UNIVERSITY OF GOTHENBURG
DEPARTMENT OF PSYCHOLOGY

Estimating the informative: Anchoring in estimations of observed ratios

Björn Alfons Edmar

Individual paper 15 credits
Bachelor´s thesis in psychology
PX1500
Spring semester 2021

Supervisor: Martin Hedesström

Estimating the informative:
Anchoring in estimations of observed ratios

Björn Alfons Edmar

Abstract. The aim of this study was to examine if anchoring effects occur in ratio judgements where the object of the estimation is visually examined by the estimator. In addition to this, impact of personal experience concerning the source of the anchor along with any potential reduction in anchor susceptibility due to abstract reasoning capabilities was also examined. This was done through an experimental survey on 159 participants recruited via Amazon Mechanical Turk. No reliable results were found. Primacy effects due to experimental design and issues in sampling are discussed.

Decisions come in many different shapes and sizes, but they all revolve around choosing one alternative in favour of others. Some choices have few alternatives, such as if you should ask for a particularly charming person´s phone number and others have many alternatives, such as selecting a listing price for an apartment. How and why we choose the alternatives we do have been a subject of study for a long time, but considering that the way we reach our decisions can have serious ethical and moral implications, this attention is merited. Consider for example the study by Englich et al (2006) where judges were shown to be affected by dice rolling in how long their sentencing were or Goldin & Rouse (2000), where they found that blinding recruiters view of auditionees appearance increased the likelihood of females being selected to join symphonic orchestras. Hopefully we can all agree that these factors: gender and dice rolls, shouldn't take part in our evaluation of alternatives (unless we´re playing Yahtzee or choosing a partner), but in order to make our decision processes better, we need to understand how they work.

Our decisions are generally viewed as a product of our judgements regarding a certain situation, in other words, how you reach a decision or choose an alternative will be dependent on how you view and judge your alternatives. The most prominent theory regarding how judgements are made is the dual process theory which postulates that there are two types, systems or patterns of thought, that dictate how we make judgements and subsequent decisions. The first type, commonly known as "system 1" is used when reaching decisions or making judgements based on heuristics, reflexes, fast thinking and intuition. The second type, commonly known as "system 2" is used when reaching decisions or making judgements based on reason, reflection and more analytic assessments (Evans, 2008). But as Evans (2008) suggests, this description of two systems that have different attributes might be misleading since many attributes categorized in these two systems are not actually related, and we might be better of classifying type 1 and type 2 thinking as thought processes that either tax a capacity-limited central working memory resource or not. Any further mention of type 2 or type 1 thinking will follow this definition, where type 2 thinking requires access to working memory resources and type 1 does not.

When we utilize our working-memory, or more specifically the central executive part of our working memory, we direct attention towards certain objectives that require

conscious thought (Baddeley, 1996, 2012). A simple but illustrative example of this can be made through mathematics (Kahneman, 2011). Calculating equations such as 2+2 requires no actual calculation or conscious thought for most adults, we simply know that it is 4. We do not need to focus on the numbers individually and add them together to get the result, further, the calculation does not require any cognitive exertion, that is, you are not accessing a limited cognitive resource when you consider the calculation. Contrast this previous calculation with the following: 352.3+18.75. Intuitively we recognize that the equation is not a difficult one to solve, but the answer is not accessible to us without calculation, that is, without access to limited working-memory resources. Thus, these types of calculations are taxing, even though they might not be hard, and doing many of them for a longer period of time will leave you feeling cognitively depleted. This is also an example of how type 1 thinking (2+2) differs from type 2 thinking (352.3+18.75).

Most biases such as availability, representatives, (Tversky & Kahneman, 1973) and affect heuristics (Slovic et al., 2007) are a product of type 1 thinking, that is, coming to a conclusion by not accessing and taxing precious working-memory resources. But some, such as anchoring and adjustment, the bias that made the judges in Englich et al (2006) give longer sentences to people when they rolled higher dice rolls, is not, in the sense that it requires access to working-memory resources in order to evaluate information and make estimations.

Anchoring and adjustment

Anchoring and adjustment, the phenomena of biasing estimations closer to a previously given numerical value, has been given much attention since it was first described in Kahneman and Tversky´s seminal 1974 paper (Tversky & Kahneman, 1974). The replications and examination of this now traditional anchoring effect often look quite similar (Furnham & Boo, 2010). You present a question, such as "is the Burj Khalifa (the world's tallest building) higher or lower than X meters?". After that comparative judgement has been made, the respondents are asked to estimate the absolute height of the building, that is, answering the question: "How tall is the Burj Khalifa?". The value presented along with the initial question, be it low or high, will then skew the respondent´s guess of how tall the building is in the direction of that number. For example, if I presented the number 12 and asked you if the Burj Khalifa was higher or lower than the value (12 meters) before I asked you how tall the building was, chances are that you would guess significantly lower than if I presented you with the number 3582. In this particular example we are guessing a height, which, if we need to guess, we don't know beforehand. If we already knew the height, the anchor shouldn't affect our guess.

Even though almost 50 years of research has been dedicated to examining the effects and the limitations of anchoring, some aspects of the phenomenon are still unknown (Furnham & Boo, 2010). One of the limits of anchoring is illustrated above, if we have sufficient knowledge about a subject we won't be as susceptible to anchor values (Smith et al., 2013), but this is not the only factor that dictates the effectiveness of anchors. There are many factors, but the main one which I will discuss in this paper is anchor dimension and relevancy (Chapman & Johnson, 2002). Anchor dimension refers to on which scale the anchor is presented. If anchors are presented in a sufficiently different dimension than the following estimation, an anchoring effect will be hard to detect. For example, if I wish to examine an anchoring effect in how people judge the

weight of an object, the anchor should be expressed in the same unit as the estimation, id est, an anchor expressed in pounds won´t be as effective on an estimation of kilograms and an anchor expressed in height won't be as effective on an estimation of length (Strack & Mussweiler, 1997). These two examples are still somewhat closely linked in the sense that they both express weight and distance so an anchoring effect might still occur, but the further the anchor dimension and relevancy deviates from the estimation, the weaker the anchor becomes. Further, it is also of note that anchoring is a more or less strictly numeric effect, that is, the anchor needs to be expressed as a number, not something that just symbolize a value. For example, Mochon & Frederick (2013) found that an object that corresponds to an amount of money doesn't work as an anchor in the same way the actual amount of money does. That is, anchoring with the value of a pack of AAA batteries does not garner the same anchoring effect as anchoring with 6 dollars.

There are differing, although not contradictory, theoretical accounts of how anchoring and adjustment work (Mussweiler & Strack, 1999; Chapman & Johnson, 2002; Frederick & Mochon, 2012; Wegener et al., 2010). The initial definition of anchoring was as a heuristic, but that definition has been revoked by Kahneman himself since it does not reflect the process of attribute substitution (Kahneman & Frederick, 2002), that is, substituting a hard, cognitively taxing type 2 question with a simple, untaxing type 1 question. In this study I approach anchoring as a type 2 bias in which a judgement is skewed due to insufficient adjustment away from a biased starting point. You can also approach anchoring as a type 1 bias in which a judgement is skewed based of contaminative numerical information received prior to the judgement, but not necessarily consciously related the estimation (Chapman & Johnson, 2002) . The example with the Burj Khalifa is an example of the former of these perspectives, the biased estimation occurs due to the evaluation of the number 12 as a possible answer to the question and thus serves as an anchor for that estimation. If we had for example just asked the participants to think of the number 12, and then asked them to guess the height of the Burj Khalifa, an adjustment from the given number would not necessarily have occurred. There could however still exist an anchoring effect, but in this instance that effect would have been semantic, associative or possibly subliminal. This illustrates the latter approach to anchoring as described above, that is, anchoring as a numeric primacy effect (Mussweiler & Englich, 2005; Mussweiler & Strack, 2001; Reitsma-van Rooijen & L. Daamen, 2006; Wilson et al, 1996) . The literature on the anchoring effects that deviate from the traditional Kahneman-Tversky paradigm in the sense that the adjustment phase of the evaluation is skipped is not without its faults. Replicating subliminal anchoring effects is hard (Röseler et al., 2021) and replicating the much cited study by Wilson et al (1996) has also proved difficult (Brewer & Chapman, 2002).

We do however have convincing evidence that anchoring effects from the traditional paradigm regarding general knowledge estimations, such as "Height of the tallest redwood tree", "Year the telephone was invented", "length of the Mississippi river", et cetera (Jacowitz & Kahneman, 1995) are robust and replicable (Klein, 2014). This incredibly rich empirical evidence of anchoring leaves us in a position open to much exploration without having to make an unflinching commitment to any one theoretical explanation, especially since the occurrence or absence of the anchoring effect is what is of main concern. A more detailed account of the different types of anchoring processes, the conditions under which they occur and why they occur is beyond the scope of this paper (see: Chapman & Johnson, 2002; Furnham & Boo, 2010; Simmons et al., 2010 for more information).

In conclusion, anchoring and adjustment is one of the most robust and well researched phenomena ever found in the judgment and decision making subfield of psychology. However, as we have seen, much research is focused on estimations that have the possibility of being very skewed. Either in the sense that they are done on a practically limitless scale such as weight, or that they concern estimations in which the estimator lacks the knowledge required to make a reliable and accurate estimation. Estimations in these circumstances are quite lax in the sense that there is a potential of great variation within the estimations. The main purpose of the present study is to broaden research in anchoring and adjustment to harsher settings, that is, settings where the naturally occurring variance is low. One of the issues concerning anchoring lie in appraising the frequency of anchoring effects in real-world decision making, and by broadening experimental research to include increasingly stringent estimation settings we can get a better picture of how anchoring can occur in everyday decisions.

This study has three goals, the first is to examine whether or not anchoring effects occur in ratio judgements where the estimator has visually observed the ratio of interest. In our particular setting, the estimator is factually acquainted with the object of the estimation, and an accurate estimation should be made if the estimation process remains unbiased. A real-world example of such a task could be estimating the field goal shooting percentage of a basketball player during a game you just watched. Since you are factually acquainted with the shooting performance of the player you should be able to make an accurate estimation given that you have paid attention to the task. There is a difference between this type of estimation, id est ratio estimations, and estimates regarding general knowledge or likelihoods, in the sense that the estimator has access to all the necessary information to make an accurate estimate. In contrast, if you ask me to guess the height of the Burj Khalifa and I have no prior knowledge of the existence of such a building, my estimate will be a "true" guess rather than an informed one. Thus, this kind of performance evaluative task must be seen as a task where the estimator has a high degree of knowledge about the subject of their estimation. And since knowledge in a subject makes the estimator less prone to skewed estimations (Smith et al, 2013), we can assume that the variability in these estimations should be lower and therefore more resistant to potential anchoring effects. This study aims to further explore anchoring effects in subjects that has had access to the amount/value they are trying to estimate. One of the only examples (that I could find) of this kind of estimation in a research setting can be found in a study by Fath, Larrick, and Sol (under review) in which participants viewed a video of a person performing a task and were afterwards asked to estimate the person´s performance on that task in terms of percentages of correct answers. The results of their studies indicate that anchoring effects, although modest ones, are found in settings where the estimator bases their estimate on their own visual evaluation of a performance. Subsequently, the first goal of this study is to corroborate the results found by Fath, Larrick and Soll, but in a slightly different setting. This leads us to the first hypothesis of this study, *H1: Anchoring effects due to a low anchor can occur in ratio judgements where the estimator has had access to all necessary information to make an accurate estimation*. That is, I hypothesize that the treatment groups in this study will give lower estimations than the control group (see procedure for further information). If this were to be the case, it would imply that anchor values affect our evaluation of objects that solely rely on our visual inspection, rather than common knowledge or personal preferences.

The second goal of this study is to examine whether personal experience of the source of the anchor affects the susceptibility to the anchor when estimating a ratio. That

is, examining how the degree of familiarity the estimator has to the source of the anchor value affects their following estimation. I hypothesize that experience of the source of the anchor value can reduce the susceptibility towards biased judgements since the salience of information that relates to personal experience should be more accessible and therefore more easily recalled (Mussweiler & Strack, 2001; Wilson et al, 1996). However, the opposite could also be true. Information which is more easily recalled could also serve as a stronger anchor and thus bias the estimation more. It comes down to how we use the information, and since the anchoring and adjustment bias is, at least partly, a type 2 process (Kahneman, 2011), recall could reduce the susceptibility to anchoring when making ratio judgements. Our second hypothesis is thus, *H2: Personal experience of the source of the anchor will moderate anchor susceptibility*. Note that this hypothesis lacks direction, this is due to the fact that we are unable to know how this personal experience will be treated in regard to the subsequent estimation. That is, since no clear theoretical backing for a one-way hypothesis can be made, I decided to not make one. It should however be noted that experience of the source of the anchor could both increase and decrease anchor susceptibility, and thus cancel each other out. For a more detailed examination of how experience of the source of the anchor affects our susceptibility to that anchor, a more specific research design that cater to that question is required.

The third goal of this study is to examine if abstract reasoning capabilities affects the susceptibility to anchoring. The relationship between cognitive abilities such as intelligence, and numerical anchors have, according to Furnham & Boo (2010), been contradictory and equivocal. An example of this kind of ambiguous results can be found in Teovanović (2019) where no obvious relationship between intelligence and anchoring susceptibility could be found. However, a relationship between a high degree of the Big-five personality trait Openness (Costa & McCrae, 1999; DeYoung, 2015) and anchoring susceptibility was found. This can be viewed as puzzling considering the positive relationship generally found between Openness and intelligence (DeYoung et al, 2005), especially if you suspect intelligence to be a moderator of anchoring susceptibility (Bergman et al., 2010; Teovanović et al., 2015). It should however be noted that personality traits have an equally ambiguous relationship with anchoring susceptibility and further research on individual differences in anchoring susceptibility is needed (Norem, 2019). This leads us to the final hypothesis*, H3: Participants who score high in abstract reasoning will be less susceptible to anchoring.* That is, they will make more accurate estimations than participants who score low in abstract reasoning.

The two latter hypotheses aim to further research in bias reduction and is therefore dependent on the first hypothesis being true (if no bias occurs no reduction in that bias can be examined). Much research, as indicated by the first hypothesis *H1*, aim to discover or illustrate certain biases in certain settings but does little to examine how these biases can be reduced, which is arguably a more important issue. If the aim of descriptive judgement and decision making research is to lay the groundwork for normative decision making procedures, it is highly relevant to examine potentially moderating variables.

Method

Participants

The sample consisted of 159 participants from the United States recruited via Amazon Mechanical Turk (MTurk). The participants received 50 cent in compensation for their participation. The sample had a mean age of M= 30.5, SD = 7.78. Of these 159 participants 106 were male, 52 were female and one identified as non-binary. An a priori power calculation was conducted using the software G*Power. Since previous studies in this particular judgement setting was hard to come by a good effect size prediction was hard to make. In the end an effect size of d = .6 was expected from the sample based on relevant studies presented in a meta study by Townson (2019). Since the study follows a 1(ratio estimate) x 3 (treatment conditions) one-way factorial design the Cohen´s d point estimate was recalculated as a Cohen´s f following procedures from Cohen (1988) and resulted in an expected effect size of f = .3. For a power of .8 a sample size of n = 111 would be required, given an alpha value of α = .05 and an effect size value of f = .3, meaning that the probability of rejecting a false null hypothesis would be 80%. Thus, collecting a sample with a size of n = 159 was considered to be sufficient in order to answer the present research questions.

Instruments

The instrument used to gage abstract reasoning skills was an abridged version of the matrix reasoning item bank, or MaRs-IB, by Chierchia et al (2019). The MaRs-IB is an open source abstract reasoning test following a standard pattern matrix design. This abridged version, created particularly for this study, contained 20 items consisting of standard 3x3 image matrices with the last image missing. The objective of the respondent is to select one out of four image alternatives to complete the pattern. Participants were not required to answer to all the items, in fact, they only had 150 seconds (02:30 minutes) to complete as many of the items as possible (for an example of these pattern matrices see Appendix B or Chierchia et al (2019)).

The dependent variable was an estimation of how many correct answers an individual achieved on a math test. The estimation was done by viewing a table consisting of 8 rows and 5 columns representing 40 questions presented in 5 blocks (see Appendix A). The correct answers were coloured in green and incorrect answers in red. The ratio of incorrect to correct were 3:7, corresponding to 70% correctly answered questions. This estimation (of correct answers on the math test) serves to measure how accurate a ratio judgement is. Estimates closer to the true score of 70% correct will be judged as less biased and more accurate than estimates that deviate from the true score of 70% correct.

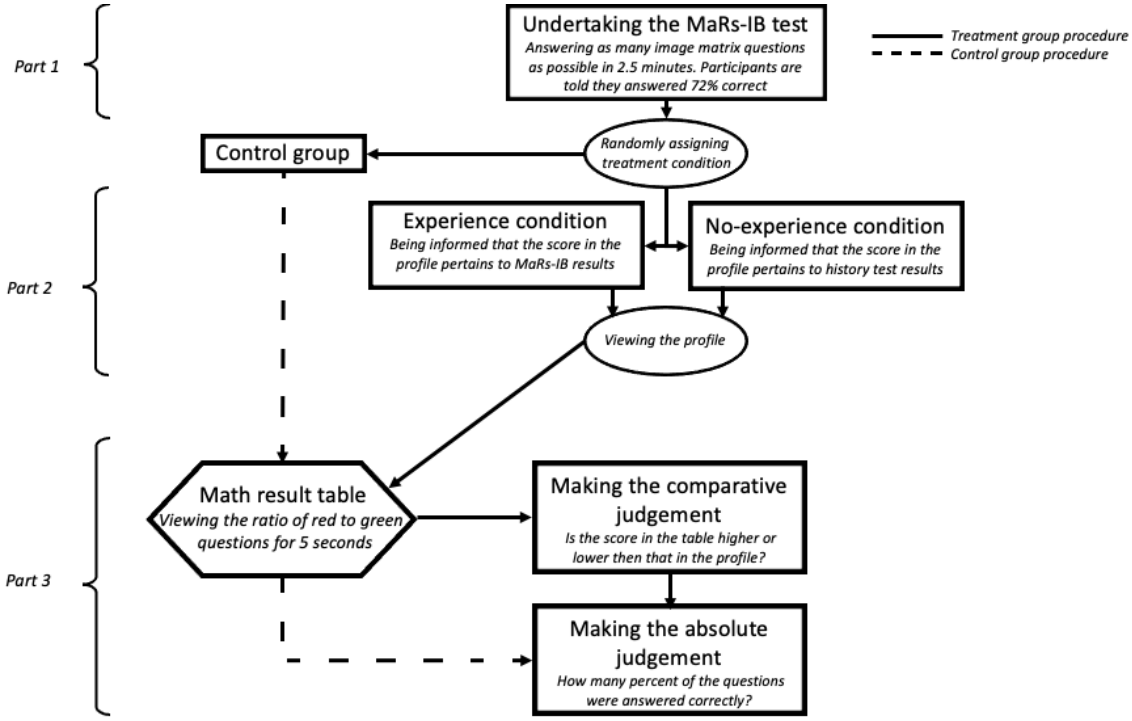
Procedure

This study used a survey with two experimental conditions and one control group in order to examine the hypothesises described above. The survey consisted of three parts,

I) abstract reasoning assessment, II) treatment group procedures, and III) ratio evaluation and estimation tasks. Figure 1 gives an overview of the experimental procedure.

Figure 1

Experimental procedure



Note. Figure 1 illustrates the procedure from start to finish. The dotted line represents the control group's task order and the black line represents the two treatment groups' task order. Note that in part 2 the two treatment groups did not view different profiles. They viewed the same profile but with different information regarding what test the score came from.

The first part of the survey aimed to assess the participants' abstract reasoning skills by asking them to complete an abridged version of the MaRs-IB test (see Instruments). After completion of the MaRs-IB, the participants received feedback on their performance on the test. All participants were told that they scored 72% correct. The believability of this feedback was judged to be high based on a pilot study consisting of 274 responses on a test similar to the MaRs-IB, where the participants where asked if they felt that their result of 72% correct was believable. Feedback on the MaRs-IB was given in order to contextualize information that would be received in the second part of the survey.

The second part of the study only involved the treatment groups, consequently the control group skipped the second part of the study, illustrated by the brackets in Figure 1. In the second part the participants were shown a fictitious profile of a previous participants which contained the participant's name, age, education and most importantly, their result on one out of two tests. The test result reflected the fictitious previous participants score on either the same test the respondents undertook in the first part of the

survey, the MaRs-IB, or a history test which the respondents had no previous knowledge about. The test result indicated that the previous participant had 30% correct answers. These two treatment groups are labelled as the "No-experience condition" and the "Experience condition", where the no-experience condition involved the history test and the experience condition involved the MaRs-IB test. As you can see in Figure 1 the actual profile shown to the participants was the same, only the previous information regarding the contents of the profile differed. Both the profile and the accompanying information messages can be viewed in Appendix A.

The third part of the study involved the target estimation and any potential anchoring effects. All groups viewed a table representing a person's result on a math test for 5 seconds. The table consisted of 40 questions, 28 questions were marked in green indicating a correctly answered question and 12 questions were marked in red indicating an incorrectly answered question, the ratio of red to green questions were thus 3:7 (for further description of the table see Instruments and Appendix A). The participants in the treatment groups were told that the math test results were those of the participant previously shown to them in the profile. After viewing the table representing the results on the math test, all groups were asked to estimate how many percent of the questions in the math test were answered correctly. Before making this estimation however, the treatment groups were asked to make a comparative judgement regarding the score shown to them in the profile and the score indicated by the table. This comparison concerned whether or not the score indicated by the table was higher or lower than the score showed to them in the profile; in effect asking if the score indicated by the table was higher or lower than 30%, but since the participants are not shown the profile while they are making this comparative estimation they would have to recall the test result shown to them in the profile correctly in order to answer the question confidently (An option of estimating equal performances on the two test was also available for exhaustive purposes). This way of inducing a potential anchoring effect follows the standard Kahneman/Tversky paradigm (Kahneman & Frederick, 2002), that is, by first asking for a comparative judgement framed by the anchoring value and following that question by asking for an absolute estimation. In our case, the anchor value used to induce any potential anchoring effects was the score shown to the treatment group participants in the profile of the previous participant, that is, a value of 30% correct.

Without any biasing information, the task of estimating the amount of correctly answered questions was judged to be easy due to the precision of the estimations in the pilot study, where the respondents estimated a mean of 71.6% correct answers (IQR = 68.5-74.3) when the true score was 70%, even when they only viewed the table for one second. Further, the most frequent guesses were of 70 and 80 percent correct, representing 40 out of 99 responses. Thus, the table was judged to be easily estimated.

The a priori statistical procedure I aimed to utilize in order to examine hypothesis *H1: Anchoring effects due to a low anchor can occur in ratio judgements where the estimator has had access to all necessary information to make an accurate estimation* and *H2: personal experience of the source of the anchor will moderate anchor susceptibility,* was a one-way analysis of variance (ANOVA) examining the mean difference between the different groups described above in their estimation of math test performance through the table described in Instruments. The third hypothesis, *H3*: *Participants who score high in abstract reasoning will be less susceptible to anchoring* were to be examined by including the standardized score of the participants on the MaRs-IB as a covariate in the original analysis of variance, effectively making it an analysis of

covariance (ANCOVA). The identification of potential outliers was to be done by utilizing z-scores. By assuming the central limit theorem to hold due the size of the sample collected, no z-scores greater than 3.29 were expected (Field, 2013). To supplement the z-scores the standard deviations found in the pilot study were to be used to give nuance to potential outliers detected by the z-scores.

Results

The statistical analysis began by examining the first hypothesis *H1* in order to establish if anchoring effects extend to the harsher settings of a ratio judgement. To test this hypothesis, I must compare the mean ratio estimation between the groups. The dependent variable is thus the participants estimation of how many correct answers the person doing the math test had (in percentages), and the independent variable is participant condition. By visually reviewing the data issues in normality was found, so in order to follow through on the outlier detection procedure described in above, a Ln-transformation of the dependent value was done. After standardizing the residuals of the Ln-transformed dependent variable, three observations emerged as potential outliers with z-scores exceeding 3.29. These three observations had made estimations of 1% (z-score = 5.3), 2% (z-score = 4.29) and 4% (z-score = 3.3) with regard to the percentage of correct answers on the math test and thus serve as both intuitive and statistical outliers.

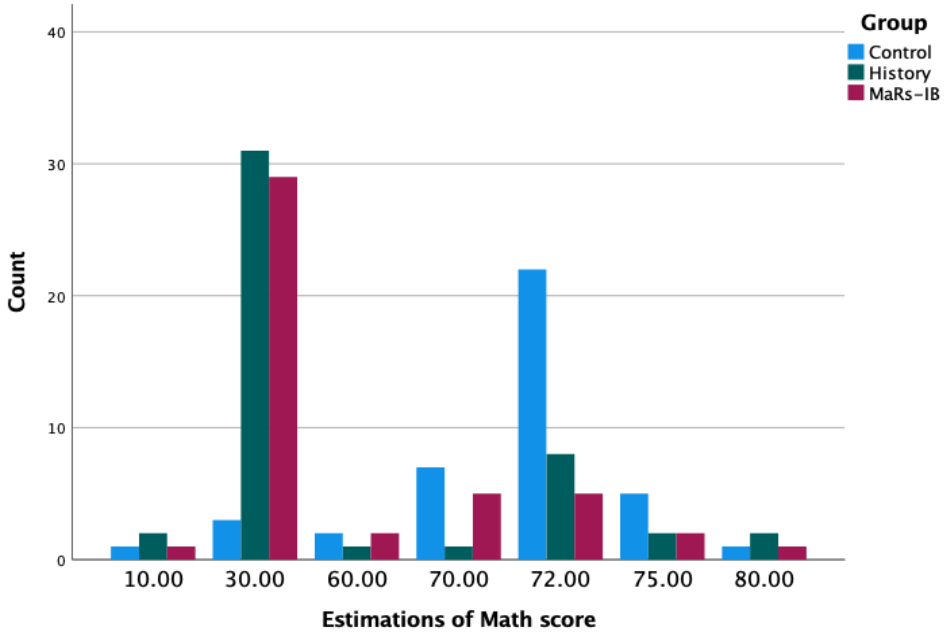
By conducting a one-way ANOVA on the non-transformed dependent variable (participants estimation of math test performance) after the removal of the three observations which had a standardized residual larger than $\pm$ 3.29 (resulting in a sample of n=156 derived from the sample of 159) yielded staggeringly significant results. The omnibus test had an F statistic of F = 14.8, 2, a p-value of $p < .000$ and a fixed effect omega squared effect size of $\omega = .16$. In addition to this, Levene´s test showed a p-value of $p = .85$ indicating equal variances across groups. Post hoc pair ways comparisons with Sidak corrected p-values revealed the driving factors of the significance found in the omnibus test. The mean difference between the control group and the history treatment group was $M = 20.56$, $SE = 4.1$, $p < .000$ and the mean difference between the control group and the MaRs-IB treatment group was $M = 17.73$, $SE = 4.1$, $p < .000$. These results indicate that the treatment groups estimated around 20 percentage points lower than the control group, and subsequently confirms our first hypothesis *H1*, in that participants exposed to the low anchor made lower estimations. Our second hypothesis *H2*, that personal experience of the source of the anchor is a moderating factor of anchor susceptibility, is not confirmed by these results since no significant difference between the two treatment groups was found. That an anchoring effect of this magnitude, effect size and significance would occur even in the harsh estimation settings of the current treatment is astonishing. This would indicate that even when we visually examine an object with an intent of objectively evaluating it, we still fall prey to adjustment effects due to anchor values. However, the results found in this initial model do not tell the whole story.

First and foremost, the distribution of the estimations from the different groups indicated heavily skewed estimations. Out of the 159 responses, roughly 56% were of the same two estimates, 30% correct and 72% correct, corresponding to 63 and 35 responses respectively. If you recall, these are the same numbers that was presented in the survey pertaining to the respondents own score on the MaRs-IB test (of 72% correct, which all

participants received), and the test results from the profile of another individual's test result (30% correct) shown to the respondents in the treatment groups (see Appendix A). As you can expect, the treatment groups were the ones most often estimating 30% correct and the control group was the one most often estimating 72% percent correct. Figure 2 shows the frequencies of different estimations based upon the treatment group that the respondent belonged to.

Figure 2

Frequency of the estimations of math scores by treatment groups



Note. The y-axis "count" describes the number of observations that made the particular estimation shown by the x-axis. The observations are clustered for ease of view.

Additionally, the data violated the conditions for a traditional ANOVA in the sense that homogeneity and normality issues remained even after Ln-transformations and outlier trimming, indicating that a change in analysis procedure was needed. In lieu of the normal ANOVA procedure I instead decided to follow the advice of Tomarken & Serlin (1986) and do a Welch´s ANOVA with Games-Howell *post hoc* corrections. In addition to using a more robust method of ANOVA I also incorporated bootstrapping based on 1000 samples to generate less uncertain confidence intervals of the pairwise comparisons (Kulesa et al, 2015).

The distribution of estimates also calls in to question how we should define outliers. The a priori estimation of an outlier based upon the standard deviation of the estimates in the pilot study indicated that all estimates lower than 40% could be viewed as outliers. However, it is unclear if a blanket rule on what is considered an outlier is appropriate if we wish to examine potential anchoring effects. In order to give a more nuanced analysis, many outlier definitions were examined. Regardless of how outliers were defined many observations would be lost. Due to this reduction in sample size and the following unequal distribution of group sizes I made a new dummy variable indicating

whether the subject had received a treatment or not (i.e., belonged to the control group) in order to replicate any potential effects found in the ANOVAs. Table 1 illustrates how the model values change depending on the outlier exclusion criteria. The Welch´s ANOVA followed a 1x3 design containing the two treatment groups and the control group. The independent samples t-test concerns the collapsed treatment group and the control group.

As indicated in Table 1, the only model where significance is observed in both tests is the second, where all observations lower than 31% were excluded. This exclusion left a sample of n = 77, where the control group made up 39 responses, the history treatment 16 responses and the MaRs-IB treatment 22 responses. The independent sample t-test with the collapsed treatment groups consisted of the same observations, but with the two treatment groups combined to make a collapsed treatment group of n = 38.

Table 1

Original ANOVA with Levene´s test for homogeneity, Welch´s ANOVA and independent samples t-tests on five samples: Statistics, p-values and effect sizes.

		ANOVA			**Levene´s test**		
Model:	Outliers	*F*	*p*	*ω*	statistic	*df*	*p*
1 (n=156)	z< ± 3.29	14.8, 2	>.000 ***	.16 [.056, .26]	.16	2, 151	.85
		Welch´s ANOVA			**Independent samples t-test**		
		F	*p*	*ω*	*t*	*p*	*d*
2 (n=77)	< 31%	3.86, 2	.034 **	.112 [0, .25]	2.68, 44	.01 **	.62 [.16, 1.07]
3 (n=76)	< 35%	3.161, 2	.059 *	.085 [0, .22]	2.43, 44	.019 **	.571 [.11, 1.03]
4 (n=74)	< 40%	1.88, 2	.173	.03 [0, .14]	1.89, 45	.067 *	.45 [0, .91]
5 (n=72)	< 50%	1.55, 2	.363	.011 [0, 12]	1.17, 51	.246	.288 [0, .75]

Note. The table illustrates five models with different outlier criteria. The significance of the p-values is indicated by asterisks as follows, p < .1 = *, p < .05 = **, p < .01 = ***. The brackets specify the 95% confidence intervals.

The Welch´s ANOVA in the second model in Table 1 implies that the means among the groups could not be considered equal, $F(2,25.9) = 3.86$, $p = .034$, $\omega = .112$. The Games-Howell *post hoc* tests revealed that the only observed significant difference were between the control group and the experience condition treatment group (the group who observed the MaRs-IB profile), $M = 8.58$, $SE = 3.2$, $p = 0.032$ with a bootstrapped 95% CI [2.52, 15.3]. In addition to confirming our first hypothesis *H1*, this model also alludes to personal experience of the anchor value as being a moderating factor in the sense that without experience of the source of the anchor, no significant difference from the control group is observed, thus providing some evidence of the second hypothesis *H2*, that personal experience of the source of the anchor moderates the susceptibility to that

anchor, in this case making it more effective. In other words that experience of the source of the anchor increases the anchor susceptibility.

A bootstrapped t-test on the dummy variable consisting of treatment vs control replicated the results in the Welch´s ANOVA with a mean difference of $M = 6$, $SE = 2.2$, $p = .01$ and a point estimate of Cohen´s d, $d = .62$ with a 95% CI [.16, 1.07]. Bootstrapped confidence intervals on the independent sample t-test showed a 95% CI [1.84, 10.72], $p = .018$, the variance in the two samples was not considered equal.

The a priori outlier exclusion criteria of any estimates lower than 40% (model 4) did not support the notion that the means within the groups were different. However, the independent sample t-test was significant at $a < .1$, $t(45) = 1.89$, $p = .067$ showing a mean difference of $M = 3.28$, $SE = 1.75$ with a bootstrapped 95% CI [-.11, 6.95] and a Cohen´s d point estimate of $d = .45$ indicating that the treatment group made slightly lower estimations, $M = 68.47$, $SD = 9.6$, compared to the control group, $M = 71.76$, $SD = 4.11$. It should however be noted that the significance observed in model 4 (outliers < 40%) is completely dependent on two observations, which if excluded completely eliminates any significance, as observed in model 5 (outliers < 50%). *Post hoc* power analyses were not conducted in accordance with (Hoenig & Heisey, 2001).

Discussion

This study had three goals, I) to examine whether an anchoring effect could be found in ratio judgements where the estimator has visually observed the ratio of interest, II) to examine whether personal experience of the source of anchoring value affected anchor susceptibility, and III) to examine whether abstract reasoning moderated anchoring susceptibility. An while the design of the study allowed us to pursue these goals, the data collected was insufficient. The three hypotheses that were established to cater to the specific goals of this study could not be properly examined and subsequently, no conclusions about the veracity of these hypotheses could be made. The average estimation of the amount of correct answers was exceedingly accurate regardless of whether the respondents in the study had received and anchor or not. However, it could be argued that none of these hypotheses could be tested reliably due to the distribution of the data. Most notably, the third hypothesis concerning abstract reasoning skills and anchoring susceptibility could not be properly examined due to the distribution of the estimations. This inability to make inferences on the original hypotheses does however not necessarily render the data unusable in any analytic capacity.

The distribution seen in Figure 2 is quite interesting through an anchoring perspective. It is quite clear that an adjustment process has not taken place in many of the estimations, in the sense that the estimations are not skewed in the direction of the anchor, but a reiteration of the anchor value itself. An overwhelming number of respondents estimated that the amount of correct answers on the math test were 30%. At first glance this suggest that the respondents misunderstand the question and simply entered the answer that were shown to them in the profile, or that the low anchor in the profile actually made them answer how many percent of the answers were *incorrect* rather than correct. The notion that the respondents didn't actually understand the questions is furthered by the incongruency between the comparative and absolute judgements. Many respondents say that the profile score was better than the math score only to estimate that the math score was around 70%. This does not make any sense, since 70% is more than 30%. This

would suggest that the comparative judgement of scoring better, worse or the same compared to the score in the profile was arbitrary. The issue with that assertion is that only 1 out of the 38 respondents who did the comparative judgement reported that the scores were about the same. If the respondents chose randomly, this distribution of choices is unlikely, but since I did not randomize the order of those alternatives I cannot be certain of this, and since the question was formulated as better/worse it is conceivable that the respondents just guessed and viewed the third alternative as a non-option.

Explaining the strange distributions of estimates as a misinterpretation of the question seems like a likely story. The issue with this interpretation is that the questions in the survey was very clearly formulated and many of the participants did indeed interpret them correctly. A better explanation, in my opinion, is that the participants who made the 30% correct answers estimation, simply did not read the questions at all. If this were to be the case, it would also illustrate a big mistake made when gathering the data. When the survey was advertised it was described as a short survey about memory, examining recollection and abstract reasoning. Given that the participants would read the questions, this formulation seemed to describe the tasks given to the respondents quite well. Conversely, if the participants would not read the questions, this framing of the task as a recollection and memory task would serve as an explanation as to why so many participants estimated a math score of 30%, they recalled that they had seen 30% before, and simply entered that number. This is of course only applicable to the participants in the treatment groups since they were the only group exposed to the number 30. This explanation also lends itself to explain the abundance of estimations of 72% correct answers. Like 30%, 72% was observed by the respondents (as feedback of their own score on the MaRs-IB) and therefore subject to the possibility of recollection fuelled reiteration by the respondent when they were given the task of estimating the amount of correct answers on the math test. However, since 72% is a plausible answer, this estimation is not as alarming as that of 30% and thus didn't skew the estimation to the same degree.

Regardless of the numerical accuracy in the estimation, the main point of discussion is whether or not the participants knew what they were estimating, that being the amount of correct answers on a math test. If we entertain the possibility that they did, it would imply that the priming of the feedback and the profile values together with the framing of the task resulted in very skewed estimations. Whether or not to call this an anchoring effect is debatable, I would not categorize reiteration of numerical values as anchoring. Personally, I believe that the combination of priming and framing resulted in the observed estimations, but I do not believe that these estimations reflect the question posed to the participants, rather, the estimation simply reflect the priming value, be it 30 or 72. I encourage the reader to view the Appendix containing the questions in the survey, ponder the plausibility of the estimations seen in Figure 2, and make their own judgements regarding the objective of the participants estimations.

Conclusions, limitations and further research

This study could not find any convincing evidence for anchoring effects in ratio judgements where the estimator had visually evaluated the ratio, or observe a difference in susceptibility to anchoring effects due to personal experience of the source of the anchor value or degree of abstract reasoning capabilities. However, this is not due to the potential of the hypotheses posed in the study to be untrue, rather, it is due to the

distribution of the data that was collected. This leads to the first limitation of this study, the sample size. It is possible that a greater sample size would generate more usable observations in order to examine any potential anchoring effects and therefore mitigate the overall distribution of the data. However, even if the hypothesized effects occurred, the possible inferences would be limited. First of all, the treatment in this study lacks variance, it concerns a single table of green and red rectangles which doesn't lend itself to further generalizations beyond that particular table. Secondly, the shortened instrument used to gage abstract reasoning has not been proven to be an adequate substitute for its full procedure. And thirdly, the current model didn't incorporate any a priori probability values even though much research on the subject is available.

The second limitation of this study is the absence of validation checks in the estimation process and the framing of the survey. I should not have framed the survey as a memory task, and I should not have accepted estimations that didn't reflect a contingency between absolute and comparative judgements. In hindsight, giving the participants their results on the MaRs-IB might also have been a bad idea given the prevalence of 72% correct estimations. Simply, the design of the study was not optimal.

Nevertheless, the subject of anchoring effects in ratio judgements should be examined more since literature on the subject was hard to come by. And while this study was flawed, the treatment of estimating the ratio of colours to each other seems to be a good and simple way of observing any potential effect. Formulating a study where participants view and estimate the ratio of colours in an array of tables following the traditional anchoring paradigm might serve as a great way of incorporating variation in the treatments as well as isolating a potential anchoring effect in a harsher setting than usually examined.

References

Baddeley, A. (1996). *Exploring the Central Executive*. 24.

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*, *63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Bergman, O., Ellingsen, T., Johannesson, M., & Svensson, C. (2010). Anchoring and Cognitive Ability. *Economics Letters*, *107*, 66–68. https://doi.org/10.1016/j.econlet.2009.12.028

Brewer, N. T., & Chapman, G. B. (2002). The fragile basic anchoring effect. *Journal of Behavioral Decision Making*, *15*(1), 65–77. https://doi.org/10.1002/bdm.403

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the Irrelevant: Anchors in Judgments of Belief and Value. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases* (1st ed., pp. 120–138). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.008

Chierchia, G., Fuhrmann, D., Knoll, L., Piera Pi-Sunyer, B., Sakhardande, A., & Blakemore, S.-J. (2019). The matrix reasoning item bank (MaRs-IB): Novel, open-access abstract reasoning items for adolescents and adults. *Royal Society Open Science*, *6*, 190232. https://doi.org/10.1098/rsos.190232

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Costa, P., & McCrae, R. R. (1999). A five-factor theory of personality. *The Five-Factor Model of Personality: Theoretical Perspectives*, *2*, 51–87.
DeYoung, C. G. (2015). Cybernetic Big Five Theory. *Journal of Research in Personality*, *56*, 33–58. https://doi.org/10.1016/j.jrp.2014.07.004
DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2005). Sources of Openness/Intellect: Cognitive and Neuropsychological Correlates of the Fifth Factor of Personality. *Journal of Personality*, *73*(4), 825–858. https://doi.org/10.1111/j.1467-6494.2005.00330.x
Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. SAGE.
Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General*, *141*(1), 124–133. https://doi.org/10.1037/a0024006
Goldin, C., & Rouse, C. (2000). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *THE AMERICAN ECONOMIC REVIEW*, *90*(4), 59.
Hoenig, J. M., & Heisey, D. M. (2001). The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician*, *55*(1), 19–24. https://doi.org/10.1198/000313001300339897
Kahneman, D. (2011). *Thinking, fast and slow* (p. 499). Farrar, Straus and Giroux.
Kahneman, D., & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. *Heuristics and Biases: The Psychology of Intuitive Judgment*, *49*, 49–81. https://doi.org/10.1017/CBO9780511808098.004
Klein, R. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142. https://doi.org/10.1027/1864-9335/a000178
Mochon, D., & Frederick, S. (2013). Anchoring in sequential judgments. *Organizational Behavior and Human Decision Processes*, *122*(1), 69–79. https://doi.org/10.1016/j.obhdp.2013.04.002
Mussweiler, T., & Englich, B. (2005). Subliminal anchoring: Judgmental consequences and underlying mechanisms. *Organizational Behavior and Human Decision Processes*, *98*(2), 133–143. https://doi.org/10.1016/j.obhdp.2004.12.002
Mussweiler, T., & Strack, F. (1999). Comparing Is Believing: A Selective Accessibility Model of Judgmental Anchoring. *European Review of Social Psychology*, *10*(1), 135–167. https://doi.org/10.1080/14792779943000044
Mussweiler, T., & Strack, F. (2001). The Semantics of Anchoring. *Organizational Behavior and Human Decision Processes*, *86*(2), 234–255. https://doi.org/10.1006/obhd.2001.2954
Norem, N. N. C., Julie K. (2019). Are Big Five Traits and Facets Associated With Anchoring Susceptibility? - Nathan N. Cheek, Julie K. Norem, 2020. *Social Psychological and Personality Science*. https://journals-sagepub-com.ezproxy.ub.gu.se/doi/full/10.1177/1948550619837001
Reitsma-van Rooijen, M., & L. Daamen, D. D. (2006). Subliminal anchoring: The effects of subliminally presented numbers on probability estimates. *Journal of Experimental Social Psychology*, *42*(3), 380–387. https://doi.org/10.1016/j.jesp.2005.05.001
Röseler, L., Schütz, A., Blank, P. A., Dück, M., Fels, S., Kupfer, J., Scheelje, L., & Seida, C. (2021). Evidence against subliminal anchoring: Two close, highly powered,

preregistered, and failed replication attempts. *Journal of Experimental Social Psychology*, *92*, 104066. https://doi.org/10.1016/j.jesp.2020.104066

Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors? *Journal of Personality and Social Psychology*, *99*(6), 917–932. https://doi.org/10.1037/a0021540

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research*, *177*(3), 1333–1352. https://doi.org/10.1016/j.ejor.2005.04.006

Smith, A. R., Windschitl, P. D., & Bruchmann, K. (2013). Knowledge matters: Anchoring effects are moderated by knowledge level: Knowledge level moderates anchoring effects. *European Journal of Social Psychology*, *43*(1), 97–108. https://doi.org/10.1002/ejsp.1921

Evans, J. B. T. (2008). *Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition AR Further*.

Strack, F., & Mussweiler, T. (1997). Explaining the Enigmatic Anchoring Effect: Mechanisms of Selective Accessibility. *Journal of Personality and Social Psychology*, *73*, 437–446. https://doi.org/10.1037/0022-3514.73.3.437

Teovanović, P. (2019). Individual Differences in Anchoring Effect: Evidence for the Role of Insufficient Adjustment. *Europe's Journal of Psychology*, *15*(1), 8–24. https://doi.org/10.5964/ejop.v15i1.1691

Teovanović, P., Knežević, G., & Stankov, L. (2015). Individual differences in cognitive biases: Evidence against one-factor theory of rationality. *Intelligence*, *50*, 75–86. https://doi.org/10.1016/j.intell.2015.02.008

Tomarken, A., & Serlin, R. (1986). Comparison of anova Alternatives Under Variance Heterogeneity and Specific Noncentrality Structures. *Psychological Bulletin*, *99*, 90–99. https://doi.org/10.1037/0033-2909.99.1.90

Townson, C. (2019). *The Anchoring Effect: A Meta-Analysis* [Ph.D., Michigan State University]. http://search.proquest.com/docview/2311653546/abstract/952F20B2265F45A7PQ/1

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Wegener, D. T., Petty, R. E., Blankenship, K. L., & Detweiler-Bedell, B. (2010). Elaboration and numerical anchoring: Implications of attitude theories for consumer judgment and decision making. *Journal of Consumer Psychology*, *20*(1), 5–16. https://doi.org/10.1016/j.jcps.2009.12.003

Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). *A New Look at Anchoring Effects: Basic Anchoring and Its Antecedents*. 16.

Appendix A

Questions and treatment table (Math results)

Treatment group profile and messaging

You will now be shown a profile of a participant on a previous study who also did the MaRs-IB test

*Note***.** This was shown to the MaRs-IB treatment group before they viewed the profile seen below

You will now be shown a profile of a participant from a previous study. The participant did a History test.

*Note***.** This was shown to the history treatment group before they viewed the profile seen below

Participant: Alex. E
Age: 22
Education level: some college

Participant had: 30% correct answers

*Note***.** This is the profile that the participants in the treatment group received containing demographic information as well as the anchor of 30% correct answers on the corresponding test.

Math test performance estimation instructions

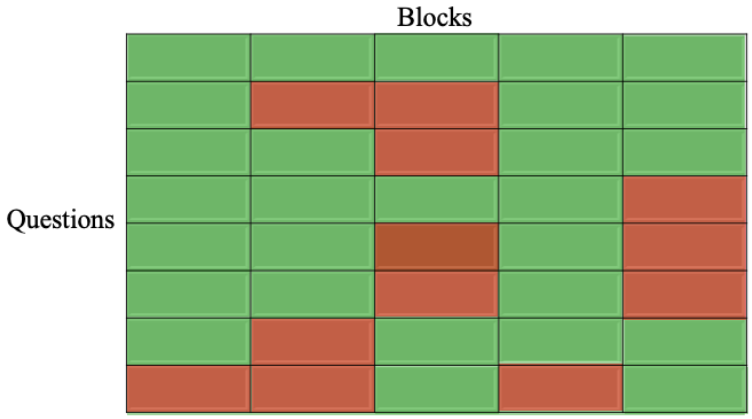
Your next task is to guess how many correct answers a person doing a test had.

On the next screen you will view the results on a Math test. The test is comprised of 40 questions arranged in 5 blocks. There are 8 questions(rows) for every block(column). Correct responses are coloured in green and the incorrect responses in red.

The table will only be visible for a short time so be prepared when you change the page. Good luck!

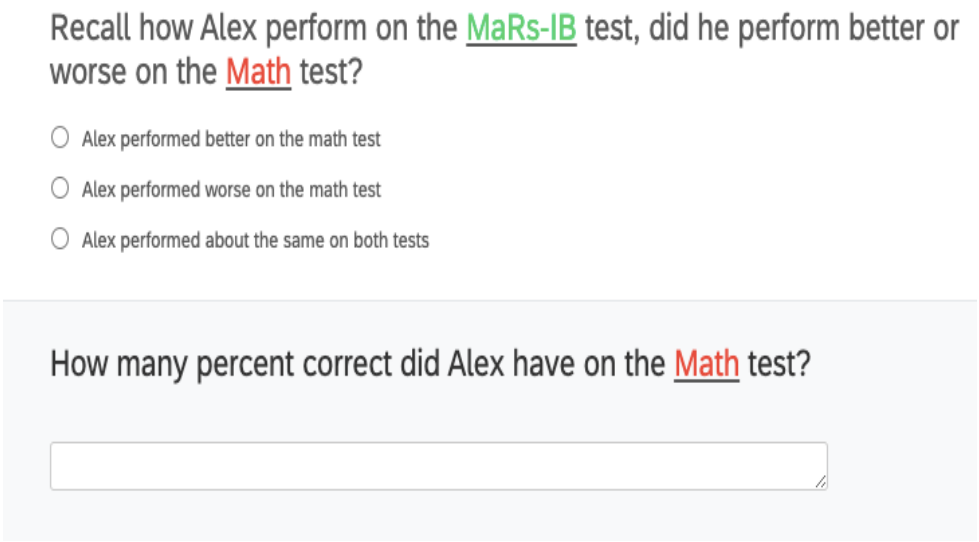
Note. This is the instructions to the estimation task. The objective being to estimate the amount correct answers, id est the number of green rectangles seen in the table below.

Treatment table, 40 question math test.

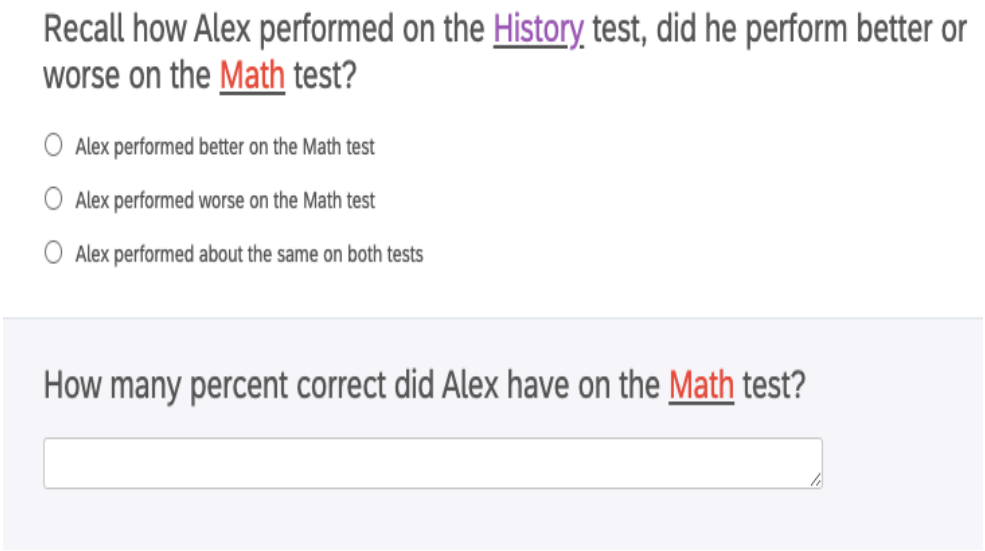


Note. This is the table shown to the participants. Each rectangle symbolizes a question. The colour of the rectangle indicates whether that question was answered correctly or not. There are 28 green rectangles and 12 red ones, representing 70% correct answers and 30% incorrect answers.

Comparative and absolute estimation for the treatment groups



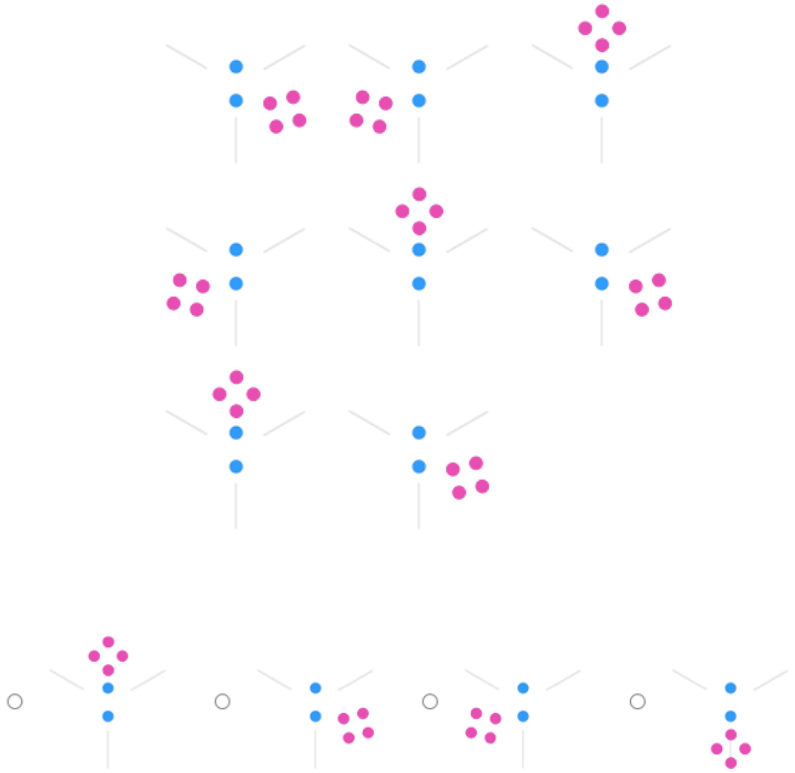
Note. This is the comparative judgement between the results in the profile and the results indicated by the table, and the absolute estimation of the amount of correct answers in the table, for the MaRs-IB treatment group.



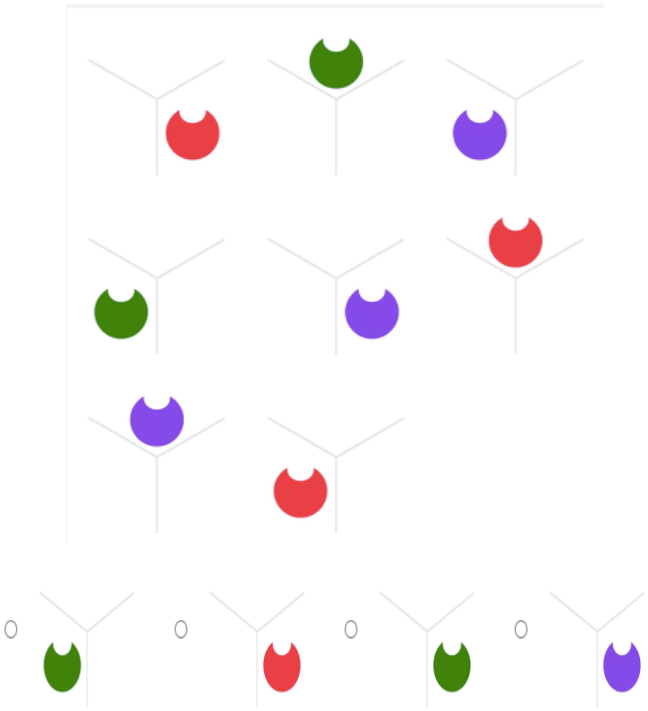
*Note***.** This is the comparative judgement between the results in the profile and the results indicated by the table, and the absolute estimation of the amount of correct answers in the table, for the history treatment group. Participants in the control group only saw the absolute estimation.

Appendix B

Samples from the MaRs-IB



*Note***.** This is an example of a MaRs-IB question. The task is to detect the pattern in the matrix and select the symbol shown in the row below the matrix that completes the pattern. In this instance the rule that dictate the image in the matrix is that for each column, the four red circles moves a "step" to the right in a rotation like motion. The correct answer is the third option from the right.



*Note***.** This is another example of a MaRs-IB question. In this instance the rule that dictate the image in the matrix is that for each column, the circle-like shape changes colour and placement, thus this task is judged to be a little harder since both placement and colour is changed. The correct answer is alternative number three from the right in the row bellow.