for accommodations: Anchoring, framing, and metric compatibility. *Journal of Travel Research*, 58(6):932–944.

Thompson, S. G. and Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in medicine*, 21(11):1559–1573.

Tufano, F. (2010). Are 'true'preferences revealed in repeated markets? an experimental demonstration of context-dependent valuations. *Experimental Economics*, 13(1):1–13.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.

Wansink, B., Kent, R. J., and Hoch, S. J. (1998). An anchoring and adjustment model of purchase quantity decisions. *Journal of Marketing Research*, 35(1):71–81.

Wegener, D. T., Petty, R. E., Detweiler-Bedell, B. T., and Jarvis, W. B. G. (2001). Implications of attitude change theories for numerical anchoring: Anchor plausibility and the limits of anchor effectiveness. *Journal of Experimental Social Psychology*, 37(1):62–69.

Whyte, G. and Sebenius, J. K. (1997). The effect of multiple anchors on anchoring in individual and group judgment. *Organizational behavior and human decision processes*, 69(1):75–85.

Wilson, D. B. (2001). Practical meta-analysis effect size calculator [online calculator]. https://campbellcollaboration.org/research-resources/effect-size-calculator.html.

Wu, C., Cheng, F., and Lin, H. (2008). Exploring anchoring effect and the moderating role of repeated anchor in electronic commerce. *Behaviour & Information Technology*, 27(1):31–42.

Wu, C.-S. and Cheng, F.-F. (2011). The joint effect of framing and anchoring on internet buyers' decision-making. *Electronic Commerce Research and Applications*, 10(3):358–368.

Yoon, S. and Fong, N. (2019). Uninformative anchors have persistent effects on valuation judgments. *Journal of Consumer Psychology*, 29:391–410.

Yoon, S., Fong, N. M., and Dimoka, A. (2013). The robustness of anchoring effects on market good valuations. *Available at SSRN 2352692*.

Yu, L., Gao, Z., Sims, C., and Guan, Z. (2017). Effect of price on consumers' willingness to pay: is it from quality perception or price anchoring? *Agricultural & Applied Economics Association Annual Meeting, Chicago, Illinois.*

Appendices

A Forest plots with z, complete dataset

Figure A.1: Fixed effects model, complete dataset (53 studies)

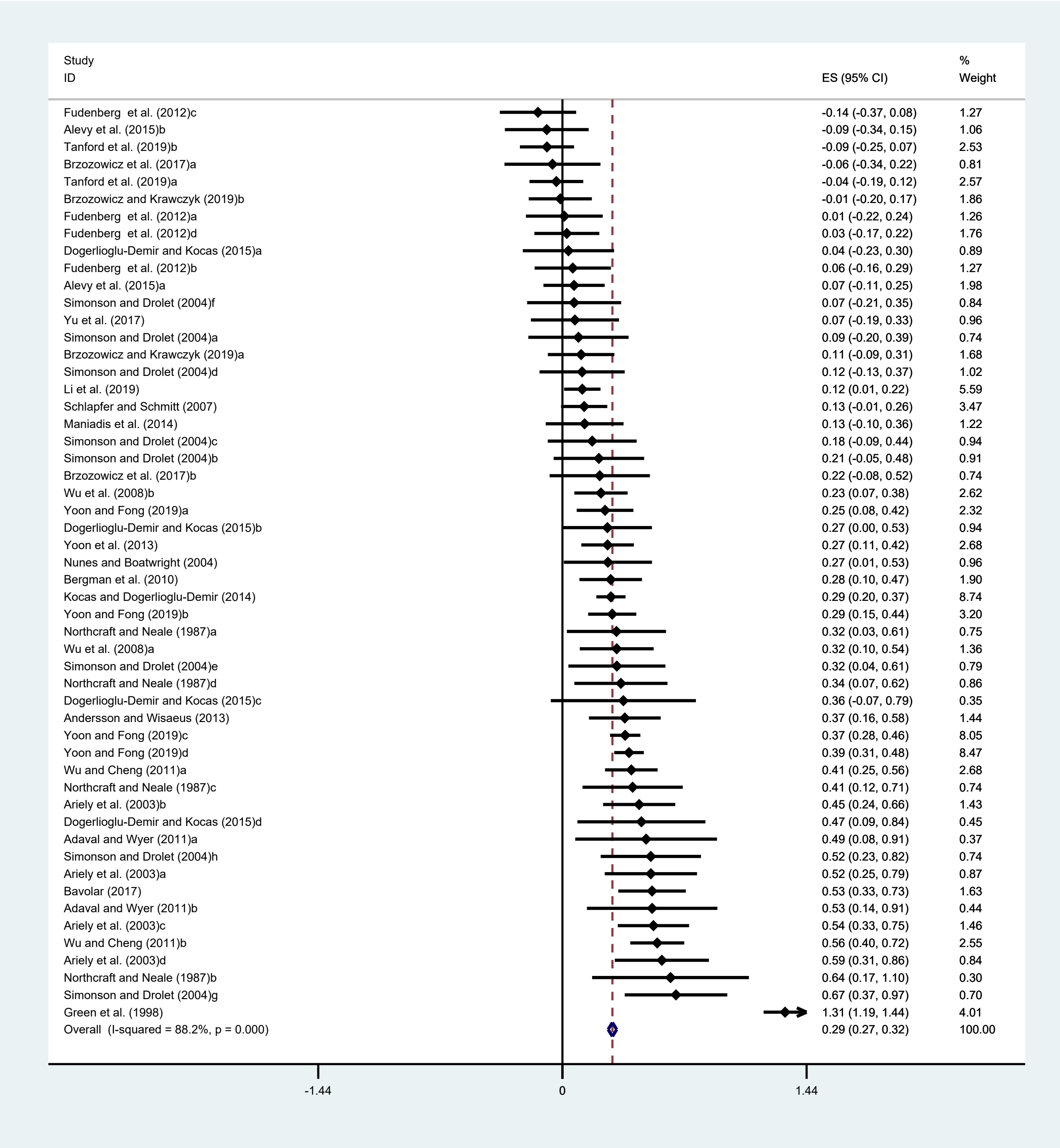
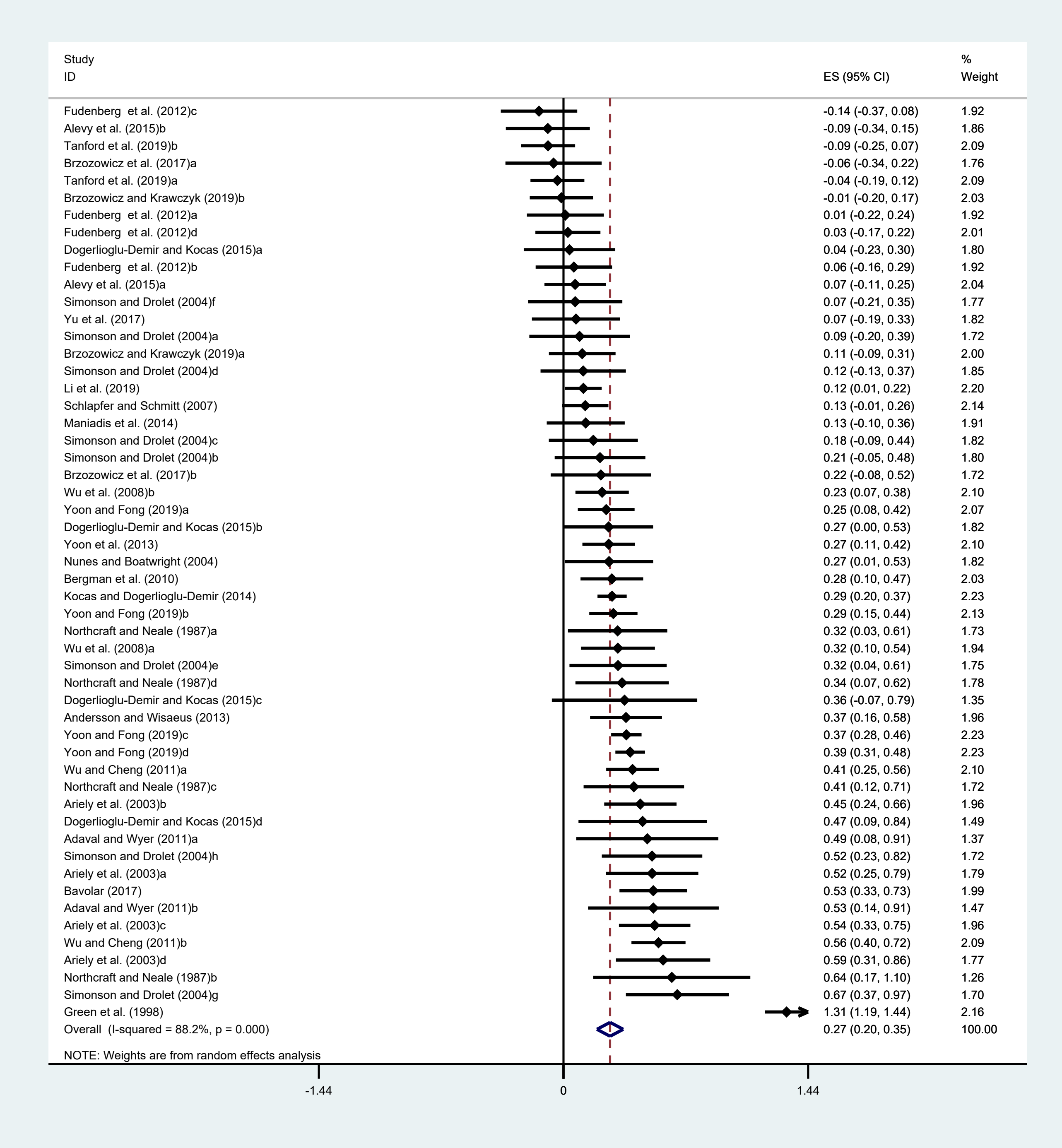


Figure A.2: Random effects model



B Meta-regression on r, complete dataset

Table B.1: Meta-regression on r, complete dataset (53 studies)

VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
Fixed anchor	0.134**	0.126**	0.126**	0.142**	0.129*
	(0.0647)	(0.0616)	(0.0611)	(0.0656)	(0.0707)
Related anchor	0.243**	0.230**	0.222***	0.257***	0.288***
	(0.0953)	(0.0882)	(0.0776)	(0.0814)	(0.0864)
Omitted var.: random anchor					
Prob. incentive	0.103	0.111	0.115	0.0894	0.0929
	(0.0862)	(0.0832)	(0.0799)	(0.0849)	(0.0913)
Full incentive	0.0415	0.0320	0.0319	-0.0420	-0.115
	(0.0843)	(0.0802)	(0.0795)	(0.0799)	(0.0809)
Omitted var.: no incentive					
Class experiment	0.0800	0.0948	0.0914	0.131*	0.176**
	(0.0790)	(0.0693)	(0.0665)	(0.0697)	(0.0730)
Field experiment	0.155	0.157	0.125	0.128	0.210**
	(0.178)	(0.177)	(0.0764)	(0.0821)	(0.0823)
Omitted var.: lab experiment					
WTA	-0.115*	-0.109*	-0.109*	-0.0789	-0.105
	(0.0617)	(0.0586)	(0.0582)	(0.0612)	(0.0649)
Incompatible	-0.123	-0.130	-0.133	-0.191*	-0.247**
	(0.102)	(0.0999)	(0.0986)	(0.103)	(0.109)
Non-canonical	-0.117	-0.108	-0.101	-0.0870	-0.0510
	(0.0768)	(0.0723)	(0.0623)	(0.0665)	(0.0703)
Raw data	-0.140**	-0.143**	-0.142**	-0.166***	
	(0.0587)	(0.0576)	(0.0568)	(0.0601)	
2012 or later	-0.162**	-0.154**	-0.154**		
	(0.0620)	(0.0578)	(0.0572)		
General population	-0.0529	-0.0348			
	(0.180)	(0.172)			
Non-USA	0.0319				
	(0.0806)				
Constant	0.290***	0.299***	0.295***	0.207**	0.111
	(0.0956)	(0.0926)	(0.0897)	(0.0900)	(0.0898)
Observations	53	53	53	53	53

Standard errors in parentheses. $***p < 0.01$, $**p < 0.05$, $*p < 0.1$.

C Meta-analytic results, reduced dataset with (Green et al., 1998) being excluded (52 studies)

- Fixed effect model: overall effect size is 0.247 with 95% confidence interval [0.222, 0.271]
- Random effect model: overall effect size is 0.241 with 95% confidence interval [0.190, 0.290]
- Estimate of between-study variance $\tau^2 = 0.024$
- I-squared (variation in effect size attributable to heterogeneity) = 72.3%

Table C.2: Sub-group random-effect estimates of the overall ES, reduced dataset (52 studies)

Category	Overall ES	95% CI	# of Studies
Random	0.205	[0.136, 0.272]	23
Fixed	0.199	[0.114, 0.281]	19
Related	0.415	[0.328, 0.495]	10
WTP	0.238	[0.180, 0.295]	35
WTA	0.251	[0.144, 0.352]	17
Canonical	0.236	[0.172, 0.299]	27
Non-canonical	0.249	[0.166, 0.329]	25
Students	0.229	[0.162, 0.294]	36
General population	0.265	[0.186, 0.340]	16
Not incentivized	0.263	[0.198, 0.326]	30
Prob. incentivized	0.240	[0.136, 0.338]	14
Fully incentivized	0.163	[0.040, 0.281]	8
Lab	0.204	[0.127, 0.279]	21
Class	0.305	[0.183, 0.417]	17
Field	0.235	[0.156, 0.311]	14
Compatible	0.252	[0.200, 0.302]	46
Incompatible	0.141	[-0.011, 0.288]	6
No raw data	0.315	[0.267, 0.361]	31
Raw data	0.126	[0.044, 0.206]	21
USA	0.238	[0.179, 0.300]	39
Non-USA	0.249	[0.150, 0.344]	13

Table C.3: Meta-regression on z, reduced dataset (52 studies)

VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
Fixed anchor	0.0484	0.0652	0.0640	0.0683	0.0622
	(0.0628)	(0.0583)	(0.0574)	(0.0605)	(0.0677)
Related anchor	0.228**	0.251***	0.249***	0.278***	0.306***
	(0.0882)	(0.0817)	(0.0727)	(0.0747)	(0.0815)
Omitted var.: random anchor					
Prob. incentive	0.117	0.104	0.104	0.0854	0.0953
	(0.0787)	(0.0760)	(0.0734)	(0.0763)	(0.0847)
Full incentive	0.00971	0.0266	0.0259	-0.0215	-0.0955
	(0.0765)	(0.0725)	(0.0714)	(0.0713)	(0.0747)
Omitted var.: no incentive					
Class experiment	0.123	0.0973	0.0963	0.120*	0.164**
	(0.0737)	(0.0637)	(0.0607)	(0.0628)	(0.0684)
Field experiment	0.0805	0.0831	0.0733	0.0618	0.145*
	(0.177)	(0.176)	(0.0718)	(0.0753)	(0.0792)
Omitted var.: lab experiment					
WTA	-0.0669	-0.0818	-0.0818	-0.0541	-0.0813
	(0.0592)	(0.0551)	(0.0543)	(0.0554)	(0.0606)
Incompatible	-0.0915	-0.0846	-0.0854	-0.104	-0.174*
	(0.0923)	(0.0914)	(0.0897)	(0.0941)	(0.103)
Non-canonical	-0.0819	-0.0987	-0.0978	-0.0863	-0.0382
	(0.0711)	(0.0663)	(0.0581)	(0.0608)	(0.0660)
Raw data	-0.159***	-0.152***	-0.152***	-0.170***	
	(0.0541)	(0.0530)	(0.0519)	(0.0536)	
2012 or later	-0.0947	-0.114**	-0.114**		
	(0.0600)	(0.0533)	(0.0525)		
General population	0.0220	-0.0105			
	(0.176)	(0.169)			
Non-USA	-0.0556				
	(0.0779)				
Constant	0.309***	0.296***	0.296***	0.230***	0.126
	(0.0860)	(0.0840)	(0.0813)	(0.0800)	(0.0831)
Observations	52	52	52	52	52

Standard errors in parentheses. $***p < 0.01$, $**p < 0.05$, $*p < 0.1$.

Figure C.3: Funnel plot with z, reduced dataset (52 studies)

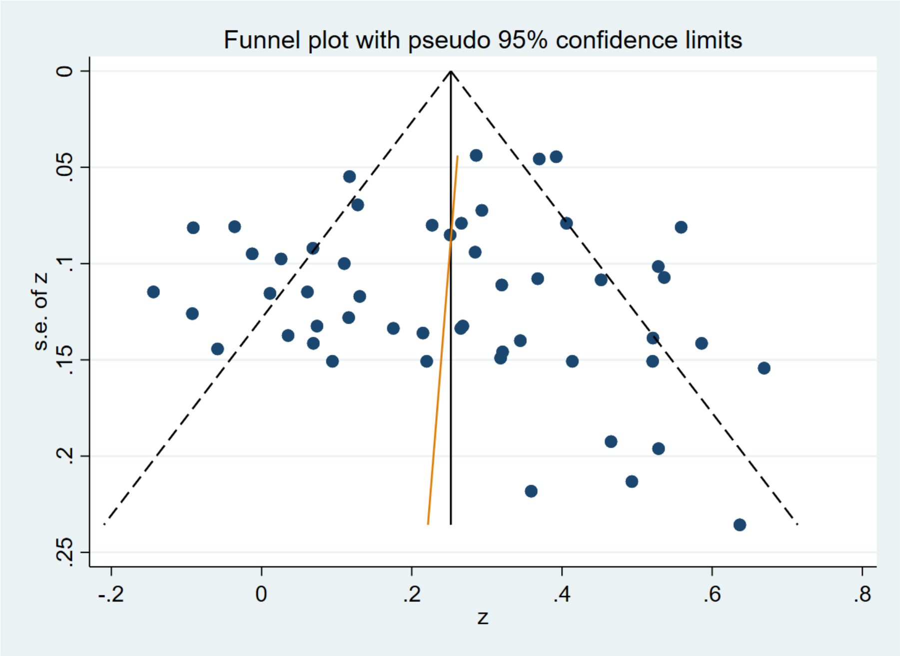
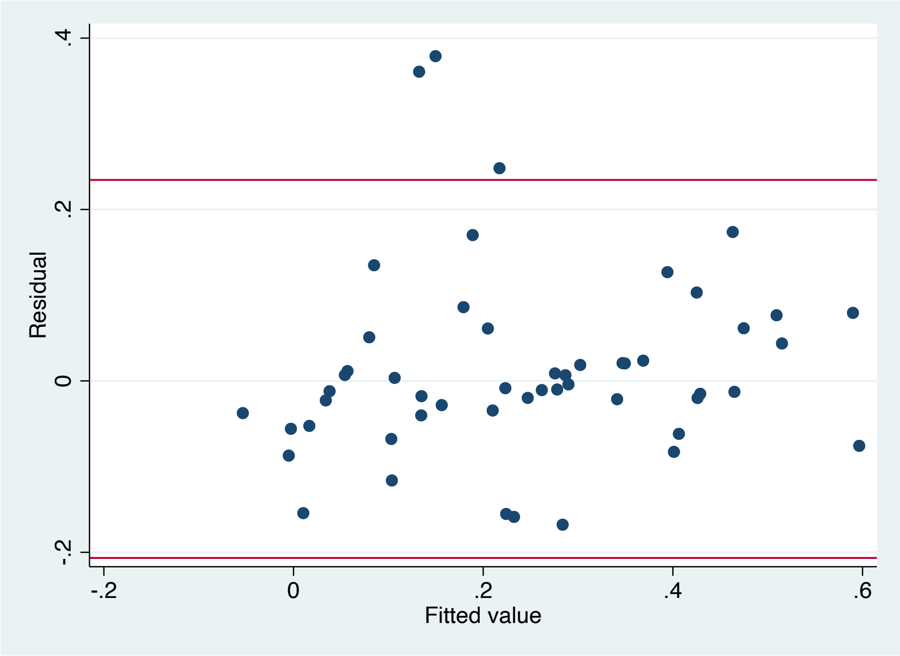


Figure C.4: Residual plot with z, reduced dataset (52 studies)



D Descriptions and Meta-analytic results, reduced dataset (24 studies)

Figure D.5: Summary of number of studies in each category, reduced dataset (24 studies)

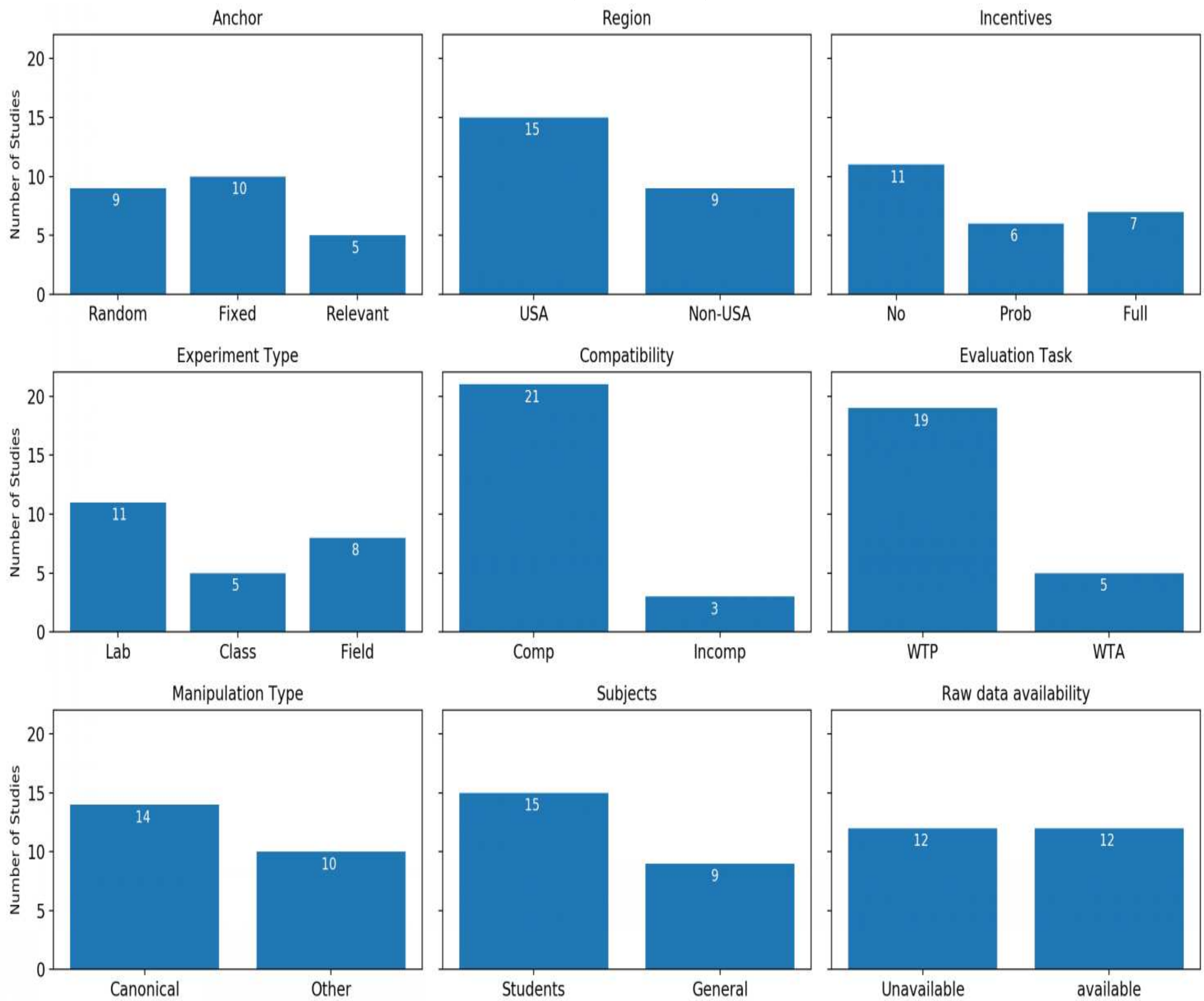
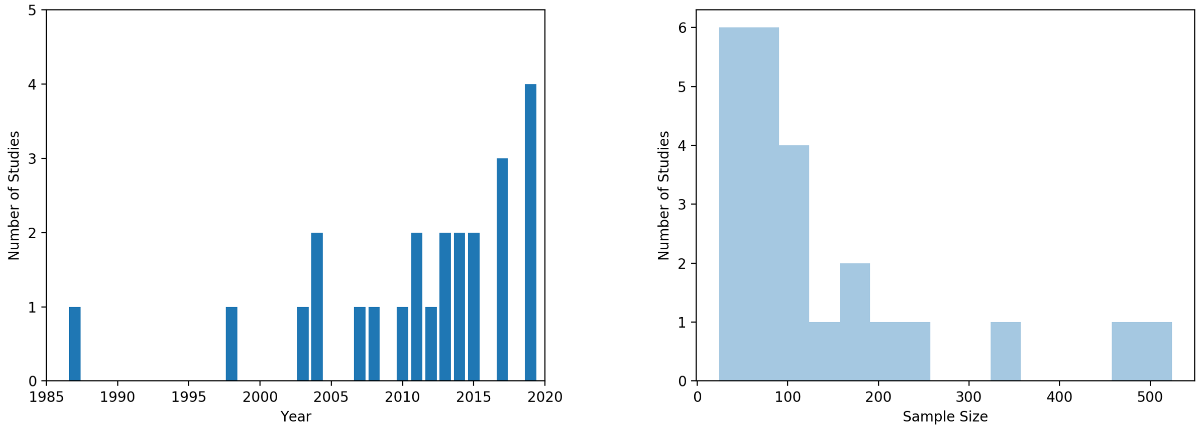


Figure D.6: Number of studies against publication year and sample size, reduced dataset (24 studies)



(a)

(b)

Figure D.7: Effect size as a function of publication year and sample size, reduced dataset (24 studies)

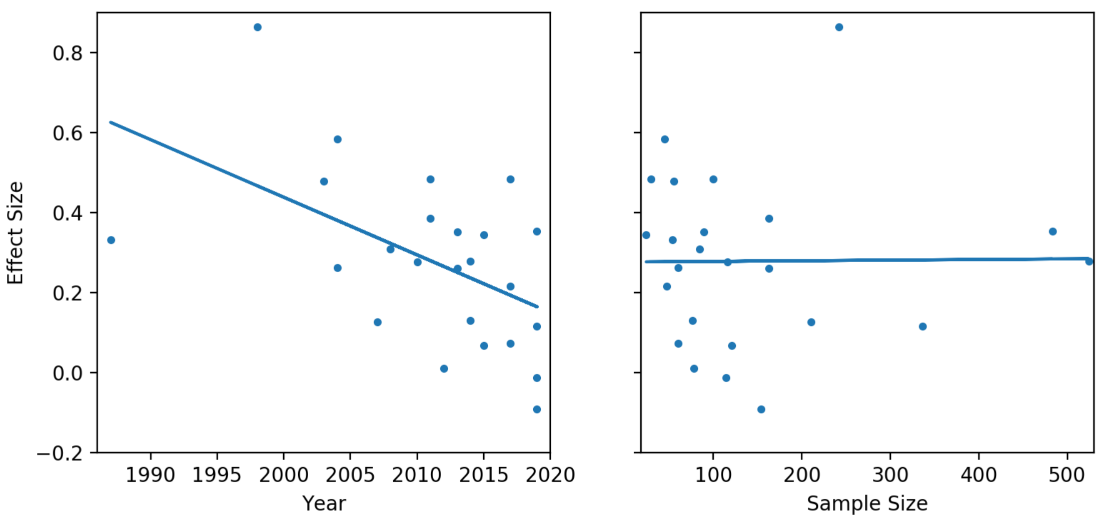


Table D.4: Meta-regression on z, reduced dataset (24 studies)

VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
Fixed anchor	0.157	0.139	0.142	0.120	0.107
	(0.216)	(0.197)	(0.190)	(0.205)	(0.191)
Related anchor	0.103	0.0629	0.0558	0.0967	0.100
	(0.277)	(0.236)	(0.227)	(0.241)	(0.234)
Omitted var.: random anchor					
Prob. incentive	-0.0775	-0.0841	-0.0728	-0.163	-0.193
	(0.314)	(0.294)	(0.282)	(0.297)	(0.261)
Full incentive	-0.0507	-0.0840	-0.0847	-0.228	-0.249
	(0.243)	(0.210)	(0.203)	(0.196)	(0.170)
Omitted var.: no incentive					
Class experiment	0.340	0.371	0.347*	0.336	0.331
	(0.237)	(0.209)	(0.182)	(0.194)	(0.188)
Field experiment	0.239	0.247	0.139	0.162	0.168
	(0.482)	(0.456)	(0.181)	(0.193)	(0.186)
Omitted var.: lab experiment					
WTA	-0.155	-0.133	-0.150	-0.171	-0.183
	(0.201)	(0.181)	(0.164)	(0.175)	(0.164)
Incompatible	-0.490	-0.520*	-0.519*	-0.585**	-0.605**
	(0.288)	(0.261)	(0.252)	(0.266)	(0.246)
Non-canonical	-0.160	-0.126	-0.108	-0.0977	-0.0970
	(0.227)	(0.192)	(0.172)	(0.184)	(0.179)
Raw data	-0.0899	-0.0933	-0.0803	-0.0406	
	(0.182)	(0.172)	(0.160)	(0.168)	
2012 or later	-0.234	-0.220	-0.213		
	(0.153)	(0.138)	(0.130)		
General population	-0.147	-0.120			
	(0.489)	(0.457)			
Non-USA	0.0660				
	(0.184)				
Constant	0.461	0.488	0.467	0.389	0.390
	(0.313)	(0.286)	(0.265)	(0.280)	(0.272)
Observations	24	24	24	24	24

Standard errors in parentheses. $***p < 0.01$, $**p < 0.05$, $*p < 0.1$.

E Using year of publication as a continuous variable

Table E.5: Meta-regression on z using year as a continuous variable

VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
Fixed anchor	0.175**	0.172**	0.171**	0.196**	0.181**
	(0.0859)	(0.0815)	(0.0806)	(0.0839)	(0.0883)
Related anchor	0.218*	0.213*	0.214**	0.292***	0.328***
	(0.128)	(0.120)	(0.105)	(0.103)	(0.107)
Omitted var.: random anchor					
Prob. incentive	0.0735	0.0772	0.0768	0.101	0.105
	(0.112)	(0.107)	(0.103)	(0.107)	(0.113)
Full incentive	0.0466	0.0439	0.0441	-0.0546	-0.132
	(0.111)	(0.107)	(0.106)	(0.100)	(0.0998)
Omitted var.: no incentive					
Class experiment	0.104	0.110	0.110	0.162*	0.213**
	(0.102)	(0.0920)	(0.0881)	(0.0889)	(0.0910)
Field experiment	0.0861	0.0884	0.0909	0.177*	0.267**
	(0.242)	(0.239)	(0.108)	(0.105)	(0.103)
Omitted var.: lab experiment					
WTA	-0.182**	-0.179**	-0.180**	-0.0954	-0.126
	(0.0885)	(0.0844)	(0.0835)	(0.0772)	(0.0802)
Incompatible	-0.167	-0.169	-0.169	-0.250*	-0.307**
	(0.134)	(0.132)	(0.130)	(0.130)	(0.135)
Non-canonical	-0.150	-0.147	-0.148*	-0.109	-0.0713
	(0.0998)	(0.0941)	(0.0820)	(0.0840)	(0.0870)
Raw data	-0.166**	-0.167**	-0.167**	-0.181**	
	(0.0753)	(0.0740)	(0.0728)	(0.0761)	
Year of publication	-0.0119**	-0.0118**	-0.0118**		
	(0.00575)	(0.00561)	(0.00553)		
General population	-0.00388	0.00241			
	(0.234)	(0.225)			
Non-USA	0.0132				
	(0.0997)				
Constant	24.13**	23.96**	24.00**	0.203*	0.0988
	(11.59)	(11.31)	(11.16)	(0.114)	(0.112)
Observations	53	53	53	53	53

Standard errors in parentheses. $***p < 0.01$, $**p < 0.05$, $*p < 0.1$.