Simple Belief Elicitation: an experimental evaluation *

Karl Schlag† James Tremewan‡

November 25, 2020

Abstract

We present a method for eliciting beliefs about probabilities when multiple realisations of an outcome are available, the "frequency" method. The method is applicable for any reasonable utility function. Unlike existing techniques that account for deviations from risk-neutrality, this method is highly transparent to subjects and easy to implement. Rather than identifying point beliefs these methods identify bounds on beliefs, thus trading off precision for generality and simplicity. An experimental comparison of this method and a popular alternative, the Karni method, shows that subjects indeed find the frequency method easier to understand. Significantly, we show that confusion due to the complexity of the Karni method leads to less cognitively able subjects erroneously stating a belief of 50%, a bias not present in the frequency method.

Keywords: Belief elicitation, Experiment, Frequency format, Cognitive ability.

JEL-CLASSIFICATION: D81, C91

*An earlier working paper version with the same main title (Schlag and Tremewan, 2014) did not contain any experiment.

†Department of Economics, University of Vienna, Vienna, AUSTRIA.

‡Corresponding Author: Department of Economics, University of Auckland, 12 Grafton Road, 1010 Auckland, NEW ZEALAND; Email: *james.tremewan@auckland.ac.nz.*

1 Introduction

Experimental economists are increasingly recognising the value of directly eliciting the beliefs of their subjects. For example, direct measures of a subject's beliefs can help us disentangle whether deviations from homo economicus behaviour is due to social preferences or bounded rationality, test whether belief updating is Bayesian, and learn about whether peer effects are caused by imitation or information transmission.

There are, by now, a large variety of incentive compatible methods for eliciting beliefs, with various strengths and weaknesses.[1] Some of the most commonly used are not incentive compatible for risk-averse subjects, (e.g. linear or quadratic scoring rules) while those that do not suffer from this flaw tend to be either time-intensive, (e.g. callibrating elicited beliefs; Offerman et al., 2009) or challenging for subjects to fully understand (e.g., variations of the Becker-DeGroot-Marschak mechanism; Karni, 2009). Here we present a procedure for eliciting beliefs about probabilities that is robust to risk-aversion, requires minimal labtime, and is simple for subjects to understand. We refer to this procedure as the "frequency" method.

In this paper we lay out the theoretical properties of the frequency method and demonstrate its practical and empirical properties in a laboratory experiment. On the theory side we establish robustness to risk attitudes and point out the inferences that can be made from subjects' reports. On the practical and empirical side we showcase the benefits of this method by comparing its performance in a laboratory experiment to that of the elicitation method of Karni (Karni, 2009). In particular we evaluate ease of implementation, understanding of subjects, and reasonableness of elicited reports. We emphasize that we do not conduct a "horse race" designed to determine the "best" method for eliciting beliefs. Instead, our objective is to highlight the properties of the frequency method by comparing it to the popular Karni mechanism. The Karni mechanism is likely to be viewed by experimenters as a leading contender for use in their own experiments, and is therefore an appropriate comparison for us.

Using the frequency method, subjects report better understanding of the belief elicitation task and complete it in shorter time. There are fewer reports of the focal probability of 0.5, which in the Karni method is correlated with low cognitive ability.[2] The two methods do not differ in terms of the

[1] For detailed and comprehensive discussions see Schlag et al. (2015), Schotter and Trevino (2014), and Charness et al. (2020).

[2] This complements the results of a related study by Burfurd and Wilkening (2020) who design an experiment to evaluate the relationship between cognitive ability and the empirical properties of the Karni method. They find that there is greater variation in

proportion of subjects best-responding to their stated beliefs, or average distance from the empirical probability. However, the frequency method results in more correct answers in a Bayesian updating task.

Most of the literature on belief elicitation focuses on payments based on the actual outcome of a single event. However, in many laboratory experiments, there will be not just one but many independent realisations of the random variable of interest. Take, for example, a one-shot prisoners' dilemma experiment where the experimenter is interested in beliefs the subjects hold about the probability of defection. If there are 20 subjects per session, each stated belief can be matched with the 19 realizations of the decisions of others. The two methods we discuss in this paper, the first for eliciting probabilities, the second for quantiles, take advantage of these multiple realisations. In doing so we remove the need to refer in experimental instructions to numerical probabilities of single events, which many subjects may have difficulty understanding (see Section 6 for evidence of this).

In the frequency method, the subject is asked to guess the empirical frequency of each outcome. A prize is then awarded if and only if their guess coincides with the realized frequencies. For the case of only two outcomes, this method has been used before (Wilcox and Feltovich, 2000; Bhatt and Camerer, 2005; Hurley and Shogren, 2005; Costa-Gomes and Weizsacker, 2008; Blanco et al., 2010; Le Coq et al., 2015), however its properties do not appear to have been well understood by the experimental community. Wilcox and Feltovich (2000) and Blanco et al. (2010) state only that beliefs about the modal frequency of outcomes are elicited, while Costa-Gomes and Weizsacker (2008) say that it is valid only when the true subjective probability coincides exactly with one of the possible empirical distributions. For the special case of binary outcomes, a correct interpretation was reported in Hurley and Shogren (2005) but not given much prominence in the paper, and as a result appears to have been largely overlooked.[3]

Not only does the frequency method elicit beliefs about modal frequencies, but we also show that it also enables the researcher to identify a region in which the belief of the subject should lie. Inference does not require postulating any assumptions on the utility function beyond assuming that the subject strictly prefers getting the prize to not getting it. This method reveals regions of beliefs for events that have an arbitrary number of possible

report accuracy between individuals who are classified as high and low ability in the Karni mechanism relative to an unincentivized benchmark (introspection).

[3]Another reason the frequency method may have been largely disregarded is that the main message of Hurley and Shogren (2005) is that they failed to recover induced beliefs. However, on closer inspection, this failure is attributed to the induction process rather than the elicitation method.

outcomes. With binary events this region is an interval of width $1/(n+1)$, where n is the number of realizations of the variable in question. For example, for $n = 19$, such as in the prisoners' dilemma example given above, the size of the interval is 5%. In this case, given that subjects tend to answer questions about percentages in multiples of five (Manski, 2004), there is no practical loss of precision. For binary events we show that this method is most precise in a well defined sense. We also show that this method can be used to estimate bounds on subjects' beliefs about means and variances of distributions.

The frequency method stands out for the simplicity of its implementation. We demonstrate its practical and empirical properties in the laboratory using the elicitation method of Karni as a benchmark. We choose the Stag Hunt game as a simple environment in which we can evaluate the relationship between elicited beliefs and actions. Questions are added to compare subjects' understanding of the two elicitation methods. In an Urn task we compare elicited beliefs to an objectively true probability. We use the Cognitive Reflection Test to provide additional insights into differences in the cognitive requirements of each method.

We also test for the first time a novel and related method, using multiple realisations of a random variable to elicit beliefs about quantiles of a distribution. As with the frequency method for eliciting probabilistic beliefs, it is extremely straightforward to explain to subjects, and is equally valid for all non-trivial utility functions. However, in contrast to the probability elicitation method, we find it performs poorly in terms of the internal consistency of the elicited beliefs.

The paper proceeds as follows: Section 2 describes the theory underlying the frequency method; Section 3 describes our experimental design and Section 4 provides the results; Section 5 gives a brief outline of the quantile elicitation method and an overview of its performance; in Section 6 we discuss the implications of our findings and conclude.

2 Theory

In this Section we present a method for eliciting probabilities and derive tight bounds on the "true" underlying probabilities.

Let Y be a random variable with k possible outcomes $s_1, .., s_k$, where p_i is a subject's subjective belief about the probability that outcome s_i will occur. We elicit information about a subject's belief by incentivizing their reports using n independent realizations of Y. Specifically, subjects are asked to report $b = (b_1, .., b_k)$, b_i being a non-negative number for all i, and are paid a prize of value R if and only if for all $i \in \{1, ..., k\}$ the reported value b_i is

equal to the number of times s_i occurs out of n independent realisations of Y. We call this the frequency method.

Note that multiple independent realizations of a variable Y are naturally available in laboratory experiments where subjects are randomly matched with one of n other subjects to play a game. Each s_i is a strategy available to the subject's matched partner. The subject is then told that they will be awarded a prize if they can correctly guess the number of people in their partner's role who play each strategy. So the choices made among these potential partners constitutes the n independent realizations.

From the standpoint of the subject making the report, the prize will be awarded with probability

$$f(b) = \frac{n!}{b_1! \cdot .. \cdot b_k!} \prod_{i=1}^{k} p_i^{b_i}.$$

It follows immediately that the subject maximizes expected utility if and only if they maximize the probability f of receiving the prize. Hence, and without loss of generality, we are interested in the relationship between the maximizers of f and the underlying subjective beliefs. In the following we provide a complete characterization of this relationship.

Let B be the set of feasible reports, so

$$B = \left\{ b \in \{0, 1, .., n\}^k : b_i \geq 0 \forall i, \sum_{i=1}^{k} b_i = n \right\}.$$

Proposition 1 *Consider $b \in B$. Then b maximises f over all B if and only if*

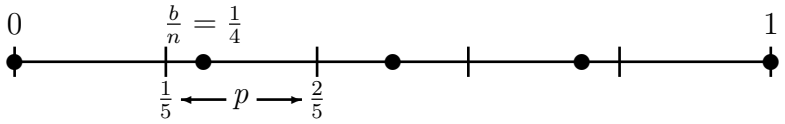
$$\begin{aligned} \frac{b_i}{b_j + 1} &\leq \frac{p_i}{p_j} \leq \frac{b_i + 1}{b_j} \ \forall j \neq i \text{ when } p_j, b_j \neq 0 \\ b_j &= 0 \text{ if } p_j = 0. \end{aligned} \tag{1}$$

In particular, if b maximizes f then

$$\frac{b_i}{n + k - 1} \leq p_i \leq \frac{b_i + 1}{n + 1} \text{ holds for all } i.^4 \tag{2}$$

Figure 1 demonstrates this result for $k = 2$ and $n = 4$. The dots show the possible reports (divided by n) and the surrounding intervals show the possible values of p given the reports. In the figure we see that only those beliefs on the boundary between two regions give rise to two different optimal

Figure 1: Reported and consistent true beliefs for $k = 2$ and $n = 4$



reports. More generally, our proof of Proposition 1 reveals that any subject with beliefs that satisfy (1) with strict inequalities has a unique best report.

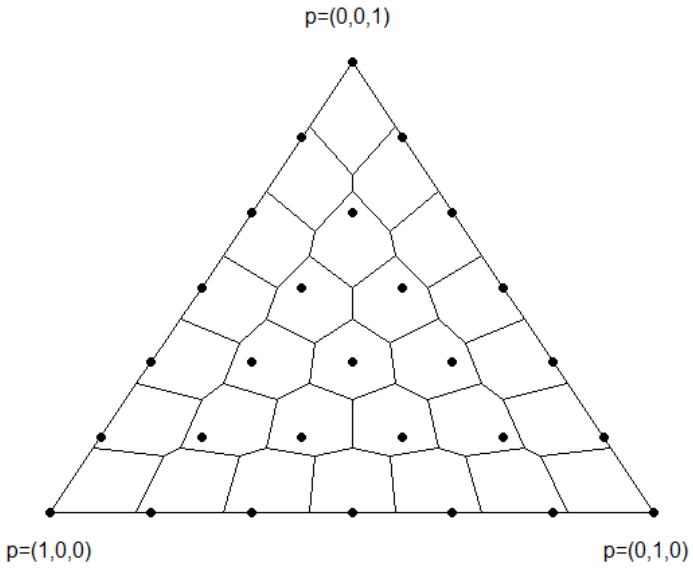
For $k = 2$ we hasten to point out that one cannot extract more precise information for an an arbitrary utility function in the following sense. Consider any alternative payment rule with the same input, that is a subject's stated belief about the number of times that an outcome will occur. For a given utility function u let P_b^u be the set of beliefs under which it is optimal under the alternative rule to report b, $b \in \{0, 1, .., n\}$. Then $\cup_{b\in\{0,1,..,n\}} P_b^u = [0, 1]$. Let $d(P_b^u)$ be the maximal distance between any two points belonging to P_b^u (where d is its width if P_b^u is an interval). It is easy to verify that $\max_{b\in\{0,1,..,n\}} d(P_b^u) \geq 1/(n + 1)$. Let the *minimal precision* of a rule be the negative of the maximal difference between any two probabilities that lead to the same report. Hence, we find that there is no payment rule with a strictly higher minimal precision than the one we have presented. In fact, it is easy to see that the inferred true probabilities of any rule with this value of minimal precision are unique. We summarize.

Proposition 2 *Any alternative rule that elicits the frequency of the occurrence of a single event (so $k = 2$) has a strictly lower minimal precision than that of the frequency guessing method.*

In general the set of feasible probabilities is constrained by $p_i \geq 0$ for all i, by $\sum_{i=1}^{n} p_i = 1$ and by the constraints given in (1). Figure 2 shows how these constraints divide the simplex into regions of feasible combinations of "true" beliefs given each report, for $k = 3$ and $n = 6$.

[4]For the special case of k=2, this result appears in Hurley and Shogren (2005).

Figure 2: Reported and consistent true beliefs for $k = 3$ and $n = 6$



We also note that once probability distributions have been elicited using this method, bounds on means and variances can also be computed (see Schlag and Tremewan (2014) for details).

3 Experimental Design

Our experiment consisted of three parts: a Stag Hunt game, an Urn Task, and a public goods game. Each subject participated in all three parts. The Stag Hunt game and Urn task were used to compare probabilistic beliefs elicited using the frequency and Karni methods. The public goods game was used to test the internal consistency of quantiles elicited with our new method. That part of the experiment will be presented later in Section 5.

Subjects first participated in a Stag Hunt game. Subjects chose an action in the game, stated beliefs about the probability of others choosing Stag, and answered four questions about their comprehension of the belief elicitation task. Subjects received €2 for sure if they chose A, €3 if they chose option B and their partner also chose B, or €0 if they chose option B and their partner chose A. There were two different treatments. In the treatment that elicited beliefs using the frequency method, subjects were asked how many out of

20 randomly chosen subjects from the session (themselves and their partner excluded) chose B, and told they would receive €2 if their guess was correct. In the treatment using the Karni method, the instructions were based closely on those from Dal Bó et al. (2017), with minor changes made to fit our game.

After beliefs were elicited (referred to as "Part 2" in the instructions), we asked each subject the following four questions:

1. How well do you feel you understood the task in Part 2?
2. How easy was it for you to come up with your answer to the task in Part 2?
3. How unsure or how confident are you that you gave the best answer?
4. In Part 2 you were asked about
    - how many participants out of 20 chose Option B. When you chose between Options A and B, how important was it for you to think about how many participants would choose Option B? [frequency treatment]
    - the chances that a randomly selected participant chose Option B. When you chose between Options A and B, how important was it for you to think about the chances that a randomly selected participant would choose Option B? [Karni treatment]

Answers to all four question were elicited on a seven point Likert scale. In addition, we also asked subjects if it would it have been helpful to ask the experimenter a question when making their decision in the belief elicitation stage (subjects were informed at the beginning of the experiment that they would not be able to ask for help with any of the instructions).

Next, subjects participated in an Urn task. Subjects were shown two urns, the first with nine purple balls and one green ball, and the second with nine green balls and one purple balls. They were told that the computer will select one urn randomly, then draw a ball from that urn, show the colour to the subject, then replace the ball. In the frequency treatment, subjects were told that the computer will draw 20 balls with replacement from the same urn, and they must guess the number of those 20 balls that have the same colour as the first. In the Karni treatment we used the Karni method to elicit the probability that a new ball drawn from the same urn is the same colour as the first, again using instructions as close as possible to those in Dal Bó et al. (2017).

The experiment continued with the publics good game described in Section 5 and concluded with a questionnaire requesting basic demographic data and responses to the Cognitive Reflection Test (CRT).

Subjects were recruited using ORSEE (Greiner, 2015), a total of 84 for each of the two treatments. Half of the subjects in each session were assigned to each treatment. Subjects were not shown any results until the end of the experiment, so each subject can be viewed as an independent observation. The experiment was programmed in jtree (Powell, 2019). The experiment lasted approximately minutes and subjects received on average €12.50.

4 Results

We report on the result of our experiment.

4.1 Stag Hunt games

In the Stag Hunt game, 38% of subjects chose B, a proportion which did not differ significantly across treatments (exact z-test (Suissa and Shuster, 1985), $p = 0.529$). The distribution of beliefs by treatment are shown in Figure 3. For comparability, data for the Karni treatment are grouped such that each bin contains the probabilities consistent with a specific response in the frequency treatment, i.e. in intervals of width $\frac{1}{21}$. We note that 88% of responses in the Karni treatment are multiples of 0.05. Therefore, the fact that the frequency method elicits intervals rather than point beliefs results in minimal loss of precision.

A Mann-Whitney test finds no statistical difference between the treatments ($p = 0.923$), however the spike at 0.5 in the Karni treatment is striking. Indeed, the proportion of subjects stating a probability of exactly 0.5 in the Karni treatment is substantially higher than those choosing 10 balls in the frequency treatment, with the difference strongly statistically significant (Frequency: 0.05; Karni: 0.27; exact z-test, $p < 0.01$).

Following the literature, we compare the elicitation mechanisms in two ways. First we compare the average distance of stated beliefs from the actual proportion of subjects choosing B. We then consider the proportion of subjects "best-responding" to their stated beliefs, assuming risk-neutrality. Here we need a point belief, so use the midpoint of the interval elicited in the frequency method.[5]

[5]Using $\frac{x}{20}$ where x is the number of balls stated by the subject makes no difference to the results.

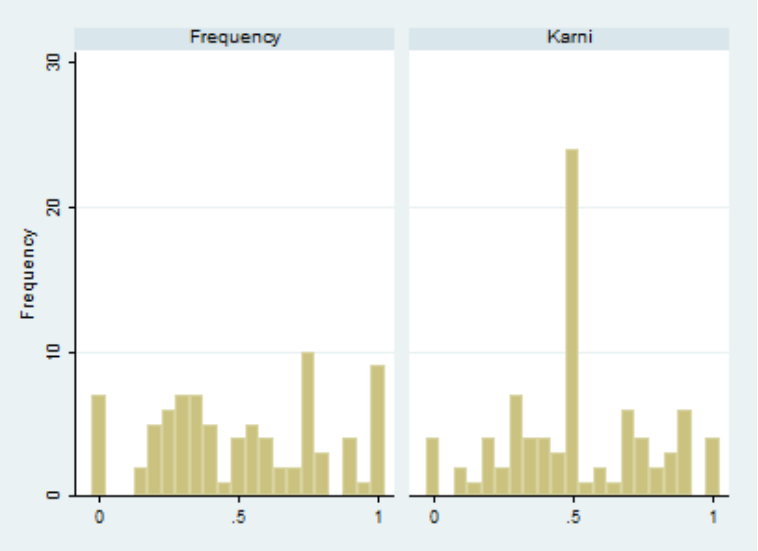


Figure 3: Distribution of beliefs about the probability of choosing Option B.

There was no statistical evidence that the distributions of average distances of beliefs from the actual proportion choosing B differed across treatments, whether using session-specific proportions (Frequency: 0.26; Karni: 0.24; MW $p = 0.480$), or the proportion across all sessions (Frequency: 0.25; Karni: 0.23 ; MW $p = 0.693$). There was likewise no evidence that the proportion of subjects best-responding to their stated beliefs differed across treatments (Frequency: 0.76; Karni: 0.83; exact z-test $p = 0.282$).

Of practical interest to experimentalists is the time the belief elicitation methods take to implement. From the time instructions first appeared on the screen, subjects in the Karni treatment took on average 162 seconds to enter their answer, compared to a significantly lower 56 seconds for the frequency method (Stochastic inequality test (Schlag, 2008), $p < 0.01$).[6]

The full distributions of responses to the four comprehension questions are shown in Appendix C. Subjects' self-reported understanding of the frequency method was statistically higher than the Karni method (Frequency: 6.7; Karni: 5.7; Stochastic inequality test, $p < 0.01$). The improvement in self-reported understanding remains significant when looking separately at those who perform above and below the median in the CRT.

[6]Without assumptions that are unrealistic given our data, the Mann-Whitney test can only identify a difference in distributions, not central tendencies. In this paper, when distributions are statistically different according to a Mann-Whitney test, we use the stochastic inequality test to test for a directional difference (Schlag, 2015).

There was no statistically significant difference between treatments in the distributions of how easy it was to come up with a response (MW $p = 0.493$) or confidence in responses (MW $p = 0.711$). Subjects viewed it as (weakly) less important on average to think about how many participants chose B in the frequency treatment, than thinking about the chances that a randomly selected participant chose B in the Karni treatment (Frequency: 4.8; Karni: 5.5; Stochastic inequality test, $p = 0.076$). More subjects stated that it would have been helpful to ask a question about the instructions in the Karni treatment, but the difference was not statistically significant (Frequency: 0.13; Karni: 0.21; exact z-test $p = 0.170$).

The probability of stating 0.5 in the belief elicitation task, disaggregated by treatment and number of correct responses to the CRT are shown in Figure 4. As can be clearly seen, in the Karni treatment the probability of stating 0.5 is negatively related to cognitive ability, as measured by this task. The average number of correct responses for those who stated 0.5 is lower than those who did not (1.0 and 1.6, respectively). The distributions of numbers of correct responses are different (MW, $p = 0.040$), but a Stochastic inequality test finds no evidence of a directional difference ($p = 0.223$). Non parametric z-tests, however, find that the proportion of subjects stating 0.5 is higher for those who had no correct answers compared to those who had at least one correct (CRT$=0$: 0.45; CRT>0: 0.22; exact z-test, $p = 0.047$), and also higher for those who scored at most one compared to those who scored two or three (CRT≤ 1: 0.35; CRT>1: 0.17; exact z-test $p = 0.080$). A probit regression finds a significant negative relationship between CRT score and the probability of choosing 0.5 ($p = 0.047$).

4.2 Urn Task

Our findings in the Urn Task are as follows. The distributions of beliefs are shown in Figure 5. As in the Stag Hunt game, in the urn task there is no statistical evidence of a treatment difference in distributions of beliefs (MW $p = 0.559$), and most subjects in the Karni treatment state a multiple of 0.05 (92%). The proportion of subjects stating 0.5 is again lower in the frequency treatment than the Karni treatment (0.21 and 0.33, respectively). The difference, however, is only weakly significant (exact z-test, $p = 0.087$). We note here that it is reasonable to expect a belief of 0.5 to be genuinely held by some subjects, as this is the belief that results from a failure to apply Bayes' rule, a commonly observed phenomenon (Tversky and Kahneman, 1974).

The response that maximizes expected payoff in the frequency treatment is to guess 17 balls out of 20 will be the same colour as the initial draw.

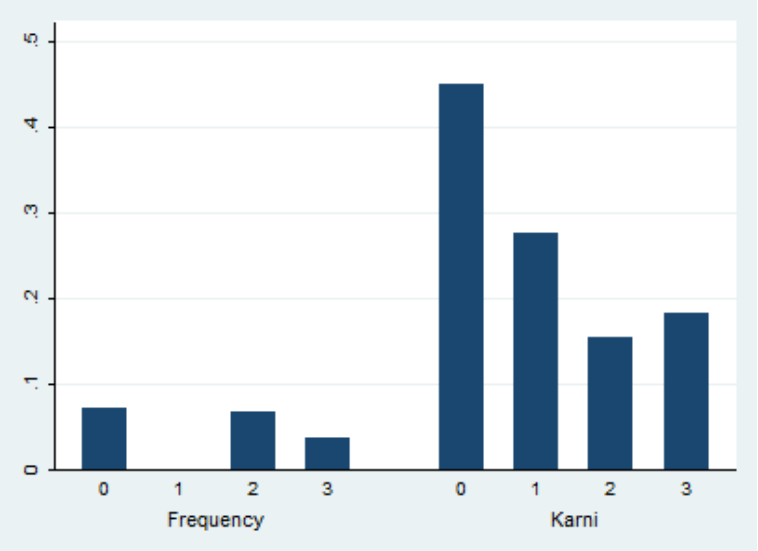


Figure 4: Proportion of subjects stating belief of 0.5 for the Stag Hunt game as a function of the number of correct answers in the Cognitive Reflection Test.

The best answer in the Karni mechanism is to state a probability of 0.82. We evaluate the accuracy of stated beliefs in two ways, first of all by comparing the distance from the correct answer, then the proportion who state (approximately) the correct answer.

For evaluating the distance from the correct answer, we first make the two methods comparable in the same way as for the histograms. We group the data from the Karni method into bins corresponding to probabilities consistent with each response in the frequency method and count the number of bins distant from the correct one. There is no statistically significant difference in these distributions of these differences (Frequency: 5.50; Karni: 5.64; MW $p = 0.567$). The proportion of subjects who state the correct answer of 17 balls in the frequency treatment is weakly greater than those in the Karni treatment who state a probability corresponding with that elicited interval, i.e. between $\frac{17}{21}$ and $\frac{18}{21}$ (Frequency: 0.08; Karni: 0.02; exact z-test $p = 0.060$). Allowing for a little more leeway, the proportion of subjects who state 16, 17, or 18 balls in the frequency treatment is again greater than those in the Karni treatment who state a probability between $\frac{16}{21}$ and $\frac{19}{21}$, but the difference is not statistically significant (Frequency: 0.37; Karni: 0.25; exact z-test $p = 0.119$). We emphasize that these results must be taken with a grain of salt for evaluating the accuracy of belief elicitation, because,

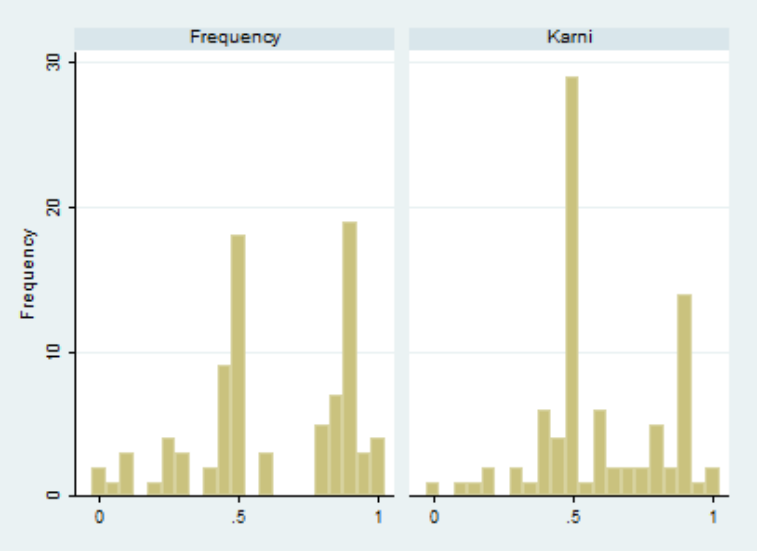


Figure 5: Distribution of beliefs about the colour(s) of the next ball(s) in the Urn Task.

as mentioned above, subjects beliefs may not be correct due to failures of Bayesian updating.

A similar negative relationship between number of correct responses in the CRT and stating 0.5 can be seen in Figure 6 for the Karni treatment but not for the frequency treatment, although we find no statistical support for this at conventional levels.

5 Quantile Elicitation

In the third part of the experiment we use a public goods game to investigate a new method for eliciting quantiles. Subjects make decisions in a two person public goods game, half with a marginal per-capita return (MPCR) of 0.65, and half with an MPCR of 0.9. To elicit beliefs about the median contribution we ask subjects to guess a number, and they will be paid €2 if that number lies between two randomly drawn contributions from other subjects in the session. Furthermore, we elicit upper quartiles by asking subjects to report a number that is higher than three randomly drawn contributions and lower than a fourth. Similarly, we elicit lower quartiles by asking for a number that is lower than three randomly drawn contributions and higher than a fourth. Details and proofs related to this method can be found in Schlag and Tremewan (2014). We also repeated the four comprehension questions for

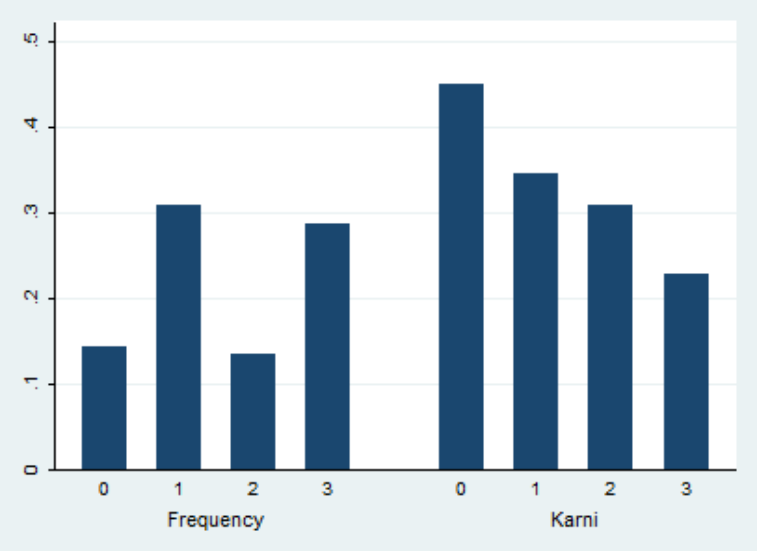


Figure 6: Proportion of subjects stating belief of 0.5 for the Urn Task as a function of the number of correct answers in the Cognitive Reflection Test.

each of the three elicitation tasks.

Average responses were in line with expectations: lower quartiles were lower than medians, which were lower than upper quartiles; all quantiles were higher for the higher MPCR. However, the difference between average LQ and average UQ was much smaller than the true interquartile range (4.2 compared to 10.5 when MPCR=0.65, and 3.6 compared to 13.5 when MPCR=0.9). More concerningly, 8% (24%) of subjects stated an UQ (LQ) lower (higher) than their stated belief about the median. This level on confusion was reflected in their responses to the question about how well they understood the task, which were closer to similar to those of the Karni method than the frequency method. As at least 24% of subjects did not respond coherently to the incentives of this method, we cannot recommend using this elicitation mechanism as implemented. It remains as an open question which method to choose for eliciting quantiles. One other contender is Qu (2012) who develops an extension of the Karni method for eliciting probability distributions.

6 Discussion and Conclusion

Reducing the complexity of instructions, and simplifying the communication of probabilistic information has not been a focus of the experimental eco-

nomics literature on belief elicitation (two exceptions are Hao and Houser (2012); Burfurd and Wilkening (2018)). Confusion and difficulties with processing probabilities without doubt increase noise and possibly introduce biases in responses. In light of this we suggest that an important route to improving the quality of belief elicitation is to better facilitate the understanding and communication of probabilities by subjects.

In our experiment, subjects reported better understanding of the frequency method than the Karni method. Three features that make the frequency method simpler for subjects are that it does not require mathematical formulae, can be explained with substantially less text, and crucially, it involves natural frequencies rather than numerical probabilities.

Probabilities can be expressed in a number of different ways: as a number, a percentage, or as a frequency. There is substantial evidence that even highly educated individuals often perceive mathematically equivalent probabilities as different when presented in the alternative formats. Lipkus et al. (2001) found that in a sample where 90% of respondents had at least some tertiary education, 40% were unable to convert a percentage to a frequency, while 79% were unable to convert a frequency to a percentage. Similar but more extreme results have been found for less educated respondents (Schwarz et al., 1997). Consequently, the format of probabilities has the potential to affect responses when eliciting beliefs.

There is evidence that people tend to be more comfortable and better able to process probabilities expressed as natural frequencies rather than other formats. Experiments by Kahneman and Tversky (1983) find that expressing probabilities as natural frequencies can mitigate the conjunction fallacy, while Gigerenzer and Hoffrage (1995) show that it also facilitates Bayesian reasoning. Cosmides and Tooby (1996) confirm the latter result and argue that human cognitive architecture has evolved to process natural frequencies rather than single-event probabilities in many situations. Schapira et al. (2001) report that participants in their study identify frequency formats as being intuitive and easy to interpret. To illustrate the primacy of natural frequency in probability related cognition the reader may try to explain the meaning of the statement "a fair coin will come up heads with probability 0.5" to someone not fluent in mathematics *without referring to natural frequencies*![7]

Although the role of the Karni treatment was simply to act as a benchmark, our experiment has revealed a serious bias in responses to this method.

[7]Another avenue we believe worth pursuing is the use of graphical aids. There has been a great deal of work on this in the fields of cognitive psychology and medical risk communication which could both complement and be complemented by experimental economics methodology. See, for example, references in Schapira et al. (2001).

Given the correlation we find between reporting 0.5 and cognitive ability, a reasonable interpretation of our results is that when subjects are confused, they simply choose the middle value. If this interpretation is correct, the bias is likely to occur in other complex elicitation methods, and is therefore worthy of further investigation. Depending on the reason for eliciting beliefs, this bias could lead to erroneous conclusions. For example, if an experimenter is eliciting subjects' beliefs about scoring above the median in a test, the observation that below average subjects consistently report that they are as likely to score above as below the median would be misinterpreted as overconfidence, rather than an artefact of the elicitation process.

This paper should not be read as a criticism of the Karni mechanism, as there may well be other ways of implementing the Karni method that reduce or eliminate this bias. However, our experiment shows that nice theoretical properties do not immediately translate into high quality data. Furthermore, the poor results from our elicitation of quantiles show that simple instructions, and simple mappings of events to payoffs, are not necessarily sufficient to obtain high-quality data in the domain of probabilistic beliefs, and all methods should be tested as thoroughly as possible. It is also clear from the spikes we see at 0.5 in the Karni treatment of our Stag Hunt game, and in both treatments for the Urn task, that when evaluating belief elicitation methods it is crucial to look at the entire distribution of elicited beliefs, rather than simply at population averages.

In this paper we have presented and characterized methods of belief elicitation which are extremely transparent to subjects and not dependent on restrictive assumptions about utility functions. The results of our experiment show that subjects understand this method better than a popular alternative, respond faster, and are less likely to choose a focal option. Simpler belief elicitation can give subjects more time and energy to focus on other tasks in an experiment with no apparent reduction in the quality of data. We encourage experimentalists to use this method in their own work, and especially to compare their empirical performance with other existing scoring rules.

References

Bhatt, M. and C. F. Camerer (2005). Self-referential thinking and equilibrium as states of mind in games: fMRI evidence. *Games and Economic Behavior 52*, 424–459.

Blanco, D. Engelmann, A. K. Koch, and H.-T. Norman (2010). Belief elicitation in experiments: is there a hedging problem? *Experimental Economics 13*(4), 412–438.

Burfurd, I. and T. Wilkening (2018). Experimental guidance for eliciting beliefs with the Stochastic Becker–DeGroot–Marschak mechanism. *Journal of the Economic Science Association 4*(1), 15–28.

Burfurd, I. and T. Wilkening (2020). Cognitive heterogeneity and complex belief elicitation. mimeo.

Charness, G., U. Gneezy, and V. Rasocha (2020). Experimental methods: Eliciting beliefs. Technical report, mimeo.

Cosmides, L. and J. Tooby (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement under uncertainty. *Cognition 58*, 1–73.

Costa-Gomes, M. A. and G. Weizsacker (2008). Stated beliefs and play in normal-form games. *The Review of Economic Studies 75*, 729–762.

Dal Bó, E., P. Dal Bó, and E. Eyster (2017). The demand for bad policy when voters underappreciate equilibrium effects. *The Review of Economic Studies 85*(2), 964–998.

Gigerenzer, G. and U. Hoffrage (1995). How to improve bayesian reasoning without instruction: Frequency formats. *Psychological Review 102*(4), 684–704.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association 1*(1), 114–125.

Hao, L. and D. Houser (2012). Belief elicitation in the presence of naïve respondents: An experimental study. *Journal of Risk and Uncertainty 44*(2), 161–180.

Hurley, T. M. and J. F. Shogren (2005). An experimental comparison of induced and elicited beliefs. *Journal of Risk and Uncertainty 30*(2), 169–188.

Kahneman, D. and A. Tversky (1983, October). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review 90*(4), 293–315.

Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica 77*(2), 603–606.

Le Coq, C., J. Tremewan, and A. K. Wagner (2015). On the effects of group identity in strategic environments. *European Economic Review 76*, 239–252.

Lipkus, I. M., G. Samsa, and B. K. Rimer (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making 21*, 37–44.

Manski, C. F. (2004, September). Measuring expectations. *Econometrica 72*(5), 1329–1376.

Offerman, T., J. Sonnemans, G. Van de Kuilen, and P. P. Wakker (2009). A truth serum for non-Bayesians: Correcting proper scoring rules for risk attitudes. *The Review of Economic Studies 76*(4), 1461–1489.

Powell, O. (2019). jtree - a javascript toolbox for running economics experiments. `https://opowell.github.io/jtree`.

Qu, X. (2012). A mechanism for eliciting a probability distribution. *Economics Letters 115*(3), 399–400.

Schapira, M. M., A. B. Nattinger, and C. A. McHorney (2001). Frequency or probability? A qualitative study of risk communication formats used in health care. *Medical Decision Making 21*, 459–467.

Schlag, K. H. (2008). A new method for constructing exact tests without making any assumptions. Department of Economics and Business Working Paper 1109, Universitat Pompeu Fabra.

Schlag, K. H. (2015). Who gives direction to statistical testing? Best practice meets mathematically correct tests. Mimeo, University of Vienna.

Schlag, K. H. and J. Tremewan (2014). Simple belief elicitation. Available at SSRN: 2449224.

Schlag, K. H., J. Tremewan, and J. J. Van der Weele (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics 18*(3), 457–490.

Schotter, A. and I. Trevino (2014). Belief elicitation in the laboratory. *Annual Review of Economics 6*(1), 103–128.

Schwarz, L. M., S. Woloshin, W. C. Black, and H. G. Welch (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine 127*, 966–971.

Suissa, S. and J. J. Shuster (1985). Exact unconditional sample sizes for the 2 times 2 binomial trial. *Journal of the Royal Statistical Society: Series A (General) 148*(4), 317–327.

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

Wilcox, N. T. and N. Feltovich (2000). Thinking like a game theorist: Comment. Mimeo, University of Houston.

Appendix A Proof

Proof. To prove the "only if" statement suppose b maximises $f(b)$. If $p_v = 0$ then it is clearly best if $b_v = 0$ because if $b_v > 0$ the prize will be won with probability 0. For any $u \neq v$ with $b_v, p_v > 0$,

$$\begin{aligned} & f(b_1, ..., b_u, ..., b_v, ...b_k) - f(b_1, ..., b_u + 1, ..., b_v - 1, ...b_k) && \geq 0 \\ \Rightarrow & \frac{n!}{b_1! \cdot .. \cdot b_k!} \prod p_i^{b_i} - \frac{b_v p_u}{(b_u + 1) p_v} \frac{n!}{b_1! \cdot .. \cdot b_k!} \prod p_i^{b_i} && \geq 0 \\ \Rightarrow & f(b) \left(1 - \frac{b_v p_u}{(b_u + 1) p_v}\right) && \geq 0 \end{aligned}$$

which gives us the set of constraints

$$b_v p_u \leq (b_u + 1) p_v \forall u \neq v. \tag{3}$$

Now $p_i = \sum_j \frac{b_j}{n} p_i = \frac{b_i}{n} p_i + \sum_{j \neq i} \frac{b_j}{n} p_i \leq \frac{b_i}{n} p_i + \sum_{j \neq i} \frac{b_i + 1}{n} p_j = \frac{b_i}{n} + \frac{1}{n}(1 - p_i)$ which implies

$$p_i \leq \frac{b_i + 1}{n + 1}. \tag{4}$$

Also, for $b_i > 0$, $p_i = 1 - \sum_{i \neq j} p_j \geq 1 - \sum_{j \neq i} \frac{(b_j + 1) p_i}{b_i} = 1 - \frac{p_i}{b_i}(n - b_i + k - 1)$, which implies

$$p_i \geq \frac{b_i}{n + k - 1}. \tag{5}$$

To prove the "if" statement assume that b satisfies (1). Consider any b' such that $f(b') > 0$, $b'_u > b_u$ and $b'_v < b_v$. Hence, $p_v > 0$. From the above equations above we obtain

$$\begin{aligned} & f(b'_1, ..., b'_u, ..., b'_v, ...b'_k) - f(b'_1, ..., b'_u + 1, ..., b'_v - 1, ...b'_k) \\ = & f(b') \left(1 - \frac{b'_v p_u}{(b'_u + 1) p_v}\right) \\ > & f(b') \left(1 - \frac{b_v p_u}{(b_u + 1) p_v}\right) > 0. \end{aligned}$$

This means that whenever we increase the report of event u by one and at the same time decrease the report of v by one then the probability of winning the prize goes down, provided the report of u was above b_u and the report of v was below b_v. Thus, for any given p we can compare $f(b)$ to any other $f(b')$, by repeating the above for all $u \in \{i : b'_i > b_i\}$ and $v \in \{i : b'_i < b_i\}$. This shows that b maximizes f over all $b' \in B$ which completes the proof. ■

Appendix B Instructions

Stag Hunt Game

In this part of the experiment you are matched with another participant. Both you and the other participant are reading the same instructions.

Both you and the other participant will have to choose between two options, "A" and "B" without communicating.

If you choose option A, you will receive 4 Euros no matter what the other participant chooses.

If you choose option B, you will receive 6 Euros if the other participant also chooses option B, and nothing if he/she chooses option A.

Which of the two Options do you prefer?

Belief Elicitation - Frequency

In this part of the experiment we will randomly select 20 participants from Part 1, excluding the participant you were matched with in that Part. How many of these participants do you think chose Option B. You will earn 1 Point if your guess is correct.

Remember: If a participant chose Option A, he/she would receive 10 Points no matter what the participant with whom they were matched chose. If a participant chose Option B, he/she would receive 15 Points if the participant with whom they were matched also chose Option B, and nothing if the participant with whom they were matched also chose Option A.

How many of the 20 randomly selected participants do you think chose Option B?

Belief Elicitation - Karni

In this part of the experiment the computer will randomly select a participant from Part 1, excluding yourself and the participant you were matched with in that Part. What is your belief about the chances that this participant chose Option B in Part 1 of this experiment?

Please state your belief in terms of a number between 0 and 100 (for example, 0 corresponds to no chance this participant choose Option B, 50 corresponds to equal chances this participant chose Option B vs Option A, and 100 corresponds to full certainty that this participant chose Option B).

In order to incentivize accurate reports of beliefs, you will be compensated according to the following scheme. This scheme makes it in your best interest to report your true belief about the likely choice. After you report a number

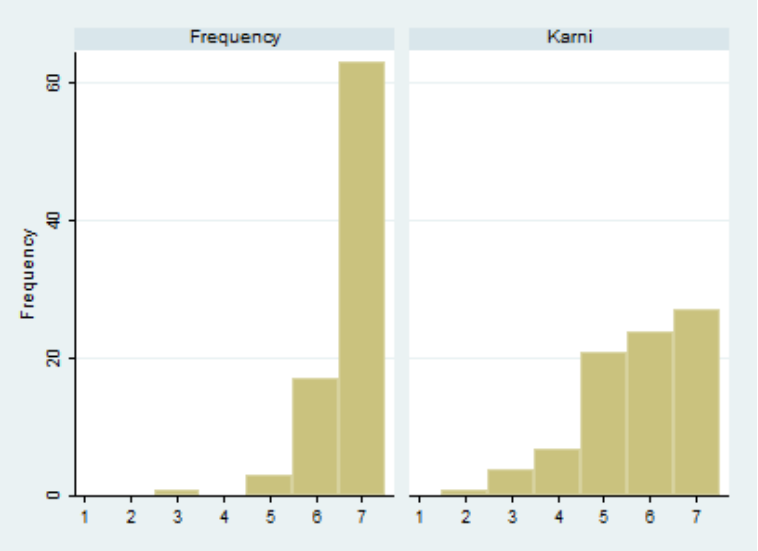


Figure 7: Distribution of answers to the question "How well do you feel you understood the task in part 2?" (1 = I did not understand at all, ..., 7 = I understood very well)

between 0 and 100, the computer will randomly choose a number between 0 and 100. If this number (call it n) is lower than the number you report, then you will be paid 2 Euros if the randomly selected participant chose Option B, and you will be paid nothing (0 Euros) if that participant chose Option A. If the random number n is greater than the number you reported, then you will earn 2 Euros with a chance of n% and nothing (0 Euros) with a chance of (100-n)%.

Remember: If a participant chose Option A, he/she would receive 10 Points no matter what the participant with whom they were matched chose. If a participant chose Option B, he/she would receive 15 Points if the participant with whom they were matched also chose Option B, and nothing if the participant with whom they were matched also chose Option A.

What is your belief about the chances that the randomly selected participant chose Option B in Part 1 of this experiment?

Appendix C Additional Results

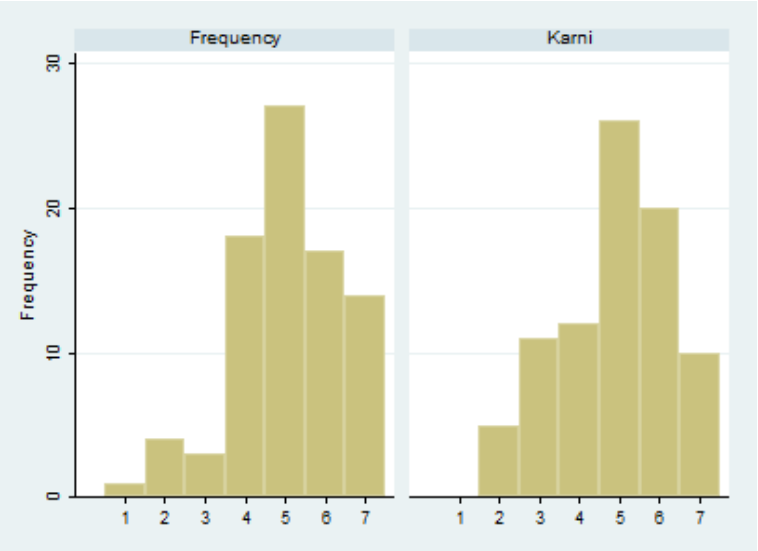


Figure 8: Distribution of answers to the question "How easy was it for you to come up with your answer to the task in part 2?" (1 = very difficult, ..., 7 = very easy)

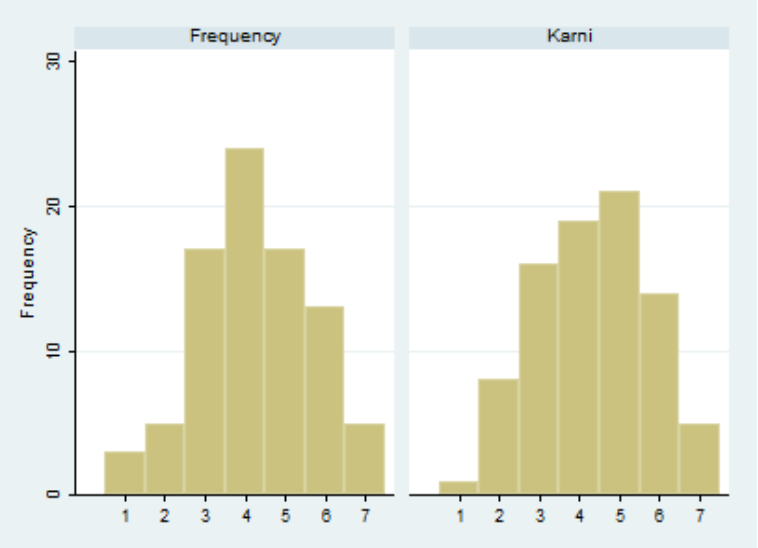


Figure 9: Distribution of answers to the question "How unsure or how confident are you that you gave the best answer?" (1 = very unsure, ..., 7 = very sure)

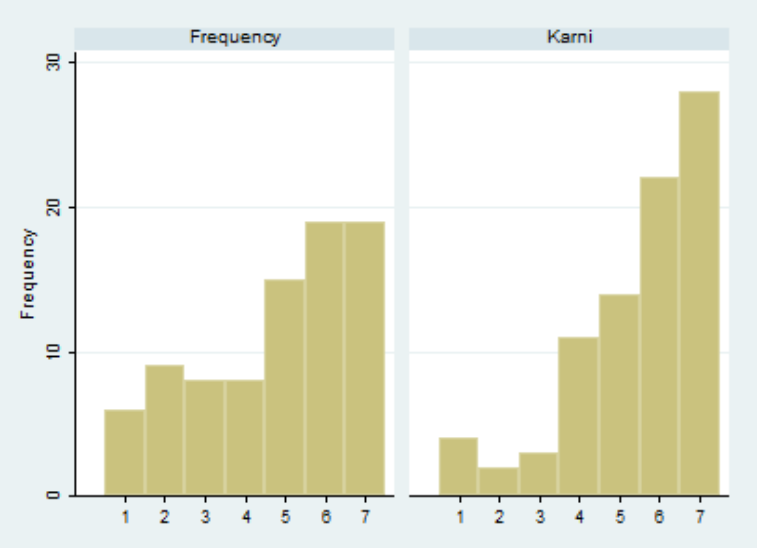


Figure 10: Distribution of answers to the question "When you chose between Options A and B, how important was it for you to think about how many participants (the chances that a randomly selected participant) would choose Option B?" (1 = Not important at all, ..., 7 = Very important)