ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Master Thesis Behavioural Economics

**Predictions in the Surprisingly Popular Method**

Name student: **Melvin Hanswijk**
Student ID number: 358204

Supervisor: Aurélien Baillon
Second assessor: Benjamin Tereick

Date final version: 1 August 2017

Abstract

Several tested methods that aggregate crowd wisdom in a way superior to majority voting require that respondents make predictions about the answers of other respondents. These predictions are required to be accurate on average, but no research has yet been done into how people tend to form them. This thesis researches whether respondents base their predictions on different possible worlds or not, one world in which the answer to a binary question is correct and one where it is false. It also tests whether researchers can influence this process by having respondents report their confidence in advance. In an experiment with three treatments and two stages, respondents, mostly students, give their main predictions in one stage and their separate predictions for each possible world in the other. A comparison between the two stages then reveals which worlds respondents base their predictions on. Roughly one-third of respondents seem to generally base their predictions on both possible worlds, another third tends to base their predictions on the most likely world, and the last third does not form different predictions for the different worlds. Respondents who had to report their confidence in advance were not more likely than other respondents to fall into any of these three categories. Several factors contributed to respondents giving internally inconsistent predictions, difficulty of the tasks and low incentives being the most important.

# Contents

# 1 Introduction

The core principle of democratic voting is also its weakness: the majority rules. When information is hard to access, or experts in a field are but few, situations where the majority is wrong can occur quite easily. Prelec, Seung, and Mccoy (2014) call problems where the majority vote is wrong 'majority-unsolvable' problems. In the same paper, they propose an algorithm to deal with this type of problems: the Least Surprised by the Truth (LST) algorithm. The method requires that the researcher asks the respondents not only the question that he wants to answer, but also a prediction question. Respondents need to predict how often each possible answer is given by the other respondents.

Recently, the same authors published another method: the Surprisingly Popular (SP) method (Prelec, Seung, & McCoy, 2017a). On the researcher's side, the methods have some differences. The Surprisingly Popular method is both easier to understand and easier to use. It also requires different assumptions than the older method. To respondents, however, the methods are identical. Both require the same prediction question next to the actual question. It is this prediction question that this thesis investigates.

Prelec et al. (2017a) generally assume that respondents imagine different worlds when they make a prediction, meaning that they imagine a world where an answer is true and a world where the answer is false. Respondents think about what the distribution of answers would be in either world, and use that to make their prediction. The authors show that the method also works if respondents do not imagine different worlds, instead basing their prediction solely on what they believe is true, but they do not test how many respondents use one process and how many use the other.

To get a deeper understanding of how the SP method works and which assumptions are really necessary, and to take a step in the direction of more accurate and efficient models, I set out to answer the following question:

*Do respondents in the Surprisingly Popular method base their predictions on both possible worlds, or solely on the one they believe is actual?*

With the follow-up:

*Can the researcher influence this by asking for confidence in advance?*

I give a more elaborate motivation and explanation for these questions in a later chapter, after describing the SP method in detail.

Because LST is identical to SP from the respondent's point of view, the predictions are made in the same manner. As such, any answers that I find regarding the prediction process of respondents in the SP method, are identical for the prediction process in the LST method.[1] I elaborate on LST later and make a comparison between the two methods. The main focus, however, will be on SP.

[1] Assuming respondents do not know which method the researcher is using.

To answer the research question, I run an experiment in which respondents have to give separate predictions for both worlds, besides the actual question and the standard prediction. I compare the standard predictions with the predictions for both worlds to see whether the standard prediction was based on one world or on two. I find that respondents are approximately equally divided over three different types. Some do not imagine different worlds at all, some can imagine different worlds, but base their prediction on only one of them, and some imagine different worlds and base their prediction on both of them. I find no evidence that asking for confidence in advance influences which of these three types a respondent is likely to be.

The remainder of this thesis is organised as follows. Section 2 is the literature review, which consists of a thorough explanation and overview of the SP method, including a comparison to LST. In section 3 I go deeper into the assumptions underlying SP. The research question and hypotheses are stated and motivated in section 4. Section 5 encompasses the entire experimental design, and in section 6 I analyse the data. In section 7 I discuss topics that were not yet discussed in the analysis and elaborate on some topics that were. Section 8 concludes.

# 2 Literature Review

In this review, I discuss the SP and LST methods. I start by explaining the SP method, followed by the formal model and the method's origin. This brings me naturally to the LST method. After explaining LST, I make a theoretical and an empirical comparison between the two methods.

## 2.1 The Surprisingly Popular method

### 2.1.1 The method

The Surprisingly Popular method (SP) (Prelec et al., 2017a) requires asking the respondents two questions. The first is the actual question that the researcher wants to answer, the second is a meta question, namely: 'What percentage of other respondents do you think gave answer *x*?'

I will use an example from Prelec et al. to illustrate. Consider the following question:

'Is Philadelphia the capital of Pennsylvania?'

Many people know that Philadelphia is a large city in Pennsylvania, and they might assume that it is indeed the capital. Depending on the respondents' backgrounds, it could well be a minority of people that knows that the capital of Pennsylvania is actually Harrisburg.

Say 40% of the population knows that Philadelphia is not the capital, while 60% thinks it is. Assuming our sample is representative, we will then have 40% of respondents answering 'No' and 60% answering 'Yes'.

For the second question, all respondents are asked to predict the percentage of respondents that answered 'Yes'. This is where the key assumption that SP relies on comes in.

The assumption is that in a world where a certain fact is true, more people will believe that that fact is true than in a world where that fact is not true. For example, when faced with the statement 'Gold is the most expensive metal on earth', more people will think this statement to be true when gold is indeed the most expensive metal on earth than when it is not.

We also assume that respondents know the distribution of answers in both possible worlds, they just do not know which world is real.[2]

In the Philadelphia example, we can discern two worlds. The first is the world where Philadelphia is not the capital of Pennsylvania; this is the actual world. The second is a counterfactual world, where Philadelphia is, in fact, the capital of Pennsylvania.

Now, assume that in the actual world, 60% of people believe that Philadelphia is the capital (which is why 60% answered 'Yes' to the first question). In the counterfactual world, however, where this fact is true, 90% believes that.

[2] This (respondents knowing the distributions) is the strongest version of this assumption. SP also works under weaker, more realistic, versions, like predictions being accurate on average. I discuss this later.

If all respondents are completely certain of their answer, the predictions will be as follows:

The 40% of respondents that correctly answered 'No', will correctly predict that 60% will answer 'Yes'.

The 60% of respondents that answered 'Yes' will not make a correct prediction. They believe they are in the counterfactual world, and their prediction will be based on that world. This 60% of people will therefore predict that 90% of people will answer 'Yes'.

The total, average, prediction of the percentage of people that will answer 'Yes' will be:

0.4*60% + 0.6*90% = 78%.

The total prediction for the relative amount of 'No' will be 22%.

The actual percentages (question 1) were 60% for 'Yes' and 40% for 'No'. The answer 'No', even though only 40% of people gave it as an answer, is therefore more popular than predicted, or 'Surprisingly Popular'. This answer is the correct answer.

When respondents are not completely certain of their answers, their individual predictions will lie somewhere between 60% and 90%, depending on their confidence. The 40% of respondents that answered 'No' will give predictions between 60% and 75%, while the 60% of respondents that answered 'Yes' will give predictions between 75% and 90%. The total, average, prediction will lie somewhere between 60% and 90%. Since the actual frequency, in a sufficiently large and representative sample, will converge to 60%, 'Yes' will be less popular than predicted. The correct answer, 'No', will be Surprisingly Popular.

### 2.1.2 Formal model

Prelec et al. (2017a) published the Surprisingly Popular method with both empirical evidence and theoretical proof (Prelec, Seung, & McCoy, 2017b). Their proof utilises a formal model with *m* possible worlds, which can be thought of as coins, and *n* possible signals, which can be thought of as the number of sides each coin has. Every world corresponds to an answer, while a signal is the information a respondent has, on which he bases his answer. To illustrate the model, and to explain the terminology that I will use from now on, I will relate it to the Philadelphia example from earlier.

Different worlds, each containing a coin, are the different answers to a multiple-choice question. There are *m* possible answers, $\{a_1,...,a_m\}$. Respondents base their vote V for a certain answer on the evidence that they possess. This evidence is summarised by signal S, with *n* possible signals, $\{s_1,...,s_n\}$. The signal is the result of the coin flip, with *n* being the number of sides that the coin has.

Question: Is Philadelphia the capital of Pennsylvania?

Answer 'No' = Actual world.

Answer 'Yes' = Counterfactual world

Signal: The information a respondent has on which he bases his answer. Let us assume, for the sake of simplicity, that every person thinks one of two possible things. For example, we could say that there is the general

knowledge that Philadelphia is the capital, and there is the not commonly known fact that Harrisburg is actually the capital. One of these facts is false, of course. Which one that is, depends on which world we are in. If everyone starts out with the general knowledge, and some people learn the other fact, we can discern two signals. $s_1$ is just the general knowledge, leading to a vote for answer 'Yes'. $s_2$ is the general knowledge plus the added fact, which leads to a vote for answer 'No'. Thus, we have the case m=n=2.

Recall that in the world where Philadelphia is not the capital, 60% of respondents think it is. In other words, in the actual world, 60% of people receive $s_1$ and 40% of people receive $s_2$. We can say that this world has a coin that lands on 'heads' 60% of the time and on 'tails' 40% of the time.

In the counterfactual world, 90% of people receive $s_1$ and 10% of people receive $s_2$. We can say that this world has a coin that lands on 'heads' 90% of the time, and on 'tails' 10% of the time.

Depending on the outcome of their coin flip, respondents get one of the two signals. Based on that signal, they give their answer to question 1. They also predict the relative frequencies of the given answers.

Respondents therefore:

1. State which world they think is the actual world, by stating their vote (= their answer on Question 1).
2. Give their prediction of the distribution of given answers, so they predict in which world other respondents think they are.

Prelec et al. (2017b) prove that with ideal respondents, the model above will always result in the correct answer being Surprisingly Popular. Characteristics of ideal respondents are that they know which worlds are possible (which coins exist), they know the prior distributions of signals in these worlds (the coins' biases), and they apply Bayesian updating to form their beliefs about which world is real (which coin is being used). The proof is extended to the general case of $m$, $n > 2$. Prelec et al. also show that not all of these characteristics are necessary for the SP method to be successful. I discuss the required assumptions in section 3.

### 2.1.3 Empirical evidence and origins

Empirically, the SP method has been shown to be an improvement on the simple majority vote, as well as on different scoring rules that incorporate respondents' confidence in their answers. The fields where the SP method has proven successful include state capitals, trivia, lesions (assessment by dermatologists), and art (estimating market value) (Prelec et al., 2017a).

To the best of my knowledge, there is no other published empirical research that provides evidence for the (still quite new) SP method as published in Prelec et al. (2017a). The SP method, however, is a cleaner adaptation of earlier methods that are based on the same underlying principle, namely that correct answers will be given more often than expected. Prelec (2004) introduces the Bayesian Truth Serum (BTS) as a method that "assigns high scores, not to the most common answers, but to answers that are more common than collectively predicted, with predictions drawn from the same population" (p.1). Like SP, BTS requires respondents to answer the question and to make a prediction. As the name implies, the Bayesian Truth Serum at that point is used as a means to incentivise respondents to give truthful answers. Weiss (2009) then uses the BTS scoring rule not as a

truth-telling incentive, but as a means of identifying experts and finding correct answers.[3] In a later study, Prelec, Seung, and McCoy (2014) formalise this use of the BTS and call it the Least Surprised by the Truth (LST) method. Four years later, the same authors publish SP. Let us first have a look at how LST works, and then make a comparison between the two methods.

## 2.2 The Least Surprised by the Truth method

### 2.2.1 The method

In the LST method, all respondents are assigned a score. This score, which is the same as the BTS score, is made up of two parts, which in the BTS framework are called the 'information score' and the 'prediction score' (Prelec, 2004; Prelec et al., 2014). The information score is dependent on the answer a respondent gives. Every answer has an information score, which is calculated based on the collective prediction of the relative frequency of the answer, combined with its actual relative frequency. This is what Prelec (2004) is talking about when he says that answers that are more common than collectively predicted get high scores. Since each answer has its own information score, respondents who give the same answer get the same information score. The other part of a respondent's BTS score is the prediction score. The prediction score is calculated based on the distribution of answers predicted by the respondent, combined with the actual distribution of answers. Respondents who better predict the distribution get higher prediction scores. For every answer, the researcher calculates the average BTS score of all the respondents who gave that answer. The answer with the highest average score is selected as the best answer.

### 2.2.2 Empirical evidence

Empirical evidence for the Least Surprised by the Truth method is slightly more varied than for SP. LST was also shown to outperform majority voting and confidence weighted voting when participants had to find the best move in certain chess problems (Weiss, 2009). Contrary to the tests in Prelec et al. (2017a), of which all but the art study had binary questions only, participants in Weiss (2009) had to choose from five answers in each question. Another difference is that solving chess problems is more a measure of skill than a measure of knowledge, as opposed to the tests that Prelec et al. used.[4] Because of their similarities, the fact that LST was successful in this study is an indication that the SP method, too, is applicable to a broader variety of question types than just knowledge based binary questions.

## 2.3 SP vs LST

### 2.3.1 In theory

Recall that the SP method selects the answer which is more popular than predicted as the right answer. Placing that in the LST/BTS framework, we could say that SP simply selects the answer with the highest information score, disregarding the prediction score altogether. At first sight, it may not seem logical to ignore the prediction score. The reasoning behind the prediction score is, after all, quite intuitive. Experts in a field are assumed to

[3] Weiss (2009) references a 2006 working paper from Prelec and Seung.

[4] It is arguable whether the assessment of lesions is more knowledge- or skill based. Either way, it is clear that solving chess problems requires a completely different type of skill. The same can be said for judging an artwork with which the respondent is not familiar.

have more meta knowledge, making them better able to predict the distribution of answers. We can therefore assume that respondents who make better predictions are more knowledgeable about the field in question. Therefore, we want to give more weight to their answers (Prelec et al., 2014; Weiss, 2009). This sounds very reasonable, and both studies provide empirical evidence showing that the method works, so why do Prelec et al. (2017a) ignore all this information? I think the answer is twofold.

First, the fact of the matter is that under the assumptions made, the Surprisingly Popular criterion is all that is needed to find the correct answer. Yes, data is being thrown away by focussing just on the information score and disregarding the prediction score, but (theoretically) we still have all the information we need to identify the right answer.

Second, the SP criterion is much easier, both to understand and to use, than the LST algorithm. Compared to just calculating the average of all predicted distributions and seeing which answer is more popular than predicted, it is quite complicated and time consuming to calculate not only the information scores, but also the prediction score for each respondent and the average BTS score for each answer.

### 2.3.2 In practice

That first point, the fact that the SP criterion alone is sufficient to get the right answer, is obviously not necessarily true in practice. After all, in none of the discussed studies did the SP method have a 100% success rate. The same is true for LST. To get an idea of what the ease of use of SP, compared to LST, cost in terms of performance, one could compare the performance of both methods when applied to the same datasets. To test SP, Prelec et al. (2017a) use the same 'state capital' studies that were used to test LST (Prelec et al., 2014). Table 1 shows the performance of both methods in three state capital studies. All three studies required participants to answer True/False questions about the capitals of all 50 states in the United States. The numbers shown are the number of questions that the method got correct, out of 50. The bold numbers are what is reported in the papers: the 2013 paper reports absolute values, the 2017 paper reports percentages.

| | Prelec et al. (2014) | | Prelec et al. (2017a) | |
|---|---|---|---|---|
| | Majority vote | LST | Majority vote | SP |
| Study 1 | **31** (62%) | **41** (82%) | 29 (**58%**) | 34 (**68%**) |
| Study 2 | **38** (76%) | **44** (88%) | 36 (**72%**) | 43 (**86%**) |
| Study 3 | **31** (62%) | **46** (92%) | 31 (**62%**) | 45 (**90%**) |

**Table 1 Performance of LST and SP in three state capital studies.**

Note that the two 'Majority vote' columns should show the same numbers, since they are about the same studies. The reason for the differences is unclear to me. It is possible that ties were treated differently: the 2013 paper states that ties were counted as 0.5 points, while the 2017 paper does not report how ties were dealt with.

In studies 2 and 3, SP and LST perform about the same. In study 1 however, there is quite a difference: LST outperforms SP quite a bit. It is possible that this has to do with the difference in incentivisation between the studies, and with the way that both methods are dependent on participants being properly incentivised. In study 1, there were no monetary incentives for respondents. In studies 2 and 3, respondents received a participation fee and could earn extra rewards by performing well on the tests, both on the True/False questions and the predictions. This fact by itself could explain the difference in performance between study 1 and studies 2 and 3.

The difference in performance of both methods in study 1 is interesting. Since there were no monetary incentives to perform well, it is likely that intrinsic motivation played a larger part than in the other studies, meaning that the variation in respondents' motivation was likely relatively high. It also seems probable that better-motivated respondents make better predictions, since they really think about the questions, while unmotivated respondents make low effort guesses. This will affect both methods differently.

### 2.3.3 Motivational issues

SP heavily relies on the average predictions being accurate within a given world. A group of respondents making random predictions can therefore be a big problem, since they will statistically push the average prediction to an even distribution. An example: Let us say that the actual distribution of answers for a given binary question is 30-70, with the 30% being correct. The collective prediction could then be around 20-80 if all respondents are really trying, making the SP method successful. If, however, due to a lack of incentives, a significant part of respondents gives random predictions, the collective prediction will be pushed towards 50-50. The collective prediction could end up as 40-60, making the incorrect answer the surprisingly popular one.

When applying the LST method to the example above, the random predictions will still mess up the information scores. Now, however, there are also the prediction scores. Respondents who made random predictions will have a low prediction score on average, while experts who made an effort will likely have the highest. Because of this, the LST method may still select the correct answer.

Another way of looking at it is that in the SP method, every respondent has the same amount of influence. Bad predictors have just as much influence as good predictors. In the LST method, good predictors have more influence, making it better equipped to deal with unmotivated respondents.

### 2.3.4 Conclusion

The advantage that SP offers in terms of simplicity and ease of use can be significant. If a researcher has the time, however, he would most likely be better off using LST. While the methods may often give extremely similar results, as in studies 2 and 3 above, the only way that I can see LST performing worse than SP is if the use of the prediction score actually has a negative impact on performance. This would only be the case in a situation where respondents who give a wrong answer make better predictions than respondents who give the correct answer, on average. It would be interesting to see whether this tends to happen for certain types of questions. Although a field may exist where this is structurally the case, I cannot imagine one.

# 3 Assumptions in SP

As mentioned in the description of the method, the Surprisingly Popular method needs two assumptions:

1. An answer will be more common if it is true than if it is not true.
2. Respondents know the distribution of answers for every world.

In this chapter I elaborate on these assumptions. I especially go into the different possible versions of assumption 2.

Implicit are the assumptions that all respondents base their predictions only on the possible (given) worlds, and at least on the world they believe most likely to be actual. This means that respondents do not imagine a world where none of the answers are true, nor do they base their predictions solely on another world than the most likely one. For binary questions, therefore, respondents have two options. They can base their predictions solely on the most likely world, or base them on both worlds.[5]

## 3.1 Assumption 1

I call assumption 1 the mcit (more common if true) assumption. It seems like a reasonable assumption to make. First, consider a simplified worldview in which a person either knows something for sure, or has no clue and guesses at random. Assume a statement of which 20% of people know whether it is true or false, while 80% is completely clueless. This 20% will then always give the right answer, while half of the 80% is expected to be wrong, and half to be right. Therefore, if the statement is true, 20% believes/knows it to be true, and we expect 60% (20+40) of respondents to give answer 'true'. If it is false, 0% of people believes/knows it to be true, and we expect 40% (0+40) of respondents to give answer 'true'. Thus, it is clear that assumption 1 holds in a world where everyone either knows the answer or guesses at random.

In reality, there is a broad spectrum between knowing something for sure and being clueless. There will also usually be people who think they know the answer even though they are wrong. Still, it is likely that more people will believe something is true if it actually is. For one, there will almost always be a group of people that is knowledgeable about the subject. Even if there is not, it is still probable that more people have heard of a certain fact if the fact is true.

There are some situations imaginable where assumption 1 does not hold. The first, which follows naturally from the above, is the scenario where nobody has any idea of whether the fact is true. The result would be that 50% of people are expected to give answer 'true', whether it is true or not. If this happens, either the respondents are not suited for the question, or the question is simply impossible to answer at that point in time. In this situation, no other method will be able to find the answer either.

[5] Note that this does not have to be a conscious decision. Respondents can base their predictions on all sorts of different things and not explicitly think of different worlds at all, but that all fits within this framework. If they take into account the possibility that they are wrong, and adjust their predictions accordingly, they are basing their predictions on two worlds. If they do not adjust their predictions for this possibility, they are basing their predictions on one world. This is true as long as respondents honestly try to make good predictions.

Another would be the scenario where some party, for example a lobbyist, is actively trying to convince people of the opposite fact than the truth. If an effective lobbyist would convince a majority of people that a true fact is false, while he would convince the same majority that the fact was true, if it actually were false, this would result in assumption 1 being violated.

In general it seems very reasonable to assume assumption 1.

## 3.2 Assumption 2

Assumption 2 can be stated in different degrees of strictness. The strict assumption 2, as Prelec et al. (2017b) use in their model for their theoretical proof, is as follows:

2a. Respondents know the distribution of answers in every world.

The only thing they do not know, is which world is the actual world (unless, of course, they are completely certain of their answer). If this assumption holds, and respondents truly know the distribution of answers in every world, this means that it is also possible to ask them for these distributions, instead of having them make predictions. The prediction question could then be replaced by the following two questions:

1. If Philadelphia is the capital of Pennsylvania, what percentage of people would answer 'Yes' to question 1?
2. If Philadelphia is **not** the capital of Pennsylvania, what percentage of people would answer 'Yes' to question 1?

Note that every respondent should give the same answers to these questions, since they know the distributions. Then, all the researcher has to do is look at what the actual distribution of answers to question 1 was, and see to which world this distribution corresponds. The advantage of this method is that participants do not have to calculate their overall predictions, they just have to state what they already know. The obvious disadvantage is that it requires an extra question.[6]

While the strict version of assumption 2 works well for a mathematical model, it seems highly unlikely that even some respondents really know the distributions for every possible world, let alone all respondents. A more realistic assumption would be:

2b. On average, respondents make accurate predictions of the distribution of answers in every world.

This seems much more reasonable. We are no longer assuming that anyone really knows the distributions, but instead assume we get reasonably accurate predictions on average. Even though the theoretical model from Prelec et al. uses the strict assumption 2a, the Surprisingly Popular method works with the more realistic assumption 2b as well. This is because the individual predictions really do not matter in the SP method, as long as the collective predictions per world are accurate.

[6] Actually, it requires respondents to give a distribution for every possible world. So if the main question has four answer possibilities instead of two, this method requires four distribution questions instead of one prediction question.

Even though 2b is a lot more realistic than 2a, one could still question if it is likely to really hold. Can people really make accurate assumptions about worlds they do not even believe are real, even collectively? Do people even imagine different worlds when they make their prediction? Since respondents in practice often do not act as ideal respondents would, we should consider that they may not. They could just base their prediction on the world that they think is more likely to be actual, or not think in terms of 'worlds' at all, and not consider that the distribution is likely to be different if their answer is wrong. The next version of assumption 2 is therefore as follows:

> 2c. On average, respondents make accurate predictions of the distribution of answers in the world they believe is actual.

While 2c is more likely to hold than 2b, it is only sufficient if all or most respondents base their prediction on just one world. If a group of respondents bases their predictions on two worlds, but only their predictions for one of those worlds is assumed to be accurate (in accordance with 2c), their predictions for the other world might influence their total predictions to such an extent that the SP method fails. It is important, therefore, that for those respondents who do base their predictions on different worlds, all of those predictions are accurate on average. For this reason, I form assumption 2d, which is both realistic and sufficient:

> 2d. On average, respondents make accurate predictions of the distribution of answers in the worlds on which they base their prediction.

I have designed an experiment to find out which method people use, and whether this can be influenced by asking them how certain they are of their answer.

# 4 Research question

As has become clear in the previous chapter, it is important to know how respondents make their predictions if we want to know what the weakest assumption is that needs to hold. If all respondents only base their predictions on the world they believe actual, all that is needed is that they make accurate predictions for those worlds on average (2c). If, however, some respondents base their predictions partly on worlds they do not believe to be real, those predictions need to be accurate on average as well (2d). If they are not, they could potentially throw off their main predictions enough to make the SP method fail. Since it assumes something about the capability of respondents to imagine things they do not believe, this assumption is quite a bit stronger and less likely to hold in practice. For this reason, it is worthwhile to find out whether the assumption is even needed.

Another benefit of knowing how people make predictions (e.g. whether they base them on different worlds), is that this knowledge can be used to make future models more accurate, increasing performance. Knowing how people make predictions is also the first step towards being able to influence this process. I am not talking about influencing the predictions themselves, but influencing how people make their predictions. An end goal could be, for example, to get all respondents to base their predictions on both worlds, using their confidence to assign weights to both worlds, just like the ideal respondents in the model of (Prelec et al., 2017b). On the other hand, if it turns out that assumptions 2b and 2d do not hold, but 2c does, we will probably want all respondents to base their predictions solely on the world they think is actual. After all, that is the only world for which they can make accurate predictions on average.

In short, knowing how people make predictions would enable us to increase the accuracy of future models, by taking into account how people think. It is also the first step to being able to influence the prediction making process, which would also give researchers more control and increase accuracy. Therefore, my main research question is:

*Do respondents in the SP method base their predictions on both possible worlds, or solely on the one they believe is actual?*

With the follow-up:

*Can the researcher influence this by asking for confidence in advance?*

As implied by the use of 'both' possible worlds, instead of 'all', I am limiting the scope of this thesis to binary questions.

## 4.1 Confidence

Humans are notoriously overconfident (Moore & Healy, 2008). This might have an effect on whether respondents imagine different worlds. A respondent who is quite sure of his answer may not realise that the distribution of answers will be different if he is wrong, for the simple fact that he does not stop to think about the fact that he may be wrong. Respondents who are really unsure about their answer, on the other hand, may be more likely to take all aspects of a question into account. They could, therefore, be more likely to base their prediction on both worlds.

If a respondent does not really think about the possibility that he is wrong, and the consequences thereof for his prediction, it might help to force him to make this possibility explicit. This can be done by asking him for his estimated probability of being correct. If he reports something below 100%, he just acknowledged that he might be wrong, even if he is still overconfident. The fact that he made this explicit might cause him to also take this into account when he needs to make his prediction.

As of yet, most of the above is speculation. In an effort to get more clarity about which factors actually influence the prediction process, and how researchers can purposefully steer respondents toward imagining the desired amount of worlds, I also to try to answer the following questions:

*Does a respondent's estimated probability of being correct influence the probability that they base their prediction on two worlds?*

*Does asking a respondent for his estimated probability of being correct influence the probability that they base their prediction on two worlds?*

The first question compares respondents with high confidence in their answers to those with low confidence, while the second compares respondents who report their confidence before they give their prediction to those who report it after.

## 4.2 Hypotheses

Regarding the main research question, it seems obvious that not all respondents will do the same. Seeing as it takes more effort to imagine two different worlds and base a prediction on both of them, than to base a prediction on the most likely world, I expect the latter to be more common. On the other hand, adjustments for the probability of being wrong could be made intuitively, or by approximation, reducing or negating the difference in effort required. Still, I cautiously expect that most respondents do not take into account different worlds when making their prediction.

### 4.2.1 The influence of confidence

Based on the above, I form the following hypotheses regarding the influence of confidence, defined as the estimated probability of being correct, on the probability of taking both worlds into account for a prediction:

H1: Higher confidence leads to a lower probability of basing a prediction on two worlds.

### 4.2.2 The influence of asking for confidence

I form the following hypothesis about the influence of asking for confidence:

H2: Asking respondents for their confidence in advance, increases the probability that they base their predictions on two worlds.

# 5 Experimental Design

To find out whether people consider different worlds when making their predictions, and whether (asking for) confidence influences this, I run an online experiment in the form of a survey. I use three treatments, all of which with roughly the same structure, and all consisting of two stages. I start by describing treatment 1 and the ideas behind it, after which I discuss the variations that form treatments 2 and 3. For the complete survey, see Appendix A. The artworks used, along with rough approximations of their market price, are shown in Appendix B. All selected artworks either have been sold in the past for more than one million euro's, or were for sale for roughly €10,000 at the time of the experiment.

## 5.1 Treatment 1

After a short explanation about the structure of the survey, the type of questions they will get and about the fact that the best predictor wins €25.-, the respondents start with stage 1.

### 5.1.1 Stage 1

In stage 1, I explain what the questions will be and that the percentages in question 2 have to add up to 100%. Respondents get to see five artworks in random order. For each artwork I ask them the same two questions:

1. Do you think the market price of the artwork above is less or more than €30,000?
2. What percentage of respondents do you think answered 'less than €30,000' for this artwork? And 'more than €30,000'?

For question 1, respondents select either answer, while question 2 is answered with two sliders with range [0, 100]. The second slider for question 2 would not be necessary for ideal respondents, but in practice it makes sure that respondents realise that both categories together are 100% (which is a requirement to continue). After answering the two questions, respondents proceed to stage 2.

### 5.1.2 Stage 2

In stage 2, I start by explaining that they will be asked about their estimated probability of their previous answers being correct, and why this should be at least 50%. I also explain that for every artwork I will ask them to consider two scenarios for which they will have to make their prediction again: one in which the artwork is cheap and one in which it is expensive.

Respondents then get to see the same five artworks as in stage 1. For every artwork, I first show them the answer that they gave to question 1 of stage 1, then I ask them three more questions:

In stage 1, for the artwork above, you answered: [less (more) than €30,000]

3. What is your estimated probability of being correct?
4. Consider the case in which the artwork is indeed worth less (more) than €30,000. In that case, what percentage of respondents do you think would answer 'less than €30,000' for this artwork? And 'more than €30,000'?

5. Consider the case in which the artwork is actually worth more (less) than €30,000. In that case, what percentage of respondents do you think would answer 'less than €30,000' for this artwork? And 'more than €30,000'?

Question 1 is a slider with range [50, 100]. Questions 2 and 3 both consist of two sliders with range [0, 100]. After stage 2, respondents have the option to submit their email address so that they are eligible for the €25.- price for the best predictor.

## 5.2 Terminology

- For all predictions, I only use the predicted frequency of 'less than €30,000'. The predicted frequency of 'More than €30,000' is never used.
- I also refer to 'less than €30,000' as 'cheap', and to 'more than €30,000' as 'expensive'.
- The answer to question 2 is a respondent's 'main prediction', or 'total prediction', abbreviated as P.
- The answer to question 3, the estimated probability of being correct, is a respondent's confidence, abbreviated as C.
- The answer to question 4, the prediction for the 'what if you're Right' scenario, is abbreviated as Pr.
- The answer to question 5, the prediction for the 'what if you're Wrong' scenario, is abbreviated as Pw.
- Combining Pr and Pw with C yields the calculated prediction, abbreviated as Pc. Pc=C*Pr+(1-C)*Pw.
- A respondent's Pr and Pw, for a given artwork, together form a prediction pair.

Even though the mcit assumption is about the frequency of answers in both worlds and not about individual predictions, we can check if a prediction pair 'follows' mcit. Following the reasoning behind the mcit assumption, a respondent who answered that an artwork is cheap should give a prediction pair such that Pr>Pw, and vice versa. A prediction pair that follows this rule is said to satisfy (strict) mcit. A prediction pair where Pr=Pw is said to satisfy weak mcit, as at least the respondent does not predict that the answer is less common if it is true. When an answer is predicted to be less common if it is true, the prediction pair violates (weak) mcit. Whether this should have any consequences, and whether such a violation is even necessarily irrational, is discussed in section 6.

## 5.3 In theory

Theoretically, if a respondent bases his predictions solely on the world he believes is actual, his answers to questions 2 and 4 will be the same: P=Pr(≠Pc). Since the prediction is based on one world, I call this **1-world thinking**. Note that these respondents do imagine different worlds, they just base their prediction solely on the more likely one. It could also be that they only imagine different worlds when prompted to do so.

For ideal respondents, as in Prelec et al. (2017b), the submitted answers will be such that P=Pc(≠Pr). This prediction process, where the prediction is based on two worlds, I call **2-world thinking**.

There will also be respondents whose answers to questions 4 and 5 are the same: Pr=Pw. This could be because they do not imagine two worlds at all, or it could be that they simply do not believe that the distributions would be different in both worlds. As discussed, the latter belief could be rational if, for example, one thinks that none of the other respondents will know the artwork. No matter what the reason is, the respondent is not imagining two worlds with different distributions, I therefore call this **non-world thinking**.

Because Pr=Pw, non-world thinking automatically results in Pc=Pr. This means that, for individual cases, three technically different types of respondents cannot be told apart:

1. Respondents who do not imagine two worlds, 'true' non-world thinkers.
2. 1-world thinkers who believe that, in this case, the distributions are the same.
3. 2-world thinkers who believe that, in this case, the distributions are the same.

Only when looking at several observations from the same respondents would it be possible to gain some insight into which of these three categories a respondent belongs to. Respondents belonging to 2 or 3 will likely show this when answering other questions, when they do believe that the distributions in both worlds will be different. When looking at observations in isolation, independent of the respondent's other answers, all three categories are classified as non-world thinking. Theoretically, all three respondents will also give P such that P=Pr=Pc.

Because all thinking types result in a different combination of answers, the survey should shed light on the proportion of respondents that use each thinking type. This will help answer the main research question. By combining these thinking types with the answers to question 3, the speculated connection between confidence and thinking type might become visible. To test whether having respondents report their confidence before they make their prediction has an influence on which thinking type they use, I make a comparison between treatments 1 and 2.

## 5.4 Treatment 2

Treatment 2 is nearly identical to treatment 1. The main difference is that question 3, the confidence question, is moved from stage 2 to stage 1, and is asked before the prediction (Q2). The respondent therefore first has to state his estimated probability of being correct, and only then make his prediction. I can then test whether this has the expected effect of increasing the probability that a respondent uses 2-world thinking.

In stage 2, I show respondents both the answer they gave to question 1 and the probability they submitted for question 3:

> In stage 1, for the artwork above, you answered: [less (more) than €30,000]
>
> Your estimated probability of being correct was: [..]%

This is followed by the same questions 4 and 5 as in treatment 1.

## 5.5 Treatment 3

Treatment 3 serves two purposes. The first is to check for order effects. The second is to get a better idea of the extent to which respondents can be steered towards 2-world thinking. The treatment is the same as treatment 2, only the prediction question from stage 1, which elicits P, is swapped with the prediction questions from stage 2, which elicit Pr and Pw. The question order now looks as follows:

### 5.5.1 Stage 1

1. Do you think the market price of the artwork above is less or more than €30,000?
2. What is your estimated probability of being correct?

3. Consider the case in which the artwork is indeed worth less (more) than €30,000. In that case, what percentage of respondents do you think would answer 'less than €30,000' for this artwork? And 'more than €30,000'?
4. Consider the case in which the artwork is actually worth more (less) than €30,000. In that case, what percentage of respondents do you think would answer 'less than €30,000' for this artwork? And 'more than €30,000'?

### 5.5.2 Stage 2

In stage 1, for the artwork above, you answered: [less (more) than €30,000]

5. What percentage of respondents do you think answered 'less than €30,000' for this artwork? And 'more than €30,000'?

### 5.5.3 Order effects

Whether a respondent gets classified as a non-, 1-, or 2-world thinker is based on the comparison between P, Pr and Pw. The objective is to find out if P was made with Pr and Pw in mind (2-world), or just with Pr in mind (1-world). To get that answer, it is important that Pr and Pw were made independently from P. To illustrate, imagine a respondent who based prediction P solely on Pr. When the respondent gets to stage 2 and is prompted to give Pr and Pw, he realises that he should have based P on both Pr and Pw. In an effort to seem more rational, he might change his Pr and/or Pw so that it seems that he had been 2-world thinking all along. If he is successful, the 1-world thinking respondent is classified as a 2-world thinker. Besides this scenario, there might be other reasons for respondents to, consciously or subconsciously, report a 'false' Pr and Pw, because they already made the main prediction earlier in the survey. Treatment 3 allows me to test if Pr and Pw are indeed being influenced in this way, by asking them before the main prediction.

### 5.5.4 Steering towards 2-world thinking

The second purpose of treatment 3 is to get an idea of the extent to which respondents can be steered towards 2-world thinking. The idea is the same as with the confidence question in treatment 2, only a step further. Respondents are forced to think about both worlds, and to give a prediction for both worlds, before they give their main prediction. With this set of questions, this order should be the most effective to get respondents to use 2-world thinking.

## 5.6 Sample

All respondents (94 in total) are either a student or recent graduate, recent being defined as 'in the past three years'. Respondents were recruited via three different 'channels':

1. Social: 53 respondents were recruited via social media groups centred around (mostly economic) studies at Erasmus University Rotterdam.
2. Strangers: 18 respondents were recruited at an online community dedicated to helping others with their research.[7]

[7] www.reddit.com/r/SampleSize

3. Personal: 23 respondents were recruited by means of a personal message with the request to participate.

Respondents from different channels could differ in their motivations to participate, and some might put in more effort than others. All three channels, therefore, had a separate link to the survey, which ran on www.Qualtrics.com. The separate links make it possible to test if there are differences between the groups of respondents.

# 6 Analysis

## 6.1 Identifying different types of thinkers

In this chapter I set out to identify the different types of thinkers among the respondents. I start with the individual observations, working under the assumption that a single respondent might use different types of thinking for different questions. After that, I try to label each respondent as a certain type of thinker. *Thinking type* refers to the type of thinking that is used in a certain instance, while *thinker type* refers to the type of thinker that a respondent is, which in turn says something about the type of thinking that they use most often.

To recap, I distinguish three different types of thinking:

- 2-world thinking: Imagining two different worlds with different distributions, and then basing the prediction on both of these distributions, combined with the estimated probability of being correct. With ideal respondents, this results in P=Pc (Pc=C*Pr+(1-C)*Pw).
- 1-world thinking: Imagining two different worlds with different distributions, but basing the prediction only on the world that is deemed most likely to be actual. Another possibility is that the respondent can imagine two different worlds, but only does so when prompted to do so. His main prediction is therefore only based on the world he believes to be actual. With ideal respondents: P=Pr≠Pw.
- Non-world thinking: Not being able to imagine different worlds at all, or predicting the same distribution for both worlds. Ideally: P=Pr=Pw.

Of course, real respondents are far from ideal. There are a lot of observations where both Pr and Pw are considerably lower (or both higher) than P, which is irrational. If there are two possible outcomes, the expected outcome should be either one of those, or somewhere in between them. The expected outcome cannot rationally be outside of the range between the two possible outcomes, but this happened quite often. It could be that respondents did not truly understand the meaning of the answers they gave, and did not notice the inconsistency. However, the most likely explanation is that, due to the difficulty of the tasks, predictions are inconsistent over time. That is to say, people's predictions would have been quite inconsistent even if they had been asked to give P twice, instead of P first and Pr+Pw after. This theory is strongly supported by the 123 observations where Pr=Pw≠P (as opposed to the 32 internally consistent observations where Pr=Pw=P). This is an obvious inconsistency: there are only two possible scenarios per prediction, so if someone gives the same prediction for both scenarios, their total prediction must be the same, regardless of their confidence. It is safe to assume that the majority of respondents, all of whom are at least bachelor students, would notice the inconsistency. Therefore, the most likely scenario is that a respondent who gets to stage 2 has forgotten what prediction he made in stage 1, makes a new prediction 'from scratch', and comes to a different conclusion than the first time.

Note that the problem is not that the respondent has forgotten what prediction he made in stage 1, this is perfectly fine. The problem is that the second time his prediction is (too) different. A possible cause of this is a lack of effort on the respondents' side, but it could just as well be that artworks are judged differently after a couple of minutes have gone by and some other works have been judged in the meantime. The respondent may also have changed his mind on how knowledgeable he expects the other respondents to be. In some cases, a respondent may even have remembered some important piece of knowledge in between the two judging

moments, whether it be the name of the artwork or the fact that it was featured on a television show recently. These are all things that could cause a respondent's P to be inconsistent with his Pr and Pw. While it may be impossible to determine the cause of a respondent's inconsistency, the fact that these time inconsistencies exist, at least in this experiment, is undeniable.

The fact that respondents' predictions are inconsistent over time means that we cannot simply look at which of the three equalities applies to each observation, but will instead have to look at which of the three fits best.

As mentioned, I first analyse the data under the assumption that a respondent can use any type of thinking for every question, independent of the type that he used for other questions.

### 6.1.1 Thinking types

Let us start with a basic analysis. Non-world thinking can be identified by the prediction pair Pr=Pw. For the other observations, I simply look whether P is closer to Pr or to Pc. Figure 1 shows the distribution of observations over the three thinking types. It also shows that in 2% of cases, it is not possible to distinguish between 1- and 2-world thinking. This happens when a respondent submits a confidence of 100%, resulting in Pc=Pr.

We now have this distribution for the entire sample. We could do the same for every treatment individually and test whether different thinking types are more prominent in different treatments. However, let us first take a closer look at the data to check if Figure 1 really paints an accurate picture.

Figure 2 shows the distribution of the differences between |P-Pr| and |P-Pc|, for each observation that we just classified as being either 1- or 2-world thinking. When a respondent uses 2-world thinking, P will be closer to Pc than to Pr. The difference will

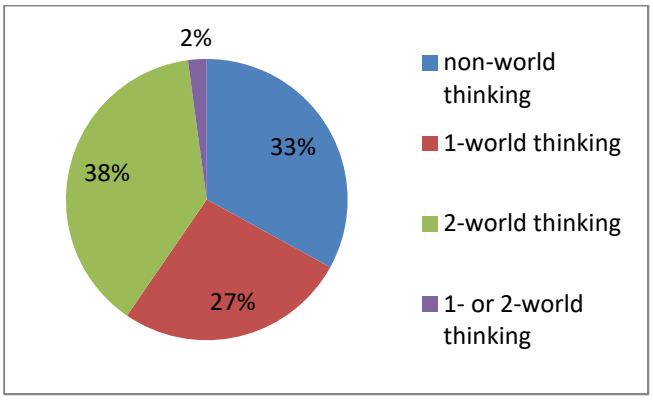



Figure 1 Distribution of observations over the three thinking types. (n=470)

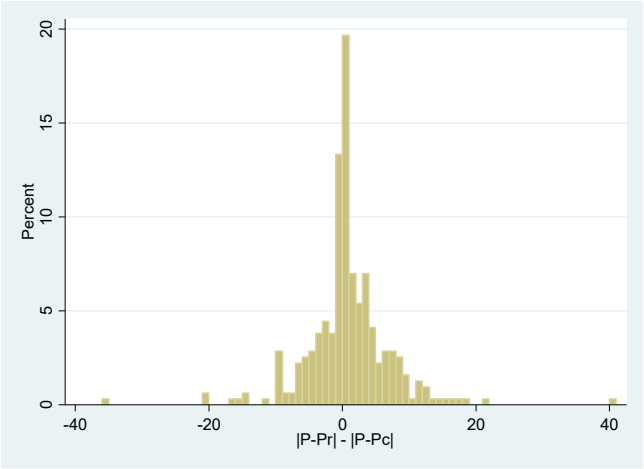



Figure 2 Distribution of the difference between |P-Pr| and |P-Pc| for each observation where 1- or 2-world thinking was used (Pr≠Pw). A positive number means that P is closer to Pc than to Pr, which is an indication that 2-world thinking was used. (n=315)

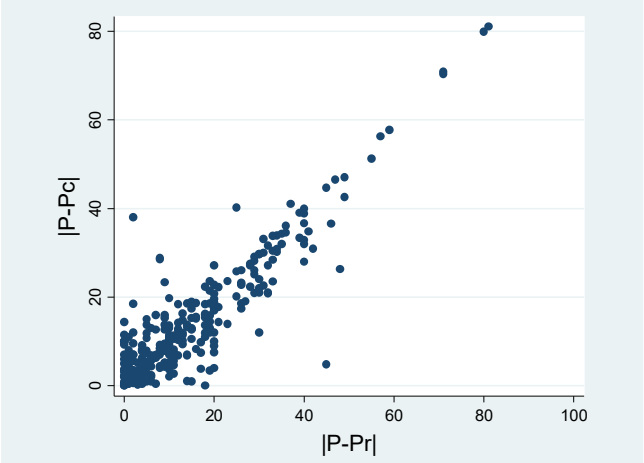



Figure 3 Scatter plot for all observations where Pr≠Pw. (n=315)

then be positive, showing up on the right side of the graph. If a respondent uses 1-world thinking, the difference will be negative.

Looking at the graph, the observations are centred roughly around 0. In other words, P is often roughly the same distance from Pc as from Pr. This can happen when P is almost exactly in between Pc and Pr, but the majority of the time it is because Pc and Pr are close together. This, in turn, has two possible causes.

One possibility is that Pr and Pw are close together, meaning that the respondent thought that there would not be a big difference in the distribution of answers, whether the artwork was expensive or cheap. A respondent might assume this if he thinks that the vast majority of other respondents do not know the artwork or its artist, causing them to judge the artwork purely by how it looks. This would result in similar distributions, regardless of the true price of the artwork.

The other possibility is that the confidence of the respondent was high. This causes Pr to be weighted heavily when calculating Pc. Either of these possibilities will cause Pc and Pr to be similar. Both of them were quite common in the sample.

So is it a problem when the difference (|P-Pr|-|P-Pc|) is close to 0? It does not have to be. If |P-Pr|=3 and |P-Pc|=0, then one can quite confidently conclude that 2-world thinking was used, even though the difference is only 3pp. If, on the other hand, |P-Pr|=20 and |P-Pc|=17, one might not want to draw that conclusion. The difference is once again 3pp, but there is a large discrepancy between the prediction made and the prediction one would have made if either 1- or 2-world thinking had been 'properly' applied to Pr and Pw, without time inconsistencies. Since Pr≠Pw, we can conclude that the respondent did not use non-world thinking. Beyond that, however, it is not possible to say with reasonable certainty what type of thinking this respondent applied or tried to apply. The problem is that both of these observations (both with a difference of 3pp) show up in the same way in Figure 1Figure 2, namely as an instance of 2-world thinking, since the analysis was based purely on |P-Pr|-|P-Pc|.

Even if, for a given observation, the difference is relatively large, say 10pp, so that one might think that the respondent clearly used 2-world thinking, it could be the case that |P-Pr|=50 and |P-Pc|=40. The conclusion that the respondent used 2-world thinking would be questionable at best. It is key, therefore, that we do not only look at the differences shown in Figure 2, |P-Pr|-|P-Pw|, but also at the separate distances |P-Pr| and |P-Pw|.

Unfortunately, Figure 3 shows that it is quite a common occurrence that both |P-Pr| and |P-Pc| are large. This means that many of the observations that, in Figure 1, were classified as either 1-world thinking or 2-world thinking, are cases where we cannot really draw this conclusion. To draw accurate conclusions, we should only take into account observations where one of the two distances is small, relative to the other. This could mean that one is 3pp and the other 0pp, or, for example, that one is 50pp and the other is 20pp. The larger both distances are, the bigger the difference between them should be to be able to state with some certainty that a certain type of thinking is used.

There is one other thing we can, and should, look for when interpreting the data. Following the reasoning behind assumption 1, the assumption that an answer will be more common if it is true (mcit), we might want to

label all observations that do not follow at least the weak version of this rule as being irrational. One could argue that the respondents cannot be expected to know this rule, or figure it out on their own, and that not doing so does not make them irrational. I agree. However, is there a rational reason for giving a prediction that states that an answer will be less common if it is true? If a respondent does not think of the reasoning behind mcit, should he then not predict Pr=Pw? Even if the respondent thinks that not a single respondent knows the answer, which is one of the possible reasons for strict mcit not to hold, he should predict Pr=Pw. I would almost argue, therefore, that any prediction that violates weak mcit is irrational. I say almost, since in this case I can think of one rational argument to give such a prediction.

A commonly held opinion among laymen seems to be that the prices of artworks are illogical or ridiculous. Artworks often seem to be expensive simply because of who the artist is, not because the work is beautiful or difficult to create. Phrases such as 'my 4-year-old nephew could make that' illustrate this feeling. Meanwhile, there are millions of beautiful artworks which do not nearly cost €30,000. Especially abstract art could be anything in the eyes of most laymen, from a million euro painting by a famous artist, to something that was in fact made by a 4-year-old playing around. Combine this with the fact that the respondents know that they are in an experiment where the researcher handpicked the artworks, and respondents may well get the idea that maybe the researcher will try to trick them. If this is the case, the researcher would be like the lobbyist discussed in the discussion of assumption 1. If a respondent thinks the researcher will be successful in tricking a significant part of the respondents, he could rationally make the prediction that an answer will be less common if it is true. Whether it is more likely that the majority of the predictions that violate weak mcit did so because of this reasoning or because of irrational thinking, is debatable. However, it is definitely something to keep in mind.

#### *6.1.1.1 Fixing the analysis*

To recap, there are three issues that need to be taken into account:

1. If both |P-Pr| and |P-Pc| are too large, we cannot properly compare stage 1 to stage 2.
2. Regarding |P-Pr| and |P-Pc|, one should be small relative to the other to be able to conclude whether 1-world thinking or 2-world thinking was used. The difference between the two distances has to be larger if both distances are larger.
3. It is possible that most predictions that violate the weak version of the mcit assumption were made irrationally.

To adjust the basic analysis for the issues above, I redo the analysis using the following rules (Rule set 1):

1. To classify as either 1-world or 2-world thinking, one of the two distances has to be at least twice as large as the other: |P-Pr|≥2*|P-Pc| or |P-Pc|≥2*|P-Pr|, and
2. the difference between the two distances has to be at least 0.5.
3. At least one of the distances has to be smaller than or equal to 20pp.
4. Prediction pairs have to satisfy weak mcit.

Rules 1 and 2 are for the classification as 1-world thinking or 2-world thinking. P has to be at least twice as close to Pr as it is to Pc, or vice versa. This adequately fulfils the need for the difference to be larger when the distances are larger. If |P-Pr|=2, then |P-Pc| has to be at least 4pp to conclude that 1-world thinking was used, while if |P-Pr|=20, |P-Pc| has to be at least 40pp. The second rule makes sure that answers like |P-Pr|=0 and

|P-Pc|=0.2, and answers like |P-Pr|=0.3 and |P-Pc|=0.6 do not get qualified as 1-world thinking or 2-world thinking, since I think they are too close to each other to draw the conclusion with reasonable certainty.

Rules 3 and 4 are requirements for all observations. If these are not met, the observation is labelled irrational. For now, I assume that all violations of weak mcit are irrational. Later, I assume that all of these observations are, in fact, rational, and check how this affects the results.

Appendix C visualises how observations are classified as a certain thinking type.

Figure 4, compared to Figure 1, shows an enormous increase in the size of the '1- or 2-world thinking' category, at the cost of the 1-world thinking and 2-world thinking categories. This is of course due to rules 1 and 2 correcting the fact that a lot of observations were classified as either type of thinking, without a reasonable indication to make that claim. Another noticeable difference is that non-world thinking now makes up 45% of the sample, as opposed to 33%. This is in great part due to the 86 observations that were deemed irrational based solely on rule 4. While rule 3 affects cases of both world and non-world thinking, rule 4 does not: since non-world thinking satisfies weak mcit by definition, rule 4 only affects cases of world thinking.

Figure 6 shows the new scatter plot. The red lines mark the area where rule 4 is effective. The blue lines mark rules 1 and 2. The area beneath the bottom blue line contains instances of 1-world thinking, while the area to the left of the top blue line contains cases of 2-world thinking.

#### *6.1.1.2 Distribution per treatment*

With the basic analysis adjusted to fit rule set 1, we can test if the three treatments have different

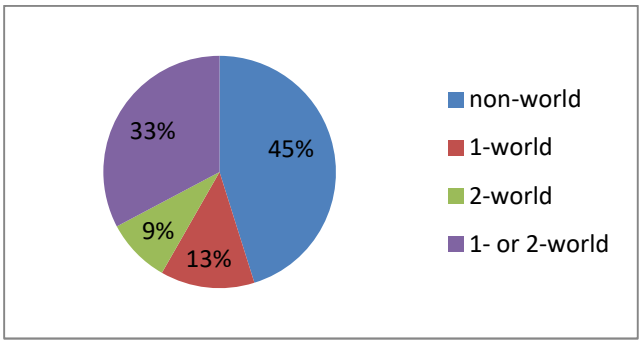



Figure 4 Distribution over non-, 1- and 2-world thinking of all observations that satisfy rule set 1. 94 observations were labelled irrational and left out because of rule 3 (30 of which also violated rule 4), then 86 more were left out because of rule 4. (n=290)

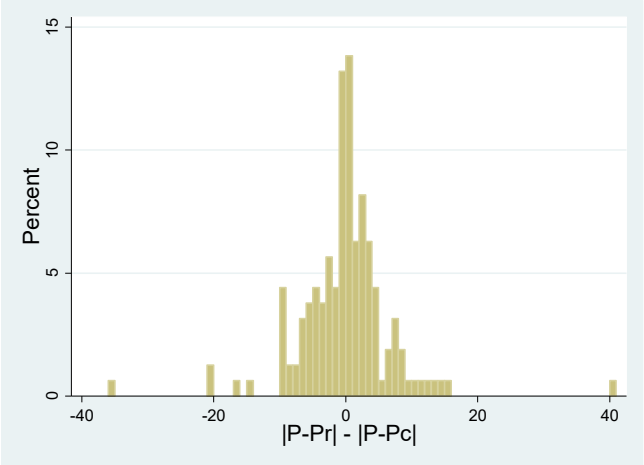



Figure 5 Distribution of the difference between |P-Pr| and |P-Pc| for each observation where 1- or 2-world thinking was used that satisfies rule set 1. (n=159)

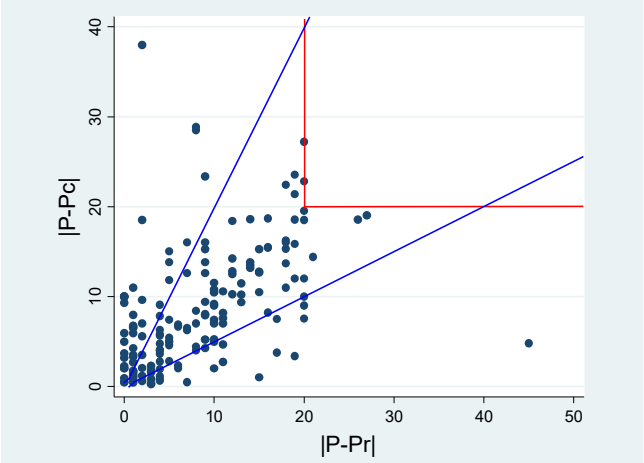



Figure 6 Scatter plot of all observations that did not use non-world thinking and that satisfy rule set 1. The blue lines visualise rules 1 and 2, the red lines visualise rule 3. (n=159)

distributions of thinking types. Table 2 and Figure 7 show the distribution of thinking types within each treatment. Surprisingly, treatment 1 has a lower proportion of non-world thinking than the other treatments. Furthermore, treatment 3 is the only treatment where the frequencies of 1-world and 2-world thinking seemingly differ considerably. This would have been in line with expectations of 2-world thinking had been the most common, but the opposite is true.

To test if the differences in the relative frequencies of non-world thinking are significant, I use Fisher's exact test. Using a 2x2 table as shown in Table 3, I first compare treatments 1 and 2. The difference is significant at the 5% level **(p=0.024)**[8]. I test the other pairs in the same way, neither the difference between treatments 1 and 3 (p=0.139), nor the difference between treatments 2 and 3 (p=0.405) is significant.

In the same manner, I test if 1-world thinking is significantly more common in one treatment than in another. The difference between the proportions of 1-world thinking in treatments 2 and 3 is weakly significant (p=0.092), the other differences are not significant.

For the proportions of 2-world thinking in each treatment, I find that there is a significant difference between treatments 1 and 3 **(p=0.038)**, but not between the other pairs.

I run the same test to check if there is a significant difference between the relative frequencies of rule 3 and rule 4 violations between the treatments. Neither rule is significantly more violated in any one treatment than it is in another, nor does the sum of rule 3 and rule 4 violations (=total proportion of

| thinkingtype2 | treatment 1 | 2 | 3 | Total |
|---|---|---|---|---|
| 0 | 29 | 52 | 50 | 131 |
| | 34.94 | 52.53 | 46.30 | 45.17 |
| 1 | 12 | 8 | 18 | 38 |
| | 14.46 | 8.08 | 16.67 | 13.10 |
| 2 | 11 | 10 | 5 | 26 |
| | 13.25 | 10.10 | 4.63 | 8.97 |
| 3 | 31 | 29 | 35 | 95 |
| | 37.35 | 29.29 | 32.41 | 32.76 |
| Total | 83 | 99 | 108 | 290 |
| | 100.00 | 100.00 | 100.00 | 100.00 |

**Table 2 Distribution of thinking types for each treatment. 0=non-world. 1=1-world. 2=2-world. 3=1- or 2-world thinking.**

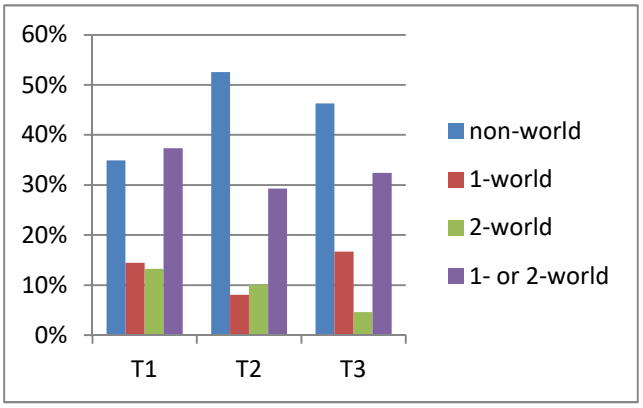



**Figure 7 Relative frequencies of thinking types within each treatment. (N=83; 99; 108)**

| nonworldthinking2 | treatment 1 | 2 | Total |
|---|---|---|---|
| 0 | 54 | 47 | 101 |
| | 65.06 | 47.47 | 55.49 |
| 1 | 29 | 52 | 81 |
| | 34.94 | 52.53 | 44.51 |
| Total | 83 | 99 | 182 |
| | 100.00 | 100.00 | 100.00 |

**Table 3 0=world thinking, 1=non-world thinking, using rule set 1. Relative frequencies per column are shown below the absolute.**

[8] 2-sided

irrational respondents) differ significantly per treatment.

#### *6.1.1.3 Dropping rule 4*

In the analysis above, 86 observations were left out solely because they violated weak mcit. The purpose of this was to confine the analysis as much as possible to rational responses only. As discussed, mcit violations are not necessarily irrational. If the respondent thinks that a significant part of other respondents will be wrong, no matter what the right answer is, for example because the researcher picked 'misleading' artworks, an mcit violation can be rational. Even if the respondent violated mcit without a rational reason, one could argue that this does not affect their ability to apply 1-world or 2-world thinking. For these reasons, I run the same analysis as before, this time without rule 4.

Figure 8 shows the new overall distribution of thinking types. Relatively speaking, there is less non-world thinking. This is because rule 4 did not filter out non-world thinking to begin with, so while the other categories increased in absolute terms, this one did not.

Unsurprisingly, Figure 9 shows that most of the newly approved observations fall within the two blue lines, meaning they are classified as '1- or 2-world thinking'. Still, there are now 10 more instances of 1-world thinking and 12 more instances of two-world thinking.

Figure 10 shows the new relative frequencies within each treatment. The difference between the proportions of non-world thinking in treatments 1 and 2 has increased, it is now significant at 1% **(p=0.004)**. Non-world thinking is now also significantly more common in treatment 2 than in treatment 3 **(p=0.048)**.

As opposed to when we still had rule 4, 2-world thinking is no longer significantly more common in

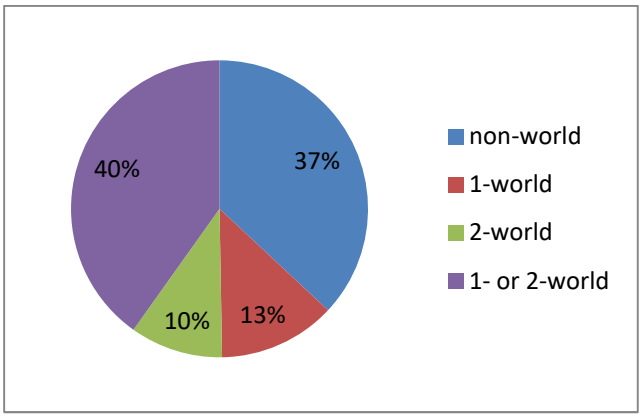



Figure 8 Distribution of thinking types after dropping rule 4 from rule set 1. (n=376)

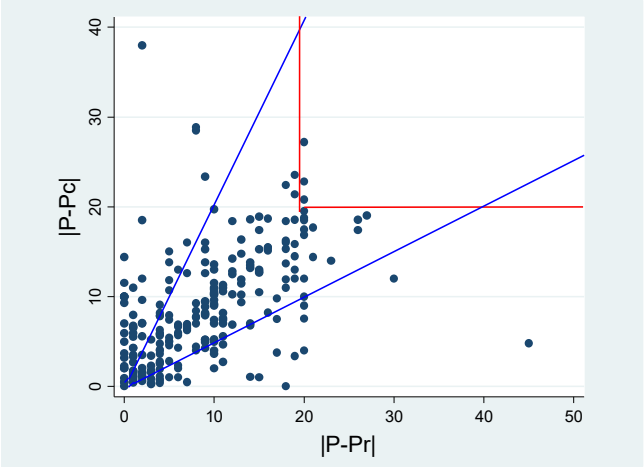



Figure 9 Scatter plot after dropping rule 4 from rule set 1. (n=376)

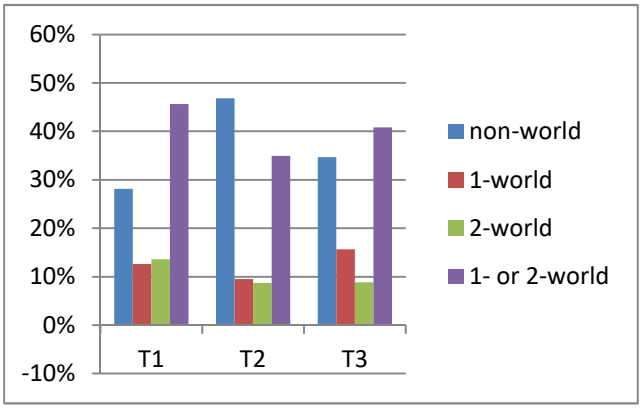



Figure 10 Relative frequencies of thinking types within each treatment after dropping rule 4 from rule set 1. (n=103; 126; 147)

treatment 1 than in treatment 2 (p=0.301). There are no other significant differences.

#### *6.1.1.4 Conclusion*

We have made the first step in answering the research question. In 37-45% of all cases, non-world thinking was used. 1-world and 2-world thinking were used in at least 9-10% and 13% of all cases, respectively, so a slight majority for 2-world thinking. However, 33-40% of all observations are classified as '1- or 2-world' thinking, meaning that while it is clear that these respondents imagined different worlds, it is not possible to say with reasonable certainty whether they based their prediction on one or on two worlds.

If we assume that the ratio between 1- and 2-world thinking is the same in the unknown category as for the known cases, we can make an approximation of the distribution.

When we exclude the weak mcit violations (rule set 1), 33% gets divided according to a ratio of 9:13. This results in the following distribution: 45% non-world thinking, 22.5% 1-world thinking, and 32.5% 2-world thinking, as shown in Figure 11.

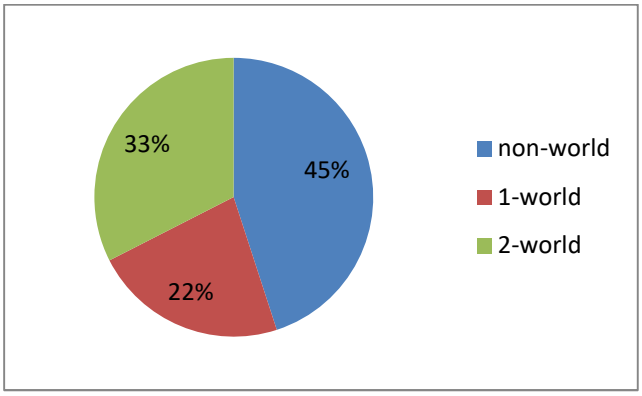



**Figure 11 Approximate distribution of observations over the three thinking types when violations of weak mcit are excluded.**

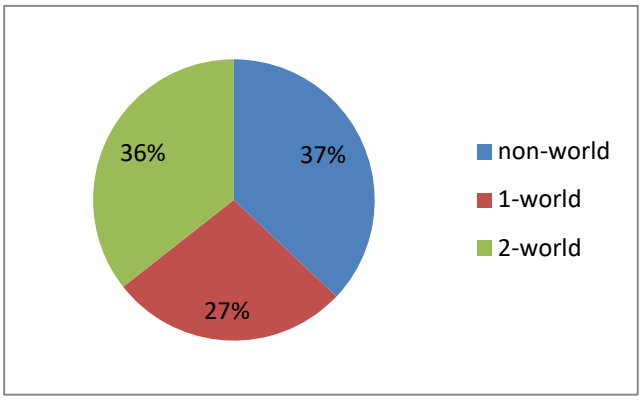



**Figure 12 Approximate distribution of observations over the three thinking types when violations of weak mcit are included.**

With mcit violations included (dropping rule 4), 40% gets divided according to a ratio of 10:13. This results in the following distribution: 37% non-world thinking, 27.4% 1-world thinking, and 35.6% 2-world thinking, see Figure 12.

We did not find any evidence for H2. So far, H1 has not been tested for.

To recap the significant findings:

- Non-world thinking is significantly more common in treatment 2 than in treatment 1.
- 2-world thinking is significantly more common in treatment 1 than in treatment 3 if observations that violate weak mcit are excluded from the analysis.
- Non-world thinking is significantly more common in treatment 2 than in treatment 3 if observations that violate weak mcit are included in the analysis.

None of these findings are in line with expectations. Of course, it could well be that these results are only significant because of the fact that the five observations per respondent were treated as independent observations. With certain thinking types being relatively uncommon, one consistent respondent, for example

one who uses 2-world thinking five times, could have a large impact.[9] After all, all five of his responses will be in the same treatment. To try and solve this issue, I set out to classify each respondent as a certain type of thinker.

### 6.1.2 Thinker types

I now try to classify each respondent as a certain type of thinker, using the same three types. I use rule set 1 without rule 4 to classify observations, just like in the last analysis.[10] Then, I use the following rules to classify respondents as types of thinkers (Rule set 2):

1. A respondent with three or more instances of the same type of thinking, is classified as that type of thinker.
2. A respondent with more instances of 1-world thinking than 2-world thinking, which, together with all instances of '1- or 2-world' thinking, add up to three or more possible instances of 1-world thinking, is classified as a 1-world thinker.
3. A respondent with more instances of 2-world thinking than 1-world thinking, which, together with all instances of '1- or 2-world' thinking, add up to three or more possible instances of 2-world thinking, is classified as a 2-world thinker.
4. A respondent with the same number of instances of 1-world thinking as 2-world thinking, which, together with all instances of '1- or 2-world' thinking, add up to three or more instances of world thinking, is classified as a '1- or 2-world' thinker.

To illustrate: A respondent with one case of 1-world thinking, two cases of 1- or 2-world thinking, and two cases of non-world thinking would be classified as a 1-world thinker as per rule 2. A respondent with one case of 1-world thinking, two cases of 2-world thinking, one case of 1- or 2-world thinking and one irrational (left out) observation would be classified as a 2-world thinker, as per rule 3.

#### *6.1.2.1 Pr≈Pw*

Since we are now looking at all five of a respondents' answers, instead of viewing all of them in isolation, there is one more thing I feel is important. Consider the following prediction pair: Pr=60, Pw=59. Is the respondent who submits these predictions imagining two different worlds? Or should this be treated as Pr=Pw(=60), a case of non-world thinking?

In principle it should be assumed that each respondent gave the answer that they wanted to give. Assuming that a respondent made a mistake and meant something else should not be done lightly, especially when the answer given can be explained rationally: in the case of Pr=60, Pw=59, it could be that the respondent thinks that almost no other respondents will know the artwork, and that therefore it will only make a slight difference whether it is cheap or expensive. Still, I make the case that in some instances, cases like these should be treated as non-world thinking. Let me first explain why I think it is possible that answers like these were submitted by mistake, and then set out in which cases I think we should assume that this is the case, so that they should be 'corrected'.

---

[9] This would not have that large of an impact if all or most respondents were consistently using the same thinking type, but this is not the case.

[10] I include mcit violations both because of the arguments stated earlier, and because it increases the amount of observations considerably.

In the experiment, the predictions had to be given by means of a 100 point slider, as opposed to by typing the number. It is plausible, therefore, that there are respondents who made predictions Pr=Pw in their mind, which due to sloppiness and laziness were submitted as Pr≈Pw. The difficulty of the questions makes it even more likely that a respondent, who in his mind predicts Pr=60, thinks 'close enough' when the (quite sensitive) slider stops on 58. After all, the 60 was only a rough prediction, and as far as he knows, 58 is just as likely to be correct. Ideally this would not happen, but given the difficulty of the task, combined with the relatively low incentives, it is realistic to assume that at least some respondents think and act in this manner.

If the above sounds implausible, or at least not likely enough to act upon, consider this example from the experiment. Respondent 37 gave the following five prediction pairs (Pr,Pw):

(56,56) (79,79) (89,89) (60,60) (22,23)

One possibility is that the respondent used non-world thinking four times, while for the last case they predicted a 1pp difference between the two worlds. The other possibility is that the respondent does not imagine different worlds with different distribution at all, and the 22 and the 23 were meant to be the same number, whether that be 22, 23 or even 25. The latter possibility seems much more likely, and therefore I classify cases like this as non-world thinking. Besides that classification, I do not change the data in any way. For other tests and calculations, I use the predictions as submitted by the respondents.

With 'cases like these', I mean respondents that gave Pr=Pw for some artworks, and Pr≈Pw for all others. I judge a prediction pair to be approximately equal (Pr≈Pw) if the difference between the two predictions is at most 3 points. This is the widest gap of which I would say it is safe to assume that the predictions are meant to be equal. In addition to that, if there are more than two Pr≈Pw pairs, I look at the corresponding answers and check if the inequalities are internally consistent. If all of them satisfy mcit, or if all of them violate mcit, there is a higher probability that the pair was unequal on purpose. In those cases, I do not classify them as non-world thinking.

In accordance with the requirements above, predictions from nine respondents are classified as non-world thinking. Now, let us have a look at the distribution of thinker types in accordance with rule set 2.

#### *6.1.2.2 Distributions*

Figure 13 shows the overall distribution of respondents over the thinker types. Nine respondents did not classify as a thinker type because three or more of their responses were irrational according to rule set 1, while seven respondents had so much diversity in their thinking types that they did not qualify as any type of thinker. Figure 14 shows the distribution over the different thinker types within each treatment.

Most noticeable in both thinker type graphs is the enormous increase in size of both the 1-world and the 2-world category, compared with their thinking type equivalents in the previous analysis (see Figure 10). This is because of rules 2 and 3, as illustrated in the examples above.

I test if certain types of thinkers are significantly more common in certain treatments, again using Fisher's exact test. Despite the fact that the difference still looks large, non-world thinkers are not significantly more common

in treatment 2 than in treatment 1 (p=0.118). Recall that this difference for non-world thinking was significant at the 1% level when we looked at the individual observations.

The other noticeable difference in Figure 14 is between the proportions of 1-world thinking in treatments 1 and 2. This difference is also not significant (p=0.157).

#### *6.1.2.3 Conclusion*

In the previous analysis, observations were treated as independent. This resulted in a distribution of thinking types, and gave an indication that asking respondents for their confidence in advance has no significant influence. In this last analysis, on the other hand, we assigned one thinker type to each respondent, assuming that each individual respondent generally just uses one thinking type. This resulted in a distribution of 31% non-world thinkers, 20% 1-world thinkers, 22% 2-world thinkers, and 27% 1- or 2-world thinkers, see Figure 13.

Dividing the 1- or 2-world thinkers according to the ratio of 'known' 1-world and 2-world thinkers results in 32.9% 1-world thinkers and 36.1% 2-world thinkers. In other words, respondents are roughly equally divided over the three thinker types, with a slight majority of 2-world thinkers, see Figure 15.

We have not found any evidence in favour of H2. H1 has not yet been tested for.

The significant results from the thinking type analysis from before are not supported by this analysis of thinker types. While some of the same differences between treatments are visible, they are no longer significant. This suggests that these results were indeed driven by the fact that observations were treated as independent. However, note also that there is now only one observation per respondent, whereas before there were five. This

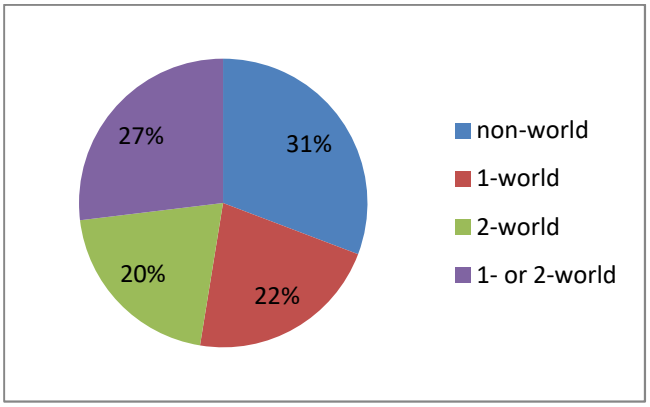



Figure 13 Distribution of respondents over thinker types. (n=78)

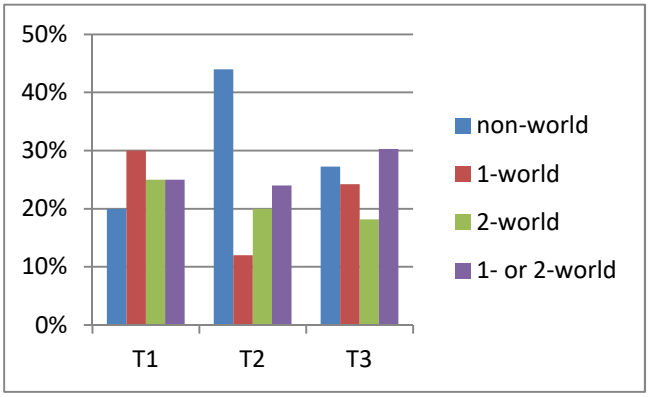



Figure 14 Distribution over thinker types within each treatment. (n=20; 25; 33)

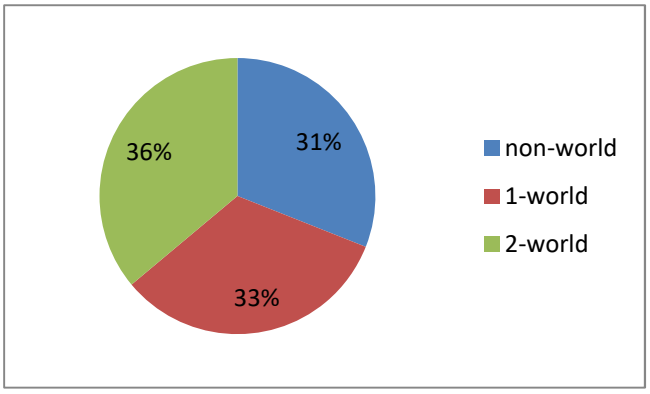



Figure 15 Approximate distribution of respondents over the three thinker types.

dramatic decrease in observations makes it harder to get significant results.

Next, let us test the hypotheses by means of regressions.

### 6.1.3 Regressions

Besides the nonparametric tests, I run some regressions to test if the assigned treatment, among other things, has an effect on the type of thinking a respondent uses. I start with a logistic regression to test which factors have an influence on the probability of a respondent using (non-)world thinking. Table 4 shows the logit regression with the treatment as the only independent variable. The coefficients are not significant, which is in line with expectations after the results of the Fisher's exact tests in the analysis to the distribution of thinker types. When running the regression without clustering the data by ID, treating the observations as independent, the coefficient of treatment 2 is significant. This is in line with the nonparametric tests on the distribution of thinking types.

Table 5 shows a more extensive logit model. Among other things, artwork is added to test if respondents are more likely to use non-world thinking for some artworks than others. Channel is added with the idea that, depending on how and where they discovered the survey, some groups of respondents may be more likely to put in more effort than others, and therefore be less prone to non-world thinking. None of the added variables are significant.

Next, I run a multinomial logistic regression to test if any of the independent variables have a significant effect on the probability that a respondent uses 2-world thinking relative to the probability that they use 1-world thinking. I use the same independent variables as in Table 5, none of the coefficients are significant. Confidence has no significant effect on the probability of using 2-world thinking relative to

| nonworldthinking | Coef. | Robust Std. Err. | z | P>\|z\| |
|---|---|---|---|---|
| treatment | | | | |
| 2 | .618414 | .4357391 | 1.42 | 0.156 |
| 3 | .1170382 | .4112558 | 0.28 | 0.776 |
| _cons | -.7605885 | .3001119 | -2.53 | 0.011 |

Table 4 Logit regression, clustered by ID.

| nonworldthinking | Coef. | Robust Std. Err. | z | P>\|z\| |
|---|---|---|---|---|
| treatment | | | | |
| 2 | .5104674 | .4465959 | 1.14 | 0.253 |
| 3 | .1063849 | .4338895 | 0.25 | 0.806 |
| artwork | | | | |
| 2 | .3033973 | .2137541 | 1.42 | 0.156 |
| 3 | -.1316446 | .200911 | -0.66 | 0.512 |
| 4 | .0747307 | .2318011 | 0.32 | 0.747 |
| 5 | .1517024 | .2150055 | 0.71 | 0.480 |
| channel | | | | |
| strangers | .6042599 | .5018677 | 1.20 | 0.229 |
| pers | -.2422834 | .3910738 | -0.62 | 0.536 |
| price | | | | |
| More than €30,000 | -.2696862 | .2061863 | -1.31 | 0.191 |
| duration | .0000844 | .0001385 | 0.61 | 0.542 |
| confidence | -.0069334 | .0106103 | -0.65 | 0.513 |
| _cons | -.2951129 | .7707118 | -0.38 | 0.702 |

Table 5 Logit regression, clustered by ID.

| 1 | (base outcome) | | | |
|---|---|---|---|---|
| 2 | | | | |
| treatment | | | | |
| 2 | .5107477 | .9786031 | 0.52 | 0.602 |
| 3 | -.3426691 | .8320305 | -0.41 | 0.680 |
| channel | | | | |
| strangers | .4441083 | 1.336388 | 0.33 | 0.740 |
| pers | -.1409781 | .8317112 | -0.17 | 0.865 |
| duration | -.0008964 | .0008403 | -1.07 | 0.286 |
| confidence | -.0015478 | .0154616 | -0.10 | 0.920 |
| _cons | .6116319 | 1.326298 | 0.46 | 0.645 |

Table 6 Multinomial logit regression, clustered by ID. 1=1-world thinker; 2=2-world thinker.

1-world thinking, nor on the probability of using 2-world thinking relative to non-world thinking. As such, I find no evidence in support of the hypothesis that a higher confidence decreases the probability of 2-world thinking (H1).

To do a similar regression for the probability that a respondent is a 2-world thinker relative to the probability that they are a 1-world thinker, some independent variables need to be left out. After all, all respondents get to see all artworks, and they are the same type of thinker for all of them, since they get classified as a certain type of thinker based on all five of their answers. It would not make sense, therefore, to try to discern the effect that different artworks, and the price categories of artworks, have on thinker types. The part of the regression concerning the probability of being a 2-world thinker relative to the probability of being a 1-world thinker, is shown in Table 6. None of the coefficients are significant. For the probability of being a 2-world thinker relative to the probability of being a non-world thinker, none of the coefficients are significant either. The results are the same when confidence is replaced by the respondent's average confidence. As such, I again find no evidence for the hypothesis that having higher confidence decreases the probability of basing predictions on two worlds (H1).

None of the regressions showed a significant effect of treatment on the dependent variables. This is consistent with the results from the nonparametric tests regarding the effect of asking in advance for a respondent's confidence. I find no evidence for the hypothesis that asking for their confidence in advance increases the probability that respondents base their predictions on two worlds (H2).

#### *6.1.3.1 Conclusion*

The regressions provide no support for either hypothesis. Higher confidence does not seem to decrease the probability of 2-world thinking, nor the probability of being a 2-world thinker. As for H2, the regressions confirm the earlier results: asking respondents for their confidence in advance does not increase the probability that they base their predictions on two worlds.

Recall that, in the thinking type analysis, non-world thinking was significantly more common in treatment 2 than in treatment 1. In the thinker type analysis, however, non-world thinkers were not significantly more common in treatment 2 than in treatment 1, although the difference was still clearly visible in Figure 14. It was not entirely clear which of two things caused these different results.

The result in the thinking type analysis could be misleading, the significance of the effect being a result of observations wrongfully being treated as independent. It could also be the case, however, that the effect is real, but that the thinker type analysis did not have enough observations to get significance.

The regressions above provide the answer. When the observations are not clustered by ID, the same significant result as in the thinking type analysis is found. However, when the observations are clustered by ID, the effect is no longer significant. Since clustering is meant to deal with cases like this, correcting for dependence within groups (respondents), and it does not reduce the number of observations, it is highly likely that the significance was the result of observations being treated as independent.

## 6.2 Order effects

As explained in the experimental design, one purpose of treatment 3 is to test for order effects. In treatments 1 and 2, Pr and Pw could be affected by the fact that the respondents already gave P for the same artworks. The submitted Pr and Pw would then not be the 'true' Pr and Pw. If this is the case, Pr and Pw in treatment 3 should have a different distribution than Pr and Pw in treatments 1 and 2. I use a Mann-Whitney U-test to test whether this is true. Table 7 shows the table for the comparison of the distributions of Pw in treatments 1 and 3. There is no significant difference (p=0.493). I run the same test comparing treatments 2 and 3, and do the same for the distributions of Pw. None of the distributions are significantly different, meaning that there is no evidence of order effects that influence Pr and Pw.[11]

| treatment | obs | rank sum | expected |
|---|---|---|---|
| 1 | 135 | 21878.5 | 21330 |
| 3 | 180 | 27891.5 | 28440 |
| combined | 315 | 49770 | 49770 |

**Table 7 Table for the Mann-Whitney U-test comparing the distributions of Pw in treatments 1 and 3. p=0.493**

| price | obs | rank sum | expected |
|---|---|---|---|
| cheap | 233 | 70157 | 54871.5 |
| expensive | 237 | 40528 | 55813.5 |
| combined | 470 | 110685 | 110685 |

**Table 8 Mann-Whitney U-test comparing P's of respondents who answered 'cheap' to P's of those who answered expensive'. p=0.000**

## 6.3 Testing the assumptions

Below I test, where possible, which SP assumptions hold and which do not. Some of these tests require the predictions for a certain world, regardless of whether the respondent believes the world is actual or counterfactual. To that end, I create $P_{cheap}$ and $P_{exp}$, the predictions for 'what if the artwork is cheap' and 'what if the artwork is expensive', respectively. $P_{cheap}$ is equal to Pr if the respondent believes the artwork is cheap, and equal to Pw if they believe the artwork is expensive.

| sign | obs | sum ranks | expected |
|---|---|---|---|
| positive | 199 | 64335 | 49297.5 |
| negative | 116 | 34260 | 49297.5 |
| zero | 155 | 12090 | 12090 |
| all | 470 | 110685 | 110685 |

**Table 9 Wilcoxon signed-rank test comparing $P_{cheap}$ and $P_{exp}$. p=0.0000**

### 6.3.1 Mcit (Assumption 1)

Since the only observable distributions are from actual worlds, we cannot test whether mcit truly holds. We can, however, test whether respondents believe in mcit. If they do, they will predict higher frequencies for answers they believe are true, on

| | 1 (exp) | 2 (cheap) | 3 (exp) | 4 (exp) | 5 (cheap) |
|---|---|---|---|---|---|
| **Actual** | **48.9%** | **34.0%** | **68.1%** | **54.3%** | **42.6%** |
| 2b | 49.6% | 49.0% | 57.1% | 50.9% | 49.7% |
| 2c | 44.4% | 58.1% | 52.3% | 43.3% | 61.8% |

**Table 10 Testing assumptions 2b and 2c. If respondents' predictions for all worlds are accurate on average, row '2b' should be similar to row 'actual'. If respondents' predictions are accurate for the world they believe is actual, row '2c' should be similar to row 'actual'. When the average prediction is significantly different from the actual value, the cell is marked orange.**

[11] There is also no significant difference between the distributions of P in the different treatments. Although this is not in line with expectations when looking at the theory, it is expected when looking at the results of the regressions and the other non-parametric tests.

average. I use a Mann-Whitney U-test to test if this is the case, shown in Table 8.

Respondents who think that an artwork is cheap, give significantly higher P's than respondents who believe that the artwork is expensive, and vice versa (p=0.000). Recall that, as always, all predictions are for the percentage of other respondents answering 'cheap'. This shows that people tend to think that more people will think something is true, when they themselves think it is true.[12] Running the same test for the individual artworks results in p=0.000 five out of five times, showing that the effect is present in every artwork.

Similarly, whether respondents believe in mcit can be tested within subjects, looking at the scenario predictions instead of the main predictions, by testing whether the newly created $P_{cheap}$ is larger than $P_{exp}$. I use a Wilcoxon signed-rank test, as shown in Table 9. The p-value of 0.0000 indicates that, regardless of their own beliefs, respondents think, on average, that more people will answer 'cheap' if the artwork is indeed cheap. Running the same test for the individual artworks yields significant results on at least the 5% level for all artworks, with the exception of artwork 2. For this artwork, $P_{cheap}$ is still larger than $P_{exp}$ on average, but the difference is not significant (p=0.482).

Both of these results provide strong evidence that respondents, on average, think that 'an answer will be more common if it is true' (mcit).

### 6.3.2 Accurate predictions (Assumption 2)

#### *6.3.2.1 Assumption 2a*

Let us start with the obvious: Assumption 2a does not hold, respondents do not know the distribution of answers in every world. If they did, all Pr's for a given artwork would be the same, as would all Pw's. In reality, both show a wide variety.

#### *6.3.2.2 Assumption 2b*

Keeping in mind the distribution of respondents over the different thinker types, with at least 20% of respondents in all types (see Figure 13), it seems somewhat unlikely that assumption 2b holds. With many respondents predicting Pr=Pw, and many giving prediction pairs where Pr and Pw are quite different, how likely is it that the average predictions for all worlds will be accurate? Still, it is theoretically possible. The problem, once again, is that only the distributions in the actual worlds are known. The best we can do is look at the available distributions, in the five actual worlds, and see if the average predictions are accurate. For these average predictions, we take into account the Pr's from those who thought the world was actual and the Pw's from those who thought the world was counterfactual. In other words, if the artwork in question is cheap, we look if $P_{cheap}$ is accurate on average. If the artwork is expensive, we instead look at $P_{exp}$. Table 10 shows the actual frequencies and the corresponding average predictions (row 2b).

[12] This could be subconsciously, respondents may not realise that their predictions follow mcit. Whether consciously or not, respondents make higher predictions when they believe the answer is true.

I use a Wilcoxon signed-rank test to test if the average predictions are significantly different from the actual values. When the hypothesis that the population mean is equal to the actual value is rejected[13], the cell is marked orange.

For three out of five artworks, the hypothesis is rejected. I conclude, therefore, that assumption 2b does not hold: respondents do not make accurate predictions, on average, of the distributions of answers in every world.

Ideally, we would determine the range, for each artwork individually, within which the average prediction is 'accurate enough' for the SP method to work, and test if the average predictions per group fall within that range. How wide such a range would be would depend on several factors, however, an important one being the (unobservable) distribution of answers in the counterfactual world. Determining these ranges falls outside of the scope of this thesis. Instead of a range, therefore, I only test if the average predictions are significantly different from the actual values themselves.

#### *6.3.2.3 Assumption 2c*

According to assumption 2c, respondents make accurate predictions, on average, of the distribution of answers in the world they believe is actual. The average predictions should now be calculated using only the Pr's of respondents who thought that the actual world was, indeed, actual.[14] Table 10 (row 2c) shows these average predictions.

In the same way as before, I find that four out of five average predictions are significantly different from the actual distributions, leading me to conclude that assumption 2c does not hold: respondents do not make accurate predictions, on average, of the distribution of answers in the world they believe is actual.

#### *6.3.2.4 Assumption 2d*

According to assumption 2d, respondents make accurate predictions, on average, of the distributions of answers in the worlds on which they base their predictions. Since we assume that all respondents base their predictions at least on the world they believe is actual, assumption 2d requires that assumption 2c holds. As such, I conclude that assumption 2d does not hold either.

[13] At the 5% level.

[14] In this case, that is the same as looking at the Pr's of respondents who chose the correct answer. If the distributions in the counterfactual worlds could be observed, we would, for those worlds, use only the Pr's of respondents who thought that the counterfactual world was actual.

## 6.4 Performance

### 6.4.1 Answers

Table 11 shows the performance of majority voting and SP, overall and per treatment. All in all, depending on how we deal with ties, SP performs slightly better, but there were not enough questions to test if it is a significant improvement. Overall, SP only resulted in the correct answer one out of five times. Looking at the previous tests, this is the result of assumption 2 not holding.

| | 1 (exp) | 2 (cheap) | 3 (exp) | 4 (exp) | 5 (cheap) |
|---|---|---|---|---|---|
| Majority | | | | | |
| SP | | | | | |
| Majority1 | | | | | |
| SP1 | | | | | |
| Majority2 | | | | | |
| SP2 | | | | | |
| Majority3 | Tie | | | Tie | Tie |
| SP3 | | | | | |

**Table 11 Performance of majority voting and SP, overall and per treatment. Correct. False.**

### 6.4.2 Predictions

To conclude, let us have a look at the prediction accuracy of different groups of respondents, starting with the different thinker types. Table 12 shows the actual distribution of answers per artwork, along with the average prediction for each thinker type. The average predictions here are based on the main predictions. I again use the Wilcoxon signed-rank test. The difference between average prediction and actual distribution is significant most often for the overall average predictions, namely four out of five times, while for non-world and 1-world thinkers it is significant only once.

| | 1 (exp) | 2 (cheap) | 3 (exp) | 4 (exp) | 5 (cheap) |
|---|---|---|---|---|---|
| **Actual (94)** | **48.9%** | **34.0%** | **68.1%** | **54.3%** | **42.6%** |
| Overall (94) | 53.5% | 44.6% | 61.3% | 52.7% | 52.3% |
| Non-world (24) | 54.3% | 45.6% | 70.4% | 46.4% | 49.4% |
| 1-world (17) | 60.5% | 36.5% | 62.1% | 53.1% | 49.0% |
| 2-world (16) | 55.6% | 46.5% | 55.9% | 55.9% | 60.5% |

**Table 12 Comparison of the prediction accuracy of thinker types. When the average prediction is significantly different from the actual value, the cell is marked orange.**

Table 13 shows the same for respondents who were correct and those who were not, regardless of thinker type. The bottom row, experts, are the respondents who got at least four out of five artworks correct.

| | 1 (exp) | 2 (cheap) | 3 (exp) | 4 (exp) | 5 (cheap) |
|---|---|---|---|---|---|
| **Actual (94)** | **48.9%** | **34.0%** | **68.1%** | **54.3%** | **42.6%** |
| Correct | 43.2% | 57.3% | 47.3% | 44.9% | 64.0% |
| Not correct | 64.3% | 38.0% | 67.9% | 59.4% | 43.6% |
| Experts (11) | 44.3% | 55.2% | 49.1% | 50.5% | 47.5% |

**Table 13 Comparison of the prediction accuracy of respondents who were correct and those who were not, per artwork. When the average prediction is significantly different from the actual value, the cell is marked orange.**

Following the LST rationale, 'correct' should outperform 'not correct', and both should be outperformed by the experts. Neither of these are the case here. In the case of the experts' relatively bad performance, this is probably because they miss out on a lot of 'crowd wisdom', as there are only 11 of them. The same thing could explain that the 'not correct' group outperforms the 'correct' group, prediction wise: for artworks 2 and 3, the 'not correct' group is twice as large as the 'correct' group. For artworks 4 and five they are also in the majority. Artwork 1, which is the only work were 'correct' outperforms 'not correct', is the only work where this is not the case.

# 7 Discussion

Most things worth discussing, especially the results, have already been discussed in their respective chapters. In this chapter, I elaborate on some of them and discuss more general issues.

## 7.1 Time inconsistencies

When designing the experiment, I made a conscious decision to split the survey into two stages. The purpose to this was to avoid, as much as possible, order effects in the form of 'fake' rationality. As explained in the experimental design, respondents might adjust their Pr and Pw to be consistent with their P, so that they seem rational. The separation into stages diminishes or solves this problem, because most respondents do not remember their answers from stage 1 once they are in stage 2. Not having the separation would also greatly diminish the difference between treatments. The questions that are supposed to steer respondents towards 2-world thinking (confidence in treatment 2, and Pr+Pw in treatment 3) would then be posed to respondents in treatment 1 already after their first prediction, instead of after they complete stage 1. These steering questions could then influence respondents in treatment 1 for four out of 5 predictions, while this treatment is supposed to be the 'pure' SP method without other influences. With the separation into two stages, this was avoided.

Unfortunately, the separation had an unforeseen result. There were indeed no order effects, but, as explained in the analysis, the predictions seemed to be inconsistent over time. For whatever reason, (deduced) predictions by the same respondent for the same artwork were often different in both stages. I named some other possible reasons in the analysis, but most important are the difficulty of the questions and the low incentives.

Judging art is difficult and subjective enough that it is quite easy to give the same artwork different verdicts on separate occasions. If the questions had been more straightforward, like knowledge based questions, or if the topic had been more accessible, like student life, there would probably be considerably fewer inconsistencies in the predictions.

Regarding the incentives, it is likely that, for most respondents, dominance[15] was not satisfied. Participants could win €25.- by being the best predictor. It is not unthinkable that for a large part of respondents, the main incentive to partake in the experiment was to help out a fellow student. The small chance of winning the money was possibly not worth putting in a lot of effort. This is true for all three channels. If, on the other hand, the incentives had been (a lot) higher, they might have drowned out the subjective costs of putting in effort and the intrinsic motivation for participating. Putting in more effort would diminish the inconsistencies, but even with high incentives the task may have simply been too difficult.

## 7.2 Limitations

Besides fixing the already mentioned problems, there are two more things I would do differently if I were to redo this experiment. The first is that I would make sure that all respondents have a (rough) idea of what the sample looks like. This information is important if one wants to make accurate predictions, and I did not inform the respondents sufficiently. The second is that I would ask some demographic questions. I thought I could overcome the need for them by limiting the scope of respondents to current students and recent graduates.

[15] In the sense of Smith's (1982) precepts for an economic experiment.

However, I find myself wondering whether a respondent's highest completed level of education, and, most of all, their field of study, have an influence on the type of thinking they use.

Another limitation of this research concerns the testing, and even the formulation, of assumption 2. I tested whether the average predictions were significantly different from the actual distributions. In most cases they were indeed significantly different, so I concluded that assumption 2 did not hold. To then be able to conclude that the SP method would fail because of this violation, assumption 2 would have to be phrased with the wording of average predictions being 'accurate enough', instead of 'accurate'.

Recall the Philadelphia example from section 2.1.1, where in the actual world, 60% of people erroneously believe that Philadelphia is the capital of Pennsylvania. In the counterfactual world, 90% believe this. The average prediction for answer 'yes' is 0.4*60% + 0.6*90% = 78%.

Since the actual frequency is 60%, answer 'yes' is less common than predicted, and answer 'no' is surprisingly popular. Now, let us see what happens if, for example, predictions for the counterfactual world are not accurate. If, instead of 90%, respondents predict 80% for the counterfactual world, this difference is likely statistically significant. The average prediction, however, which is then 72%, still results in the same answer. As long as the average prediction stays above 60%, answer 'no' will be surprisingly popular.

A violation of assumption 2 does therefore not necessarily result in a failure of the SP method. The average predictions only have to be 'accurate enough'. Since, in practice, the distributions of answers in counterfactual worlds are unobservable, determining rules or guidelines to judge whether predictions are accurate enough poses quite a challenge. This challenge fell outside of the scope of this thesis, so for the sake of simplicity I assumed that the average predictions had to be exactly accurate, but it is a possible direction for further research.

## 7.3 Results

For the most part, I did not find the expected results. I found no evidence for either hypothesis and, despite the SP method's success in previous research, assumption 2 does not seem to hold. This could be because the hypotheses are false and the assumption really does not hold, or it could be because of the problems mentioned above. It would be interesting to run the same experiment again, taking all the discussed problems into account.

The mentioned problems probably influenced the distribution of respondents over the different thinker types as well, so that the approximately equal distribution over the three types may not be accurate in general. However, I am confident that there will almost always be considerable numbers of both 1-world and 2-world thinkers, resulting in the need for assumption 2d to hold. The actual distribution is likely to depend on some other factors as well. I discuss them briefly.

## 7.4 Other factors of influence

There are a couple of other things that are likely to influence whether respondents base their predictions on different worlds, which were not relevant in this experiment, but which will likely be relevant in others. The first is the number of possible worlds (the number of possible answers to an MC question). Second is the difficulty of the questions, specifically the scenario where the respondent does not know the answer, but knows that a lot of other respondents will. Lastly, I discuss the influence of risk aversion and the incentive scheme.

### 7.4.1 Number of questions

If a question has two possible answers, it seems reasonable that a respondent imagines both worlds, makes predictions for both worlds, assigns probabilities to both worlds and computes their prediction. On the other hand, imagine how this would work if there were five answer possibilities. Respondents would have to imagine five different worlds, all with different predictions, and use all of them to compute their final prediction. Note that the prediction for one world, in that case, consists of the distribution of answers over all five answer possibilities. Respondents that want to take all worlds into account would have to make five different distributions, consisting of five frequencies each, and assign probabilities to each distribution.

In cases like these, 1-world thinking might become a lot more popular. It could also be that respondents only take into account the worlds that are (somewhat) likely to be true and the worlds that have a distribution different enough to have an impact. Worlds that are deemed implausible might be ignored, while worlds with similar distributions might be grouped together, similar to the editing phase in Prospect Theory (Kahneman & Tversky, 1979). Another possibility is that respondents make some intuitive adjustments from their 1-world predictions. Which of these methods respondents use, if any, would be a nice direction for further research.

### 7.4.2 Obviously different distributions

If a respondent does not know the answer to a question, but he knows or suspects that a large part of the other respondents will know the answer, this should increase the probability that he imagines two different worlds. As an extreme example, imagine that the question is whether a certain candidate won the presidential election in the respondent's country. The results came in a few hours before the experiment, but the respondent has not heard them yet. If all respondents are from the same country, the respondent is likely to realise that the distribution of answers will be drastically different in both possible worlds. If the candidate won, respondents who heard the results will answer 'yes', if he did not, those same respondents will answer 'no'.

In general, if it is more obvious that the distributions in both worlds are different, it should be more likely that a respondent uses 2-world thinking. It would be intersting to test whether this is true in practice.

### 7.4.3 Risk aversion

Since 2-world thinking, as opposed to 1-world thinking, is basically hedging, more risk seeking respondents are more likely to use 1-world thinking and vice versa. After all, with 1-world thinking, assuming accurate predictions, the prediction is either right or wrong. 2-world thinking results in a prediction that is always a bit off. This effect will be greatest if the incentive scheme is such that only the best predictor wins a reward. Of course, for a respondent to consciously make this decision, they first have to realise that both types of thinking are possible.

# 8 Conclusion

I set out to answer the questions of whether respondents in the SP method actually base their predictions on different worlds, and whether the researcher can influence this by asking them for their confidence.

Regarding the first question, the goals were to find out which assumptions really need to hold to guarantee that the method works, and to get a deeper understanding of how it works. I managed to answer the question with enough accuracy to achieve these goals: respondents were approximately equally divided over the three thinker types.

The consequence of having large amounts of both 1-world and 2-world thinkers is that, regarding the second assumption of SP, assumption 2c is not strong enough. Assumption 2b, on the other hand, does not need to hold. The weakest version of the assumption that needs to hold to guarantee that the method works is 2d (with the understanding that a considerable part of respondents will use 2-world thinking):

*On average, respondents make accurate predictions of the distribution of answers in the worlds on which they base their prediction.*

In this experiment, the average predictions were too different from the actual distributions, causing the SP method to fail.

Regarding the second question, I wanted to take a small step towards the goal of discovering how to steer as many respondents as possible towards the same type of thinking. The method can then be used as accurately and efficiently as possible.

I did not find any evidence for my hypotheses regarding the influence of (asking for) confidence. Confidence (estimated probability of being correct) did not have a negative effect on the probability of the prediction being based on two worlds, nor did asking a respondent for their confidence in advance have a positive effect on this probability.

There were several issues with the experiment, however, like the difficulty of the tasks and the relatively low incentives, that resulted in inconsistent and irrational answers. Therefore, even though I did not find any evidence in their favour, the hypotheses could be true in general. Likewise, it is possible that assumption 2d does hold in general, even though, based on this experiment, I concluded that it does not. Looking at the success of the SP method in other experiments, it is likely that it holds at least to some degree. It would be worthwhile to repeat this experiment, taking into account all the issues from this research.

# 9 References


Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263. https://doi.org/10.2307/1914185

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science (New York, N.Y.)*, *306*(5695), 462–466. https://doi.org/10.1126/science.1102081

Prelec, D., Seung, H. S., & McCoy, J. (2014). Finding truth even if the crowd is wrong. *Working Paper*. Retrieved from http://seunglab.org/wp-content/uploads/2015/07/FindingTruth16-copy.pdf

Prelec, D., Seung, H. S., & McCoy, J. (2017a). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532–535. https://doi.org/10.1038/nature21054

Prelec, D., Seung, H. S., & McCoy, J. (2017b). A solution to the single-question crowd wisdom problem: Supplementary Information. *Nature*, *541*(7638), 532–535. https://doi.org/10.1038/nature21054

Smith, V. L. (1982). Microeconomic Systems as an Experimental Science. *American Economic Review*, *72*(5), 923–955. https://doi.org/10.2307/1812014

Weiss, R. (2009). *Optimally Aggregating Elicited expertise: a proposed application of the Bayesian Truth Serum for policy analysis*. Massachusetts Institute of Technology. Retrieved from http://dspace.mit.edu/handle/1721.1/53066%5Cnpapers2://publication/uuid/B272DB91-4482-438F-92EF-A50D20B24EB0

# Appendix A Survey

## Treatment 1

Thank you for participating in this experiment, I really appreciate your help!

This experiment, in which **you can win €25.-**, consists of two stages. In stage 1, I will show you five pieces of art. For each artwork I will ask you one or more questions. In stage 2, I'll ask you some different questions about the same five artworks. Don't worry, you'll get to see them again. No need to remember them.

Some questions will be about the market price of the artworks. Besides that, I am interested in how well people can predict the responses of other people. Some questions will therefore ask you to guess how other people will respond.

The best predictor (the person whose predictions are the closest to the truth, on average) wins €25.-.

**Introduction stage 1**

Welcome to Stage 1.

For every artwork you see, please answer whether you think the market price is higher or lower than **€30,000.-**.

You are also asked to predict the percentage of other respondents that answered 'less than €30,000', and the percentage of other respondents that answered 'more than €30,000'. Note that these percentages have to add up to 100%.

**Questions stage 1 (per artwork)**

Do you think the market price of the artwork above is less or more than €30,000?

What percentage of respondents do you think answered 'less than €30,000' for this artwork? And 'more than €30,000'?

**Introduction stage 2**

Welcome to Stage 2.

For every artwork from Stage 1, you'll see the answer you gave about its market price. Please indicate how certain you are of this answer, by giving your estimated probability of being correct. Note that this is at least 50%, because even if you have no clue and you pick an option at random, you still have a 50% chance of being correct.

You will then be asked to consider the case that the artwork is worth more than €30,000. Now, with the knowledge in mind that the artwork is actually worth more than €30,000, please make your prediction again. Predict the percentage of other respondents that answered 'less/more than €30,000' to the first question. Note that these percentages have to add up to 100%.

Next, consider the case that the artwork is worth less than €30,000. Again, with this knowledge in mind, predict the percentage of respondents that gave each answer.

**Questions stage 2 (per artwork)**

In stage 1, for the artwork above, you answered: **x**

What is your estimated probability of being correct?

Consider the case in which the artwork is indeed worth **x**. In that case, what percentage of respondents do you think would answer 'less than €30,000' for this artwork? And 'more than €30,000'?

Consider the case in which the artwork is actually worth **x**. In that case, what percentage of respondents do you think would answer 'less than €30,000' for this artwork? And 'more than €30,000'?

**Email**

[Optional] Please fill in your email address so that I can contact you if you are the winner of the €25. If you do not fill in your email address, you cannot win. Email addresses are not used for anything other than contacting the winner.

## Treatment 2

Thank you for participating in this experiment, I really appreciate your help!

This experiment, in which **you can win €25.-**, consists of two stages. In stage 1, I will show you five pieces of art. For each artwork I will ask you one or more questions. In stage 2, I'll ask you some different questions about the same five artworks. Don't worry, you'll get to see them again. No need to remember them.

Some questions will be about the market price of the artworks. Besides that, I am interested in how well people can predict the responses of other people. Some questions will therefore ask you to guess how other people will respond.

The best predictor (the person whose predictions are the closest to the truth, on average) wins €25.-.

**Introduction stage 1**

Welcome to Stage 1.

For every artwork you see, please answer whether you think the market price is higher or lower than **€30,000.-**. Next, please indicate how certain you are, by giving your estimated probability of being correct. Note that this is at least 50%, because even if you have no clue and you pick an option at random, you still have a 50% chance of being correct.

You are also asked to predict the percentage of other respondents that answered 'less than €30,000', and the percentage of other respondents that answered 'more than €30,000'. Note that these percentages have to add up to 100%.

**Questions stage 1 (per artwork)**

Do you think the market price of the artwork above is less or more than €30,000?

What is your estimated probability of being correct?

What percentage of respondents do you think answered 'less than €30,000' for this artwork? And 'more than €30,000'?

**Introduction stage 2**

Welcome to Stage 2.

For every artwork from Stage 1, you'll see the answers you gave about its market price and your estimated probability of being correct.

You will then be asked to consider the case that the artwork is worth more than €30,000. Now, with the knowledge in mind that the artwork is actually worth more than €30,000, please make your prediction again. Predict the percentage of other respondents that answered 'less/more than €30,000' to the first question. Note that these percentages have to add up to 100%.

Next, consider the case that the artwork is worth less than €30,000. Again, with this knowledge in mind, predict the percentage of respondents that gave each answer.

**Questions stage 2 (per artwork)**

In stage 1, for the artwork above, you answered: **x**

Your estimated probability of being correct was **x**%

Consider the case in which the artwork is indeed worth **x**. In that case, what percentage of respondents do you think would answer 'less than €30,000' for this artwork? And 'more than €30,000'?

Consider the case in which the artwork is actually worth **x**. In that case, what percentage of respondents do you think would answer 'less than €30,000' for this artwork? And 'more than €30,000'?

**Email**

[Optional] Please fill in your email address so that I can contact you if you are the winner of the €25. If you do not fill in your email address, you cannot win. Email addresses are not used for anything other than contacting the winner.

## Treatment 3

Thank you for participating in this experiment, I really appreciate your help!

This experiment, in which **you can win €25.-**, consists of two stages. In stage 1, I will show you five pieces of art. For each artwork I will ask you one or more questions. In stage 2, I'll ask you some different questions about the same five artworks. Don't worry, you'll get to see them again. No need to remember them.

Some questions will be about the market price of the artworks. Besides that, I am interested in how well people can predict the responses of other people. Some questions will therefore ask you to guess how other people will respond.

The best predictor (the person whose predictions are the closest to the truth, on average) wins €25.-.

**Introduction stage 1**

Welcome to Stage 1.

For every artwork you see, please answer whether you think the market price is higher or lower than **€30,000.-**. Next, please indicate how certain you are, by giving your estimated probability of being correct. Note that this is at least 50%, because even if you have no clue and you pick an option at random, you still have a 50% chance of being correct.

You will then be asked to consider the case that the artwork is worth more than €30,000. Now, with the knowledge in mind that the artwork is actually worth more than €30,000, please make a prediction. Predict the percentage of other respondents that answered 'less/more than €30,000' to the first question. Note that these percentages have to add up to 100%.

Next, consider the case that the artwork is worth less than €30,000. Again, with this knowledge in mind, predict the percentage of respondents that gave each answer.

**Questions stage 1 (per artwork)**

Do you think the market price of the artwork above is less or more than €30,000?

What is your estimated probability of being correct?

Consider the case in which the artwork is indeed worth **x**. In that case, what percentage of respondents do you think would answer 'less than €30,000' for this artwork? And 'more than €30,000'?

Consider the case in which the artwork is actually worth **x**. In that case, what percentage of respondents do you think would answer 'less than €30,000' for this artwork? And 'more than €30,000'?

**Introduction stage 2**

Welcome to Stage 2

For every artwork from Stage 1, you'll see the answer you gave about its market price.

This time you are not asked to consider a certain scenario. Please predict the percentage of other respondents that actually answered 'less than €30,000', and the percentage of other respondents that answered 'more than €30,000'. Note that these percentages have to add up to 100%.

**Questions stage 2 (per artwork)**

In stage 1, for the artwork above, you answered: **x**

What percentage of respondents do you think answered 'less than €30,000' for this artwork? And 'more than €30,000'?

**Email**

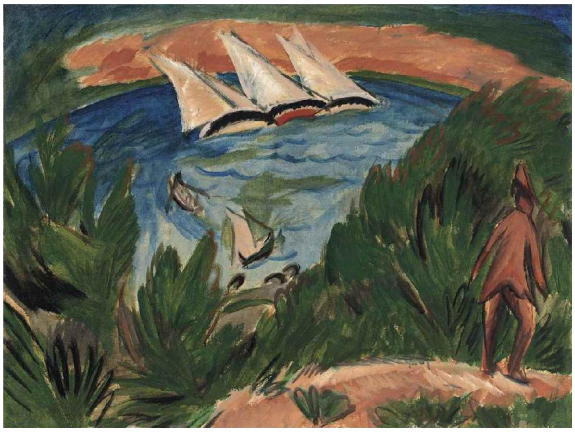
[Optional] Please fill in your email address so that I can contact you if you are the winner of the €25. If you do not fill in your email address, you cannot win. Email addresses are not used for anything other than contacting the winner.

# Appendix B Artworks

Artwork 1 Kandinsky -Studie zu improvisation 3 (millions)

Artwork 2 Gottlien - Still life with fish (€9000)

Artwork 3 Hockney - The Splash (millions)



Artwork 4 Kirchner - Segelboote im sturm (millions)

Artwork 5 Lichtenstein - Brushstrokes (€10,000)

## Appendix C Thinking type tree