# Comparing input interfaces to elicit belief distributions

Paolo Crosetto, Thomas de Haan

# Comparing input interfaces to elicit belief distributions

Paolo Crosetto and Thomas De Haan[21]

**Abstract**

We develop an intuitive, *Click-and-Drag* interface to elicit continuous belief distributions of any shape. We test this interface against the state of the art in the experimental literature – a text-based interface and multiple sliders – and in the online forecasting industry – a distribution-manipulation interface similar to the one used at Metaculus, a crowd-forecasting website. By means of a pre-registered experiment on Amazon Mechanical Turk we collect quantitative data on the convergence speed and accuracy of reported beliefs in a series of induced-value scenarios varying by granularity, shape, and time constraints. We also collect subjective data on ease of use, frustration and understanding. Results show that the click-and-drag interface outperforms all others by accuracy and speed, and is self-reported as being more intuitive and less frustrating than other interfaces, confirming our pre-registered hypothesis. Besides pre-registration, we report that the click-and-drag interface generates the least drop-out rate from the task, and scores best in a sentiment analysis of an open-ended general question. Further, we use the interfaces to collect homegrown beliefs on temperature in New York City in 2022 and 2042. On average, all subjects overshoot the real temperature for 2022 by about 2°F, and all anticipate further global warming in the order of 2.3°F; these forecasts are by and large not impacted by the interface used to elicit them. We provide a free and open source, ready to use oTree and Qualtrics plugin of our click-and-drag and all other tested interfaces available at https://beliefelicitation.github.io/.

*Keywords:* Belief elicitation, Forecasting, Scoring rules, interfaces

[1] `thomas.deHaan@uib.no` – University of Bergen, department of economics

[2] `paolo.crosetto@inrae.fr` – Univ. Grenoble Alpes, INRAE, CNRS, Grenoble INP, GAEL, 38000 Grenoble, France

[2] We would like to thank Ismaël Benslimane and Mu Numérique SAS for developing the click-and-drag interface, Aurélie Level for technical assistance, and colleagues at GAEL Grenoble for their insightful comments as well as Nikos Nikiforakis and participants at ASFEE 2022 Lyon and ESA Europe 2022 Bologna for comments. All remaining errors are ours.

## 1. Introduction

Eliciting beliefs is hard. Subjects might not entertain exact but only fuzzy beliefs on a specific event. They might have a vague idea, but be unable to provide a point estimate. If asked to provide a distribution, they might run into cognitive problems because they do not know what a belief distribution is, or how to report it. On top of all these cognitive and perception problems, lies the added problem of how the elicitation interface might facilitate or hinder the subject in expressing her intuitive belief distribution.

Eliciting beliefs and predictions has become a popular element of study across several fields and disciplines. Beliefs are key in experimental economics and psychology (Schotter and Trevino, 2014; Trautmann and van de Kuilen, 2015), in experimental research about asset markets (Haruvy et al., 2007), in macroeconomics, where expectations on key indicators play a crucial role and are elicited in recent experiments (Kryvtsov and Petersen, 2021; Armantier et al., 2016; Rholes and Petersen, 2021), in behavioral decision making, where the importance of graphical interfaces to elicit beliefs has been highlighted (Goldstein and Rothschild, 2014), in welfare economics to assess preferences for redistribution (Page and Goldstein, 2016), but also in marketing to assess priors of bayesian models (Sandor and Wedel, 2001; Delavande and Rohwedder, 2008), or in political science (Leemann et al., 2021). Belief elicitation and aggregation has also gained traction outside of academia, for instance in terms of asking professional forecasters, or 'crowds' for predictions, which is happening more and more online (e.g. https://www.predictit.org/ or https://www.metaculus.com/). The type of information elicited has also gotten more detailed. Rather than just asking for an average or a mean, either likely intervals (Schlag et al., 2015; Jain et al., 2013), modes, medians, or entire distributions (Harrison et al., 2017; Harrison and Phillips, 2014) are starting to get elicited in experiments.

There can be good reasons to choose to ask for a complete distribution when eliciting a forecast. It is possible that a policymaker is interested in more moments than just the mean of a distribution, or that there are expectations that the distribution for an indicator of interest might have multiple peaks. Another argument is that asking an entire distribution might actually be more intuitive to visualize for a forecaster than a derived measure such as the average or mode. For instance Kröger and Pierrot (2019) show that asking for a point prediction can create confusion when comparing with the complete underlying distribution.

The literature on the elicitation of beliefs is still in an early stage and features fundamental discussions on how to best ask participants about their predictions. One central discussion focuses on whether and how to incentivize participants to give their best belief estimates (Trautmann and van de Kuilen, 2015; Danz et al., 2022). A discussion missing so far in the economics literature, but present in the judgment and decision making literature (Goldstein and Rothschild, 2014) is the explicit testing of the interface used to elicit the belief distribution. If we are going to ask participants to fill in a distribution, which is not a simple task, then entering the distribution and being able to match this to the distribution they have in mind should be made as easy as possible. Frictions caused by the input interface, or frustration with interacting with the interface could create biases or inaccuracies.

We make two contributions. First, we present a first systematic comparison of several elicitation interfaces to test their performance. Second, we introduce a newly developed "Click-and-Drag" interface and test how it compares with other methods used so far for eliciting entire distributions.

We ask participants, recruited via Amazon Mechanical Turk to perform a mimic-the-distribution task. Subjects are shown a target distribution on one part of their screen and can enter a distribution on another part of the screen. Participants are asked to mimic several distributions, where the shape of the target distribution varies, as well as the time they have to work on the task (either 15 or 45 seconds). The participants are paid according to the euclidean distance between the target distribution and theirs, and are hence incentivized to mimic the target distribution as close as possible.

We run 4 treatments, corresponding to 4 different interfaces. The *Click-and-Drag* interface, where the distribution is determined by support points which the participants can create and place with the mouse. The *Slider* interface, where each bin has a slider which the participants can individually drag up or down via the mouse. The *Text* interface, where participants can numerically fill in the height of each bin in the distribution. And the *Distribution* interface, inspired by the one used on the forecasting

website metaculus.com, that starts out with an approximate normal distribution and provides 3 horizontal handles to adjust the mean, variance and skew.

The experiment was pre-registered at OSF. We find that, as pre-registered, the Click-and-Drag interface clearly outperforms all other interfaces in terms of performance and speed of convergence to a final answer. Furthermore we find that participants report the Click-and-Drag interface to be more intuitive and less frustrating than the other interfaces, including for example the Slider interface, which has been frequently used in experimental economics (Harrison et al. (2017), Harrison and Phillips (2014)).

The overall performance difference between the Click-and-Drag interface and the competition is clear. Only in one limited case does this interface get outperformed by the Distribution interface. This is when the distribution to mimic is random and quite erratic and the time given is short, 15 seconds. The Distribution interface appears here to benefit from starting from an initial normal distribution. The Distribution interface does poorly in all other instances. But it looks as though having a ready made reasonable/normal starting distribution helped for short time distribution assignments. This is a feature that could easily be added with Click-and-Drag – but feels like giving too much of a hint to subjects; starting from an empty distribution is instead free from default bias.

We also hope that this paper promotes the idea for experimental economists to more often explicitly test key features of their experimental interface. Experimental economists (and psychologists) often come up with original design solutions to test their hypotheses, however this might leave important features of the designs, which could be sources of bias and interference, untested.

## 2. The Click-and-drag belief interface

In creating Click-and-Drag, we had two main aims. First, we wanted an interface which could be understood with a few lines of instructions at most, or be picked up from practicing with a short tutorial or even no instructions at all. Second, we wanted to make an interface that scaled well, i.e. that was equally simple when creating a distribution over 5 or 50 bins, up to a (near) continuous distribution. One could imagine an oil price forecast where small differences matter for the world economy, yet the typical price variance and hence uncertainty of the future price is large.

Our solution to the aforementioned design challenges was to create an interface where one could draw a curve, and this curve would then dictate the shape of the underlying distribution. *Drawing a curve* is an intuitive exercise, and it effortlessly scales to any different number of bins by simply discretizing the continuous curve. To make this curve-drawing intuitive, we built on top of standard JavaScript libraries. In Click-and-Drag, subjects click to create a point, to which the curve is linearly attached. Points can be added, moved around, and deleted by simply clicking and dragging them. Figure 1 shows a screenshot of the interface in use; an interactive demo of the interface is available at https://beliefelicitation.github.io/.

Under the hood, the application uses the Otree platform (Python/Django framework, Chen et al. (2016)). The user interface is developed using the JavaScript libraries `Highcharts.js` and `jQuery`. The interface is inspired by the Highcharts demo tools "*Click to add a point*" and "*Basic column*". The sliders use the `noUiSlider` JavaScript library from *refreshless.com*. And the graphic elements are designed with the `Bootstrap` css library. The interface was developed by Mu Numérique SAS, Grenoble.

## 3. Experimental test

### *3.1. Task*

To test the different interfaces we set up a "mimic-the-distribution" task. In this task subjects were asked to recreate a target discrete probability distribution. The target distribution was depicted as a series of bins of varying height, shown in the top right corner of the screen. The rest of the screen was taken up by the belief elicitation interface, where subjects could enter the target distribution. The nearer subjects got to the target distribution, the higher their payoffs. Subjects were paid according to a score (between 0 and 100) which depended on the sum of all the differences between the bin heights of the targeted and entered distributions. Figure 2 shows a screenshot of the task, for the Click-and-Drag interface.

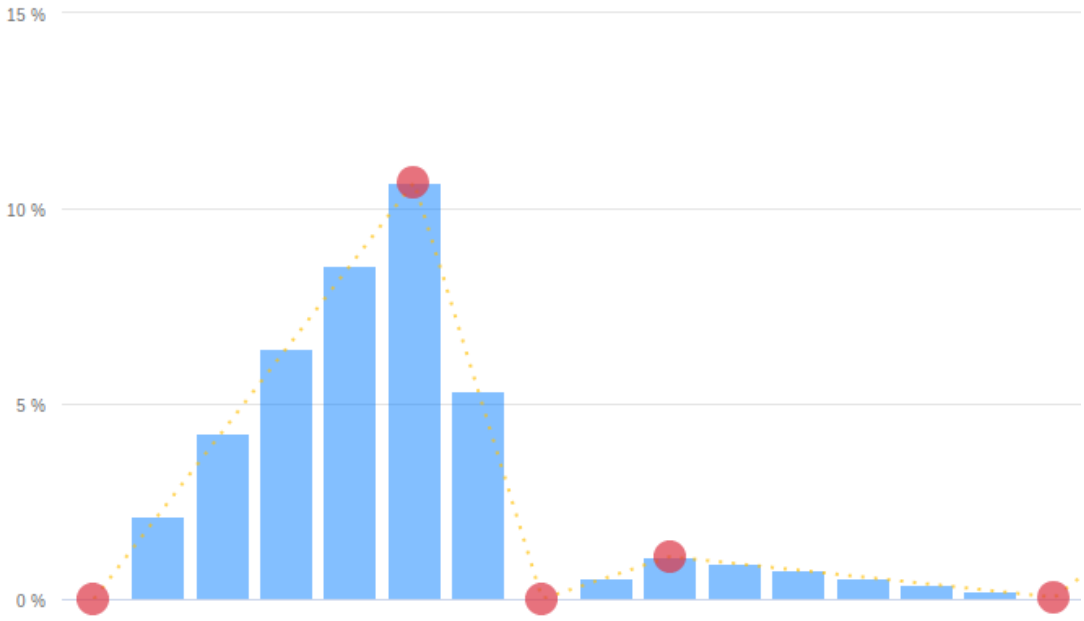


Figure 1: Screenshot of the Click-and-drag interface in action

To input their distribution, subjects could use one of four elicitation tools. Screenshots for all these tools are available in Appendix B.

**Click-and-Drag** As described in detail above.

**Text-based** This rudimentary elicitation interface has long been the norm in experimental economics, mainly for technical reasons. Subjects have to input a number for each of the discrete bins of the chosen support. The sum of the inputted numbers must be equal to 1 (or, equivalently, 100). This interface has been the one most largely used in earlier experiments for its technical simplicity. It is usually limited at 2 to 5 bins, and only seldom used beyond this threshold for practical concerns. In our implementation, the normalization is carried out by the software and no "sum to 100" constraint is imposed on the subjects. One example of the use of the text-based method can be found in the Survey of Consumer Expectations of New York (FRBNY) in the United States where, inspired by Manski (Manski, 2004; Dominitz and Manski, 2004; De Bruin et al., 2011) the New York Federal Reserve decided to ask experts the probability that future inflation would lie between two certain percentage levels. There have been critiques to this questionnaire method, for example by David (2022) who argues this type of density elicitation performs worse than a simpler inflation directional question.

**Slider-based** This interface is an incremental improvement over the text-based interface. For each bin subjects can move a slider to increase or decrease the probability mass of the bin. Normalization to 1 (equivalently, 100) can be automatic or delegated to the subject. This interface has been used when the tools available to experimental economists have evolved, and among others by Harrison et al. (2017); Andreozzi et al. (2020); Harrison et al. (2022); Fairley et al. (2019). The Slider interface is also reminiscent of – though not identical to – the Distribution Builder interface (Goldstein and Rothschild, 2014), where subjects hit on a "+" button to visually add probability mass to any of a series of bins. In our implementation the normalization is carried out by the software.

**Distribution** The rise of on-line predictions markets and crowd-prediction websites, like, for instance, Metaculus or PredictIt, has created the need for intuitive tools to enter predicted values. For instance, Metaculus uses an interface based on a bounded bell-shaped curve, that is controlled with

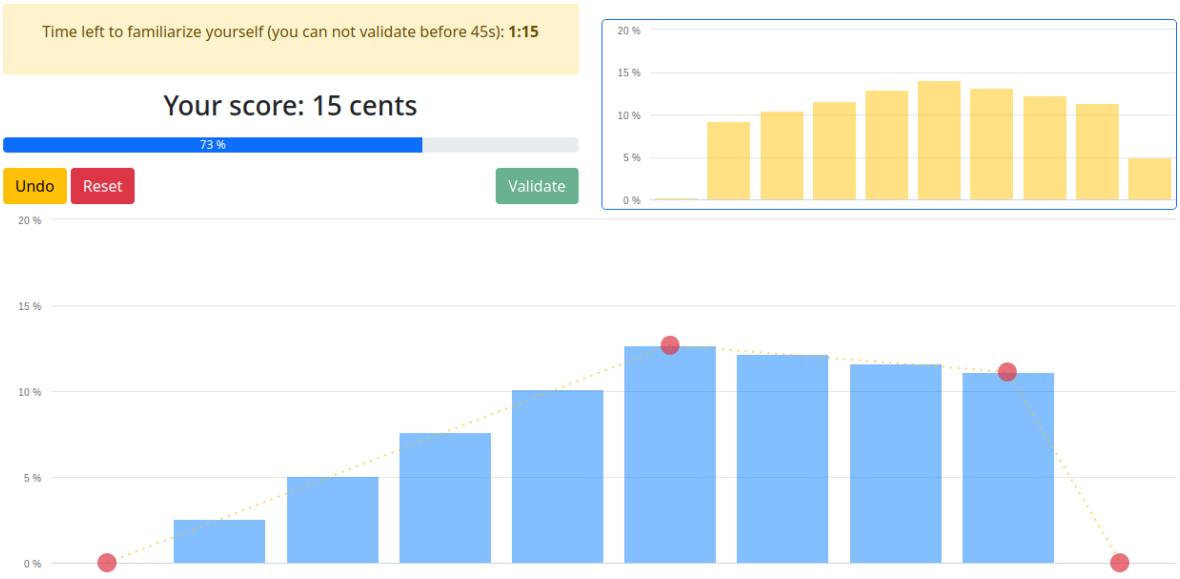


Figure 2: A screenshot of the main *mimic-the-distribution* task, for the Click-and-Drag interface

three handles. A central handle moves the distribution along the support without altering its shape. Left and right handles add mass on the left (or right) of the distribution, increasing in passing its dispersion. We reached out to Metaculus to ask for their code, but did receive a limited reply that made us unable to exactly replicate their interface. We hence created a Distribution elicitation interface based on a skew normal distribution, defined by three parameters: $\xi$ for the location, $\omega$ for the scale, and $\alpha$ for the shape. We translated these three parameters into the three-handles interface proposed by Metaculus assigning $\xi$ to the middle (location) handle, $\omega$ as the (normalized) distance between the left and right handles, and $\alpha$ as the difference between the distance of the left handle from the center and the distance of the right handle from the center. While we cannot be sure that this is exactly the function used by Metaculus, extensive testing shows us that our interface's behavior is qualitatively very similar to (a discretized version of) Metaculus' interface.

### *3.2. Treatments*

We tested the four interfaces between subjects. To assess the robustness of the interfaces to scale, difficulty, and time constraints we added three within-subjects variations.

**Target distribution shape.** We chose four different continuous shapes increasing in complexity. A truncated normal distribution, symmetric and single-peaked, is the baseline shape. We then added skew; made it bimodal; and finally added random noise to the asymmetric, bimodal distribution. Some shapes are hence harder than others – i.e. they require more tinkering with the interface – to reproduce.

**Level of discretization.** The four shapes are then discretized over 7, 15 or 30 bins. This is to assess the ability of the elicitation tool to easily scale to a finer granularity.

**Time constraint.** Subjects are given 45 or 15 seconds to complete the task. They have to spend the full time constraint on the screen – i.e., they are not allowed to submit the distribution before the time is up. This is done to measure how the different interface's performance scales when subjects are forced to reply in limited time.

Subjects faced 4 shapes × 3 discretizations × 2 times constraints, for a total of 24 screens. The screens were presented in a fixed order, in increasing level of difficulty. Screens started at 45 seconds, symmetric, 7 bins. We cycled through the different shapes, keeping the bins constant, and then through the number of bins, varying shape. The same sequence was run a second time, with all distributions mirrored, but this time with 15 seconds of allotted time. The target distributions for the first cycle of 12 screens are reported in Appendix C; the 12 screens for the second cycle featured the exact same distributions, but mirrored.

### *3.3. Measures*

We collect data on each click the subjects make. This allows us to derive two main quantitative measures of the performance of each interface.

**Accuracy.** We measure accuracy as the complement to 100% of the normalized distance of the final, submitted distribution from the target. Better interfaces have higher accuracy.

**Adjustment path.** We measure accuracy at each moment subjects interacted with the interface.[4] This allows us to see the path followed by the subjects to get nearer to the target, and hence to assess the speed of convergence to the final, submitted distribution. Some interfaces could allow subjects to quickly get the main strokes done, to then fine-tune the resulting distribution, while others might require to advance by smaller steps. Better interfaces allow for quicker convergence to the target distribution.

We also collect qualitative measures for each interface via three questions focusing on ease of use, frustration, time needed to understand the interface (on 1-to-7 Likert scales) and an open question asking for general comments. We further asked which device was used for input (keyboard, mouse, touchpad, touchscreen) as the interface performance can vary with input methods.

Finally, as a first test of the interfaces to elicit homegrown beliefs, we asked subjects to report their beliefs about the maximum daily temperature in New York City for the $4^{th}$ of July 2022 – a date that was in the near future for the experimental subjects – and for the $4^{th}$ of July 2042, 20 years in the future. We did this to have a first test of the Click-and-Drag interfaces with homegrown beliefs rather than induced values. The aim is to assess whether there are measurable biases introduced by the interfaces – i.e. whether the aggregated beliefs elicited with one interface systematically vary with respect to the same beliefs elicited with another. We do this using temperatures, that is something on which any subject can have at least a fuzzy distribution in mind, and over a 20 years period to assess whether subjects predict an increase in temperature related to global warming.

### *3.4. Sample and session details*

We implement the experiment using oTree, and run it on Amazon Mechanical Turk. The oTree application was developed by Mu Numérique SAS, Grenoble. We pre-registered and aimed for 360 Mturkers, 90 per each elicitation interface. Each subject was required to be using a PC (as opposed to a phone or tablet), and had to go through 24 screens, for a grand total of 8640 mimic-the-distribution tasks.

Following common practice, we limited recruitment to subjects geolocated in the United States, having completed at least 500 HITs on Mechanical Turk, and having an approval rate of at least 95%. After instructions (available in Appendix E), subjects faced a playground where they could practice the interface and the task for up to 1 and a half minutes. In order to weed out bots, we then implemented a set of four control questions (see Appendix F). Subjects had up to three attempts to clear the control question screen. After the first and second attempts, they were given feedback on their answers and provided with the correct replies. Subjects that failed three times the control questions were excluded from the experiment, with no bonus. If the case bots were able to pass the control question screen, they

[4] This is a click for all interfaces but Text, for which it is the moment a subject leaves one text field for the next one.

| | N | % female | mean age (sd) | mean payoff (sd) | % no error in CQ |
|---|---|---|---|---|---|
| Click-and-drag | 95 | 41% | 36.73 (9.69) | 2.92 (0.75) | 43% |
| Slider | 91 | 42% | 40.56 (10.93) | 2.35 (0.7) | 49% |
| Text | 91 | 48% | 37.07 (10.94) | 2.17 (0.89) | 46% |
| Distribution | 95 | 37% | 37.11 (11.17) | 2.23 (0.43) | 37% |

Table 1: Mechanical Turk Sample: demographics and final payoffs, by treatment

were usually unable to interact with the elicitation interfaces, since those are coded in Javascript, with quite custom code, and are much harder for bots to interact with than HTML fields. We labeled as bots and dropped from the sample all subjects not interacting at all on any screen, and thus earning zero.

We pay all screens. For each screen, recreating exactly the target distribution is worth 20 US$ cents; this means that each increase in 5% of the accuracy of the inputted distribution is worth 1 US$ cent. The maximum theoretical payoff is of 4.8 dollars, plus a 50cent fixed bonus. This makes our experiment reasonably well paid for Amazon Mechanical Turkers.

We ran the experiment over three sessions, a first session with 40 subjects to test for eventual bugs and problems (there were none), a session with what we thought would be the right number of subjects (we had to set the number of invited MTurkers taking into account an estimation of the incidence of bots) and a last session once we realized that we had more bots than anticipated, and had to fill in the missing real subjects. All sessions ran from the $20^{th}$ to the $22^{nd}$ of June 2022.

## 4. Confirmatory results

We pre-registered our main hypotheses on OSF, at https://osf.io/ft3s6. All data and analysis script to reproduce all the results in this paper are available at the OSF page of this project or in the dedicated github repository.

We hypothesized that the Click-and-Drag interface would outperform the other three tested interface in terms of accuracy, robustness to increased number of bins, different distribution shapes, and stricter time constraint, convergence in time, and that it would be self-reported as more intuitive and less frustrating. [5][6]

### *4.1. Sample*

Table 1 reports the demographics of the sample. As required, all subjects used a PC. We have slightly more usable data than pre-registered (372 vs. 360), slightly above the pre-set 90 per treatment for all treatments. The demographic distribution is not different across treatments by gender (Kruskal-Wallis test, p-value = 0.462), but it is different by age (Kruskal-Wallis test, p-value = 0.035), with the Slider treatment being significantly older than all others (Mann-Withney test, p-values: 0.015 *vs.* Click-and-drag, 0.018 *vs.* Text, 0.016 *vs.* Distribution), that are not different from each other (Mann-Withney test, all p-values $> 0.88$).

Subjects had to clear a screen of control questions to be able to perform the main task. They had three trials. After the first trial, they were given the correct answers. Between 40 and 50% of subjects

[5]We pre-registered an additional hypothesis about learning, positing that Click-and-drag would yield the best increase in performance from the first to the last time subjects saw a similar screen, an hypothesis to be tested via diff-in-diff. Unfortunately, this hypothesis was based on a previous iteration of the experimental design, based on 36 screens, where subjects would face the same screen twice, at the beginning and at the end of the experiment. For reasons of time and budget, we scrapped these screens, but we failed to update the pre-registration; we simply do not have the data to test this hypothesis, since each screen is different in one or more dimensions, and all comparisons of performance in the first and last screens would be confounded by the change in screen type (shape, time constraint, number of bins) and would not identify a learning effect.

[6]We further pre-registered a robustness check by input interface (mouse vs. touchpad vs. touchscreen) but the share of subjects using other input devices than the standard mouse + keyboard is so low that such an analysis is not warranted. See Appendix D for details.

cleared the control questions screen at their first trial. This share, a proxy for the average understanding in a treatment, is not significantly different across treatments (Kruskal-Wallis test, p-value = 0.353).

Subjects earned 2 to 3 euros on average, depending on the treatment; the payoffs are statistically different across treatments (Kruskal-Wallis test, p-value < 0.001). The difference is driven by Click-and-Drag, that generates significantly higher payoffs than all other interfaces (Mann-Withney, all p-values < 0.001). Slider and Distribution do not generate statistically different payoffs among each other (Mann-Witheny, p-value = 0.64). Slider does not generate higher payoffs than Text (Mann-Withney, p-value = 0.136), and Distribution yields higher payoff than Text (Mann-Withney, p-value = 0.026). The higher payoffs are an early indication that subjects reached higher accuracy with the Click-and-Drag interface with respect to all other interfaces.

### *4.2. Accuracy*

Table 2 gives an overview of the results of our experiment with respect to the accuracy of submitted final distributions across different interfaces, overall and by allotted time, number of bins and shape of the distributions to be mimicked.

| | Click-and-drag (N = 95) | Slider (N = 91) | Text (N = 91) | Distribution (N = 95) |
|---|---|---|---|---|
| **Overall** | | | | |
| | 60.64 [59.54,61.74] | 49.1 [47.79,50.41] | 42.79 [41.33,44.25] | 47.44 [46.56,48.32] |
| **by time constraint** | | | | |
| 45 seconds | 65.72 [64.2,67.24] | 59.96 [58.04,61.88] | 52.04 [49.82,54.26] | 48.77 [47.55,49.99] |
| 15 seconds | 55.43 [53.9,56.96] | 38.53 [36.99,40.07] | 33.42 [31.71,35.13] | 46.02 [44.74,47.3] |
| **by number of bins** | | | | |
| 7 bins | 68.53 [66.61,70.45] | 70.49 [68.51,72.47] | 64.45 [62.09,66.81] | 48.33 [46.84,49.82] |
| 15 bins | 61.59 [59.9,63.28] | 52.3 [50.34,54.26] | 43.33 [41.04,45.62] | 46.48 [45.04,47.92] |
| 30 bins | 51.79 [49.9,53.68] | 25.65 [24.18,27.12] | 19.98 [18.3,21.66] | 47.52 [45.88,49.16] |
| **by shape** | | | | |
| Symmetric | 70.28 [68.59,71.97] | 45.23 [42.49,47.97] | 40.78 [37.64,43.92] | 60.05 [58.1,62] |
| Skewed | 68.28 [66.46,70.1] | 48.21 [45.59,50.83] | 40.67 [37.79,43.55] | 56.45 [54.98,57.92] |
| Bimodal | 54.14 [51.75,56.53] | 50.43 [47.87,52.99] | 43.45 [40.68,46.22] | 37.44 [36.21,38.67] |
| Random | 49.97 [47.66,52.28] | 52.39 [49.83,54.95] | 46.26 [43.37,49.15] | 35.19 [34.17,36.21] |

Table 2: Final accuracy in percentage points, mean and 95% confidence interval, by condition for all interfaces

As pre-registered, the Click-and-Drag interface shows a better overall accuracy than all other interfaces (Mann-Withney, all p-values < 0.001). It is also notably robust to most variations. When moving from the 45 to the 15 seconds condition, it loses less accuracy when compared to Slider and Text (Mann-Withney, all p-values < 0.001), but more than Distribution, that is the only interface starting above zero, and yielding "average" results with minimal (or no) effort.[7] It is also remarkably robust to increasing the number of bins (lowest loss of performance when moving from 7 to 15 bins, and from 15 to 30 bins, against Slider and Text: Mann-Withney, all p-values < 0.001) – again with the exclusion of Distribution. When looking at the shapes, Click-and-Drag outperforms all others for Symmetric (Mann-Withney, all p-values < 0.001), Skewed (Mann-Withney, all p-values < 0.001) and Bimodal (Mann-Withney, all p-values < 0.03) target distribution shapes; it is not statistically different from Slider for messy, Random

[7] All interfaces start with an "empty" distribution (i.e. all bins set at zero), *but* the Distribution interface, that, being a discretized bell-shaped curve, cannot start at zero but starts from a symmetric, normal-like shape. The Distribution interface starts with a huge advantage, by design, and the time-constraint and dynamic adjustment path results are biased in its favor.

shapes (Mann-Withney, p-values = 0.44), but still better than Text and Distribution (Mann-Withney, all p-values $< 0.03$).

These results are driven mainly by the number of bins dimension, where the performance drop of Slider and Text is most notable. Table A.7, in Appendix Appendix A gives all the detailed results of each of the 24 screens for the interested reader.

### *4.3. Adjustment path*

Given the increasing importance of response times, and, more generally, *choice processes* in experimental economics (see for instance Spiliopoulos and Ortmann, 2017, for a review) the elicitation interfaces were designed to record the state of the distribution after each interaction with the interface – be it a click or entering a number. This allows us to track the *speed* with which subjects arrive at the final submitted distribution. We pre-registered that Click-and-Drag would be the faster tool, in the sense of being the tool with which subject take less time to cover most of the ground to the final, submitted distribution, while other tools would need more time. This is important both theoretically and practically. In theory, a tool that allows you to get near to the final answer with the first strokes is less likely to induce misreporting, or to be impacted by fatigue or carelessness, with respect to tools that need labor and effort to get where the subject wants. Practically, in applied settings belief elicitation might be done as a side task among many others, and the fact that subjects can quick sketch their beliefs is an important feature of an elicitation interface.

Figure 3 shows the accuracy of subjects, for all screens, separately for the 15 and 45 seconds conditions, in time. Click-and-Drag clearly shows a faster (i.e., in the plot, steeper) curve, especially in the first seconds. Note that Distribution here starts from a mechanical advantage, since the starting distance is *not* zero; still, its slope is the shallower of all the interfaces, indicating a slow advance towards the final submission. The result is even more clear in the 15 seconds condition, where by the $5^{th}$ second Click-and-drag subjects reach an accuracy of 20%, while Slider and Text subjects linger around 3%.

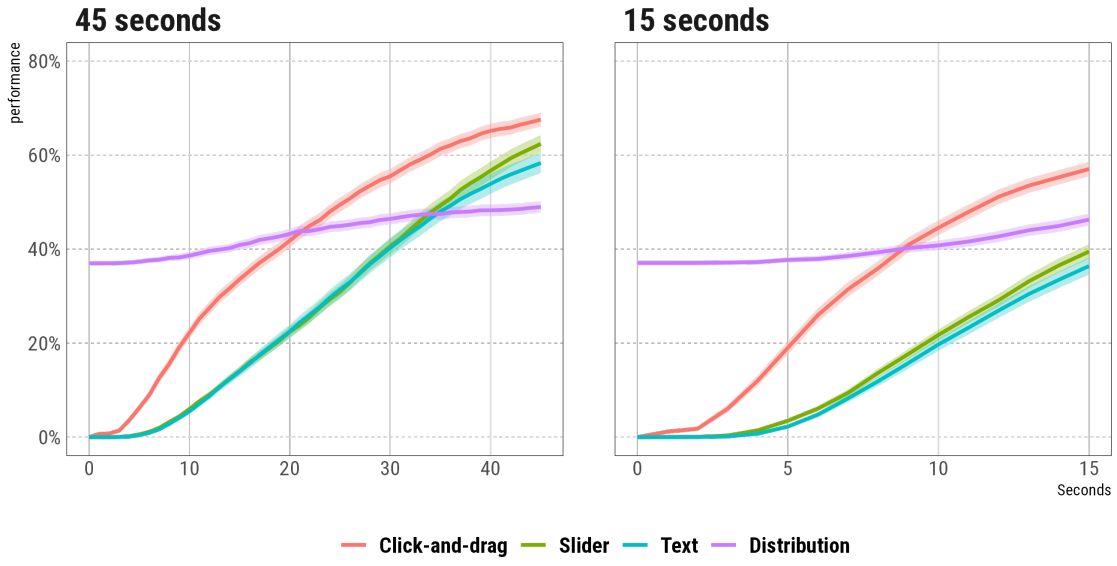


Figure 3: Performance dynamics by interface

The main result – Click-and-drag allows subject to converge faster to the target distribution, is replicated also when looking separately at screens with differing number of bins and differing shapes –

see Figures A.5 and A.6 in Appendix Appendix A.

### *4.4. Self-reported assessment*

Subjects faced three questions where they could self-report their qualitative assessment of the interface they had used in the task. All three questions were graded on a Likert scale from 1 to 7, and focused on ease of use, frustration, and ability to quickly understand. Table 3 reports the mean and 95% confidence interval of the subjects self-reported assessment on the three dimensions.

| | Click-and-drag | Slider | Text | Distribution |
|---|---|---|---|---|
| Hard to use | 3.77 [3.42,4.12] | 3.95 [3.57,4.33] | 4.16 [3.8,4.52] | 4.51 [4.18,4.84] |
| Frustrating | 3.59 [3.22,3.96] | 3.74 [3.35,4.13] | 4.24 [3.88,4.6] | 4.28 [3.93,4.63] |
| Difficult to understand | 3.48 [3.14,3.82] | 3.41 [3.06,3.76] | 3.43 [3.07,3.79] | 4.05 [3.75,4.35] |

Table 3: Likert scale (1 to 7) self-reported interface assessment, mean and 95% conf.int.

The Click-and-drag interface ranks first of all interfaces in ease of use and in generating least frustration, and third in the speed of understanding (indeed, it is easier to just input numbers, as there is likely nothing to learn there). Most results are not significant, though. Click-and-drag is perceived as less frustrating than Text (Mann-Withney, p-value $= 0.02$) and beats Distribution in all dimensions (Mann-Withney, all p-values $< 0.02$), the rest not being significant.

## 5. Exploratory results

This section includes analyses that were *not* pre-registered. This was due mainly to the fact that even in this age of replication crisis and pre-registration, discovery is still a thing, and some *ex-post* obviously important analyses did not occur to us *ex-ante*.

### *5.1. Slackers*

Elicitation interfaces can be (very) frustrating. This can lead subjects to drop out – i.e. not to finish the task, get distracted, and just let the time pass without collecting payoffs. Especially in the case of Mechanical Turk subjects in an online, unsupervised setting, subjects could just leave their browser tab open and do other things if the burden of fulfilling the task became heavier than the expected earnings. This could be particularly true for screens with 30 bins, and for the (in effect tedious) Slider and Text interfaces.

The number of persons dropping out on one or more screens – we call them *slackers* – can then be used as a proxy for the engaging (or frustrating) nature of the interface. We define subjects as having slacked on a screen if they had no or minimal interaction with the screen, and having improved the score from the starting point by less than 5 percentage points over the whole allotted time.

| | | Distribution of slackers by type | | | | |
|---|---|---|---|---|---|---|
| | Mean Slack | No | Somewhat (1-3) | Moderate (4-10) | Serious (11-15) | Severe (15-24) |
| Click-and-drag | 0.97 (1.46) | 45.26% | 49.47% | 5.26% | 0% | 0% |
| Slider | 1.54 (1.57) | 26.37% | 64.84% | 8.79% | 0% | 0% |
| Text | 4.07 (3.38) | 5.49% | 46.15% | 42.86% | 3.3% | 2.2% |
| Distribution | 11.36 (4.34) | 0% | 2.11% | 33.68% | 50.53% | 13.68% |

Table 4: Mean number of screens with no interactions, and distribution of slackers types by treatment

Table 4 reports the mean number of screns slacked, and the distribution of the number of screens (out of 24) on which a subject has slacked, by treatment. The number of slackers and the severity of

their disengagement from the task is severely affected by the treatments. The Click-and-drag interface proves to be the most engaging, with nearly half the subjects never slacking on any screen, and most of the other half slacking on only one to three screens. The Slider interface shows significantly more slacking, and the Text interface even more so. For both interfaces, it is indeed extremely frustrating to complete the task for the 15 and 30 bins targets, especially so for the Text interface. The Distribution interface shows a different pattern. In this interface, subjects do not start from zero, but from a roughly normal distribution. They are hence "spoiled" insofar as they start from a more advantageous point. Still, moving the sliders in order to approximate the given picture is sometimes hard, and subjects might have decided for a satisficing strategy keeping things "as is" as they got effortlessly "enough" money from the start. This might explain why *no* subject exerted effort on *all* screens; and the majority of subjects is classified as a serious slacker.

Click-and-Drag features the lowest mean amount of screens with no interactions (Mann-Withney, all p-values $< 0.001$), and thus proves to be, on an online sample of a possibly scarcely motivated population like MTukers, the most engaging interface, generating the least fatigue and drop-out rates.

### *5.2. Robustness of final results to slackers*

Given the above results on slackers, it is unclear whether the advantage of the Click-and-drag interface is limited to avoiding slackers. Slackers perform rather badly, and bias the mean performance of the affected interfaces downwards. It is possible that non-slackers reach a similar performance across all interfaces. To check whether the results are robust to focusing only on subjects having devoted full effort on most screens, Table 5 replicates the analysis of Table 2 limited to the subjects slacking on up to 3 screens out of the 24.

| | Click-and-drag (N = 90) | Slider (N = 83) | Text (N = 47) | Distribution (N = 2) |
|---|---|---|---|---|
| **Overall** | | | | |
| | 61.81 [60.72,62.9] | 50.7 [49.35,52.05] | 52.71 [50.7,54.72] | 67.12 [59.41,74.83] |
| **by time constraint** | | | | |
| 45 seconds | 67.19 [65.72,68.66] | 61.76 [59.83,63.69] | 63.95 [61.08,66.82] | 69.62 [57.96,81.28] |
| 15 seconds | 56.3 [54.76,57.84] | 39.69 [38.08,41.3] | 41.19 [38.75,43.63] | 64.62 [53.1,76.14] |
| **by number of bins** | | | | |
| 7 bins | 70.09 [68.21,71.97] | 72.19 [70.24,74.14] | 74.3 [71.4,77.2] | 70.62 [55.57,85.67] |
| 15 bins | 62.64 [60.97,64.31] | 54.06 [52.07,56.05] | 54.7 [51.71,57.69] | 62.9 [49.94,75.86] |
| 30 bins | 52.73 [50.85,54.61] | 26.72 [25.18,28.26] | 27.5 [24.9,30.1] | 68.88 [50.88,86.88] |
| **by shape** | | | | |
| Symmetric | 71.26 [69.65,72.87] | 47.11 [44.28,49.94] | 52.5 [48.09,56.91] | 80.43 [60.32,100.54] |
| Skewed | 69.16 [67.35,70.97] | 49.88 [47.19,52.57] | 50 [46,54] | 75 [71.97,78.03] |
| Bimodal | 55.57 [53.19,57.95] | 51.88 [49.24,54.52] | 52.99 [49.18,56.8] | 64.43 [51.48,77.38] |
| Random | 51.34 [49.02,53.66] | 53.82 [51.19,56.45] | 55.35 [51.47,59.23] | 46.83 [35.29,58.37] |

Table 5: Final performance, mean and 95% confidence interval, for subjects with limited slacking

When limiting the analysis to subject devoting near to full effort, Click-and-Drag loses a bit of its edge. It still outperforms Slider and Text in all the dimensions it did so before (Mann-Withney, all p-values $< 0.001$), but is not statistically different from Distribution (Mann-Withney, p-value $= 0.484$). Nonetheless, only two subjects are included for Distribution in this table, showing how hard it was for this interface to be used consistently – 98% of subjects slacked *more* than 3 screens when using this interface, and voiding of all meaning any statistical test. Nonetheless, the selection at play is clear: the 2 heavily self-selected subjects working hard on all screens had a performance non distinguishable from the one of the 90 (out 95) subjects devoting effort with Click-and-Drag. It takes motivation and dedication to reach good results with Distribution, and the frustration it generates discourage 98% of subjects from doing so.

*5.3. Sentiment analysis of the open-ended comments*

Subjects had the possibility of leaving an open-ended comment on their experience within the final questionnaire. This reply was not compulsory, but 149 subjects out of 372 did fill it in. The modal reply was a variation of "good" or "no problems", but some subjects made longer comments, voicing their frustration or showing appreciation for the task. We ran a sentiment analysis on the corpus of replies. Sentiment analysis is a text-mining technique using dictionaries that associate a valence (positive or negative) to any work in a dictionary, and then applying this valence to sentences. The sentiment analysis was run using the `R` package `syuzhet` (Jockers, 2015), that works by assigning an emotional value to each word in a text, in the form of a numerical value, that can be positive or negative and whose magnitude indicates the intensity of the emotion. A positive overall mean sentiment means that the message were more positive than negative.

Overall, sentiment over our experiment was positive. Nonetheless, the mean sentiment of the comments varied by interface. The Click-and-drag interface scored the highest mean sentiment, at 0.58 (st.dev. 0.59); followed by Slider (0.43, st.dev. 0.5), Text (0.4, st.dev. 0.41) and Distribution (0.36, st.dev. 0.47). Differences are not significant (Kruskall-Wallis test, p-value = 0.330) and in general all sentiments are positive; but the ranking of the sentiment analysis confirms the ranking of most of the other considered criteria.

## 6. The belief elicitation interfaces in action: predicting temperatures in NYC and climate change

On top of the pre-registered horse race of belief elicitation on the mimic-the-distribution task, we asked two *non-incentivized* direct belief elicitation questions. Subjects used the interface they had been exposed to to fill in this open belief question.

This was a first, unstructured, attempt at observing whether the interface does impact the reported beliefs. We chose to ask two questions related to maximum temperature in New York City on the $4^{th}$ of July of two given years – 2022, the year the experiment took place, and 2042, twenty years in the future. We chose this topic as one with which everyone is familiar, on which reasonable information is available to anyone, and to gauge average belief in an uncertain, but highly discussed, topic, namely climate change. Given the widely known topic, we could reasonably expect everyone of our subjects to hold a well-formed belief, and, with ∼90 subjects per treatment, we had a reasonable ex-ante expectation that the distribution of truly held beliefs would be roughly similar across treatments.

For each of the two questions, subjects could enter their belief using the interface they had been assigned to, over 31 bins, ranging from 60°F to 120°F, each bin representing a step of 2°F.

The target day was 14 days in the future at the moment of the sessions, and hence the temperature forecast for that day was unknown both to us and our subjects. But reliable, long-term weather data exist that place the target, expected maximum temperature for July $4^{th}$ in New York City to be of 83.6°F, with a 95% confidence interval of [82.3,84.9]°F. Actual maximum temperature on the $4^{th}$ of July 2022 in New York City (Central Park) turned out to be higher, at 85°F.[8]

The questions were not incentivized. This resulted in some subjects slacking – i.e., as detailed above, not interacting with the screen in any way or in a very limited way. We identify slackers as having moved the interface by less than 0.05 percentage points overall. Confirming the frustration generated by the Text interface, only 1 to 3% of subjects slacked in the Click-and-Drag and Slider treatments, but 7.6% (for the 2022 question) and 13% (for 2024) of subjects slacked in the Text interface. In the following, we report results *excluding* slackers.

The mean and 95% confidence interval of the elicited distributions, by interface, are detailed in Table 6. The mean shape of these distributions – i.e., the mean probability allotted to each bin in a treatment – and its 95% confidence interval are plotted in Figure 4.

[8] Average temperature was computed using the `R` `rnoaa` package, that allows to download accurate temperature data from the NOAA database; we downloaded the daily maximum temperature in Central Park, NYC, for all July $4^{th}$ starting from July 1900 up to July 2021.

| | Click-and-drag | Slider | Text | Distribution |
|---|---|---|---|---|
| 4th July 2022 | 87.69 [76.15,99.23] | 84.96 [57.61,112.3] | 85.83 [53.84,117.82] | 89.66 [72.81,106.52] |
| 4th July 2042 | 89.92 [78.73,101.11] | 87.07 [58.55,115.6] | 88.91 [52.43,125.4] | 91.69 [73.15,110.23] |

Table 6: Mean and 95% conf.int. of the elicited beliefs – maximum temperature on the $4^{th}$ of july

All beliefs overshoot the correct historic mean. Nonetheless, the temperature has been going up in recent years. While the overall secular average maximum temperature over 122 years is of 83.6°F, it is of 84.2°F since 2000. Overall hence, perceived temperatures are above real temperatures, even accounting for recent warming. Expected warming over 20 years is on average 2.3°F.

There is very little difference across elicitation interfaces in means. While point estimates do vary, aggregate beliefs do not significantly vary by interface, as confidence intervals are quite large. The aggregate beliefs elicited using the Click-and-Drag interface nonetheless show smaller variance, and have a smoother shape with respect to those entered using any other interface. Overall, the interface, as one would hope, seem *not* to radically affect the given beliefs in any particular way, if not for generating a smoother distribution.

## 7. Conclusion

We introduce Cick-and-Drag, a new belief elicitation interface, and test its performance against three other interfaces used in belief elicitation in the experimental literature or by a crowd-prediction website. We find considerable variance in performance of different interfaces across different task characteristics, such as time given, number of bins and shape of a target distribution. The elicitation interface appears to clearly matter.

We find that our Click-and-Drag interface overall outperforms the other three tested elicitation interfaces, and is the least impacted by changes in shape, number of bins, and time allotted to the task. Participants were able to closer mimic a target distribution within the allotted time of either 15 or 45 seconds. Especially in the case when the target distribution was one consisting of more than a few bins, Click-and-Drag very clearly outperformed both the Slider and Text input interfaces. The third competing interface, a Distribution 'shifting and stretching' interface, inspired by the interface of metaculus.com, tended to perform a little better (however overall still worse), but suffered from being found unintuitive and frustrating according to a participant questionnaire. Moreover, the Distribution interface generated the highest drop-out rate of all interfaces, being arguably harder to manipulate.

Our experiment, including the detailed data on participant performance over time, gives us a good insight into how people react to the different elicitation interfaces. One element to be possibly explored for future research is how changing the starting distribution affects the performance of the Click-and-Drag interface, as we saw that the interface where we provided participants with a ready made normal distribution, performed quite well under certain settings.

In the end our results suggest that researchers who plan to elicit belief distributions as part of their experiment, will likely benefit from adopting the Click-and-Drag interface. Moreover with an improved interface to elicit entire distributions, needing less time and feeling more intuitive, the choice to elicit a distribution, rather than just a mean or a mode, might become more attractive for both researchers and principals requiring a forecast.

We believe that the very existence of this test of the Click-and-drag interface can contribute to its adoption and robustness *on top* of the fact that the interface did indeed come ahead of the others. We know from hard data what we can reasonably expect of this interface, and this should make it easier for fellow researchers – or, indeed for the crowd-prediction industry – to safely adopt it for their belief elicitation studies.

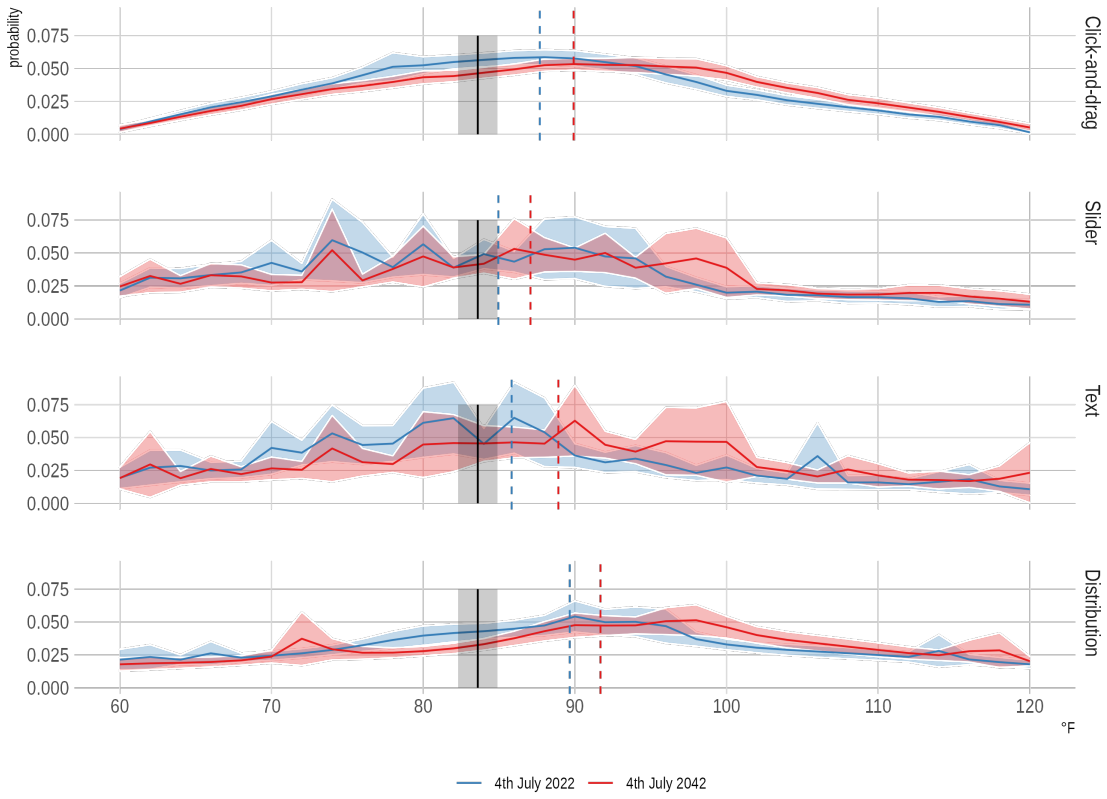


Figure 4: Distirbution of elicited beliefs on maximum temperature on the $4^{th}$ of July 2022 and 2042 in New York City, by treatment. Mean beliefs in color; true mean in black; 95% confidence interval shaded.


## Acknowledgments

Funding for this article was provided within the Priority Research Program **FAST** "*Facilitate public Action to exit from peSTicides*" financed by the French Agency for Research (ANR).


## References


Andreozzi, L., Ploner, M., Saral, A.S., 2020. The stability of conditional cooperation: beliefs alone cannot explain the decline of cooperation in social dilemmas. Scientific reports 10, 1–10.

Armantier, O., Nelson, S., Topa, G., Van der Klaauw, W., Zafar, B., 2016. The price is right: Updating inflation expectations in a randomized price information experiment. Review of Economics and Statistics 98, 503–523.

Chen, D.L., Schonger, M., Wickens, C., 2016. otree—an open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance 9, 88–97. URL: https://www.sciencedirect.com/science/article/pii/S2214635016000101, doi:https://doi.org/10.1016/j.jbef.2015.12.001.

Danz, D., Vesterlund, L., Wilson, A.J., 2022. Belief elicitation and behavioral incentive compatibility. American Economic Review .

David, C., 2022. Response bias in survey measures of expectations: Evidence from the survey of consumer expectations' inflation module. Working paper .

De Bruin, W.B., Manski, C.F., Topa, G., Van Der Klaauw, W., 2011. Measuring consumer uncertainty about future inflation. Journal of Applied Econometrics 26, 454–478.

Delavande, A., Rohwedder, S., 2008. Eliciting Subjective Probabilities in Internet Surveys. Public Opinion Quarterly 72, 866–891. URL: https://doi.org/10.1093/poq/nfn062, doi:10.1093/poq/nfn062, arXiv:https://academic.oup.com/poq/article-pdf/72/5/866/5188586/nfn062.pdf.

Dominitz, J., Manski, C.F., 2004. How should we measure consumer confidence? Journal of Economic Perspectives 18, 51–66.

Fairley, K., Parelman, J.M., Jones, M., Carter, R.M., 2019. Risky health choices and the balloon economic risk protocol. Journal of Economic Psychology 73, 15–33. URL: https://www.sciencedirect.com/science/article/pii/S0167487018302708, doi:https://doi.org/10.1016/j.joep.2019.04.005.

Goldstein, D.G., Rothschild, D., 2014. Lay understanding of probability distributions. Judgment and Decision making 9, 1.

Harrison, G.W., Hofmeyr, A., Kincaid, H., Monroe, B., Ross, D., Schneider, M., Swarthout, J.T., 2022. Subjective beliefs and economic preferences during the covid-19 pandemic. Experimental economics , 1–29.

Harrison, G.W., Martínez-Correa, J., Swarthout, J.T., Ulm, E.R., 2017. Scoring rules for subjective probability distributions. Journal of Economic Behavior & Organization 134, 430–448.

Harrison, G.W., Phillips, R.D., 2014. Subjective beliefs and statistical forecasts of financial risks: The chief risk officer project, in: Contemporary challenges in risk management. Springer, pp. 163–202.

Haruvy, E., Lahav, Y., Noussair, C.N., 2007. Traders' expectations in asset markets: Experimental evidence. American Economic Review 97, 1901–1920. URL: https://www.aeaweb.org/articles?id=10.1257/aer.97.5.1901, doi:10.1257/aer.97.5.1901.

Jain, K., Mukherjee, K., Bearden, J.N., Gaba, A., 2013. Unpacking the future: A nudge toward wider subjective confidence intervals. Management Science 59, 1970–1987.

Jockers, M.L., 2015. Syuzhet: Extract Sentiment and Plot Arcs from Text. URL: https://github.com/mjockers/syuzhet.

Kröger, S., Pierrot, T., 2019. What point of a distribution summarises point predictions? Technical Report. WZB Discussion Paper.

Kryvtsov, O., Petersen, L., 2021. Central bank communication that works: Lessons from lab experiments. Journal of Monetary Economics 117, 760–780. URL: https://EconPapers.repec.org/RePEc:eee:moneco:v:117:y:2021:i:c:p:760-780.

Leemann, L., Stoetzer, L.F., Traunmüller, R., 2021. Eliciting beliefs as distributions in online surveys. Political Analysis 29, 541–553. doi:10.1017/pan.2020.42.

Manski, C.F., 2004. Measuring expectations. Econometrica 72, 1329–1376.

of New York (FRBNY), ..F.R.B., 1999. Survey of Consumer Expectations. Technical Report. 2013-2020 Federal Reserve Bank of New York (FRBNY). URL: http://www.newyorkfed.org/microeconomics/sce.

Page, L., Goldstein, D.G., 2016. Subjective beliefs about the income distribution and preferences for redistribution. Social Choice and Welfare 47, 25–61.

Rholes, R., Petersen, L., 2021. Should central banks communicate uncertainty in their projections? Journal of Economic Behavior & Organization 183, 320–341. URL: https://www.sciencedirect.com/science/article/pii/S0167268120304157, doi:https://doi.org/10.1016/j.jebo.2020.11.013.

Sandor, Z., Wedel, M., 2001. Designing conjoint choice experiments using managers' prior beliefs. Journal of Marketing Research 38, 430–444. URL: https://doi.org/10.1509/jmkr.38.4.430.18904, doi:10.1509/jmkr.38.4.430.18904, arXiv:https://doi.org/10.1509/jmkr.38.4.430.18904.

Schlag, K.H., et al., 2015. A method to elicit beliefs as most likely intervals. Judgment and Decision Making 10, 456.

Schotter, A., Trevino, I., 2014. Belief elicitation in the laboratory. Annu. Rev. Econ. 6, 103–128.

Spiliopoulos, L., Ortmann, A., 2017. The BCD of response time analysis in experimental economics. Experimental Economics , 1–51URL: https://link.springer.com/article/10.1007/s10683-017-9528-1, doi:10.1007/s10683-017-9528-1.

Trautmann, S.T., van de Kuilen, G., 2015. Belief elicitation: A horse race among truth serums. The Economic Journal 125, 2116–2135.

## Appendix A. Detailed results by screen type

| | Click-and-drag | Slider | Text | Distribution |
|---|---|---|---|---|
| **45 seconds** | | | | |
| **7 bins** | | | | |
| Symmetric | 72.12 [66.16,78.08] | 74.41 [67.51,81.31] | 77.94 [71.62,84.26] | 61.99 [58.67,65.31] |
| Skewed | 78.07 [73.52,82.62] | 80.01 [74.39,85.63] | 73.88 [67.21,80.55] | 62.75 [59.82,65.68] |
| Bimodal | 69.12 [62.85,75.39] | 82.74 [77.4,88.08] | 72.97 [66.54,79.4] | 40.09 [37.37,42.81] |
| Random | 68.35 [62.05,74.65] | 84.79 [79.65,89.93] | 78.22 [71.99,84.45] | 30.74 [28.16,33.32] |
| **15 bins** | | | | |
| Symmetric | 73.45 [69.71,77.19] | 70.14 [65.1,75.18] | 59.02 [52.21,65.83] | 53.68 [48.86,58.5] |
| Skewed | 74.23 [70.52,77.94] | 59.08 [53.4,64.76] | 47.06 [40.15,53.97] | 56.14 [52.66,59.62] |
| Bimodal | 61.08 [55.56,66.6] | 62.95 [56.64,69.26] | 51.99 [44.44,59.54] | 39.08 [35.72,42.44] |
| Random | 61.79 [56.48,67.1] | 74.38 [69.61,79.15] | 63.26 [56.7,69.82] | 39.7 [37.31,42.09] |
| **30 bins** | | | | |
| Symmetric | 72.66 [70.6,74.72] | 36.45 [32.44,40.46] | 22.05 [16.31,27.79] | 63.48 [57.48,69.48] |
| Skewed | 63.01 [58.81,67.21] | 26.32 [20.79,31.85] | 17.85 [12.43,23.27] | 56.51 [52.82,60.2] |
| Bimodal | 52.8 [47.33,58.27] | 42.8 [38.09,47.51] | 35.8 [30.33,41.27] | 41.55 [38.07,45.03] |
| Random | 41.9 [37.55,46.25] | 33.6 [30.13,37.07] | 23.88 [18.66,29.1] | 38.89 [36.74,41.04] |
| **15 seconds** | | | | |
| **7 bins** | | | | |
| Symmetric | 70.08 [65.23,74.93] | 52.26 [46.64,57.88] | 49.39 [42.49,56.29] | 65.54 [62.59,68.49] |
| Skewed | 73.99 [69.71,78.27] | 65.47 [61.27,69.67] | 55.03 [48.86,61.2] | 59.63 [56.41,62.85] |
| Bimodal | 59.41 [54.53,64.29] | 60.51 [55.88,65.14] | 50.36 [44.54,56.18] | 35.65 [33.6,37.7] |
| Random | 57.08 [51.95,62.21] | 66.33 [62.31,70.35] | 58.56 [52.69,64.43] | 28.1 [25.59,30.61] |
| **15 bins** | | | | |
| Symmetric | 66.71 [63.61,69.81] | 33.53 [28.93,38.13] | 28.76 [23.28,34.24] | 56.8 [51.73,61.87] |
| Skewed | 63.36 [59.21,67.51] | 43.16 [39.23,47.09] | 30.97 [26.42,35.52] | 52.02 [47.93,56.11] |
| Bimodal | 42.82 [37.73,47.91] | 37.51 [34.47,40.55] | 31.64 [26.81,36.47] | 35.1 [32.42,37.78] |
| Random | 48.21 [44.68,51.74] | 38.19 [35.66,40.72] | 32.57 [28.92,36.22] | 37.78 [35.58,39.98] |
| **30 bins** | | | | |
| Symmetric | 66.39 [62.44,70.34] | 10.07 [7.49,12.65] | 7.25 [4.3,10.2] | 59.02 [53.18,64.86] |
| Skewed | 56.86 [52.36,61.36] | 20.26 [17.55,22.97] | 18.75 [14.8,22.7] | 51.1 [47.19,55.01] |
| Bimodal | 38.9 [33.12,44.68] | 17.53 [15.28,19.78] | 16.05 [13.18,18.92] | 32.65 [29.4,35.9] |
| Random | 23.03 [19.28,26.78] | 17.87 [16.05,19.69] | 17.56 [14.2,20.92] | 36.06 [33.86,38.26] |

Table A.7: Final performance, mean and 95% confidence interval, by condition for all interfaces

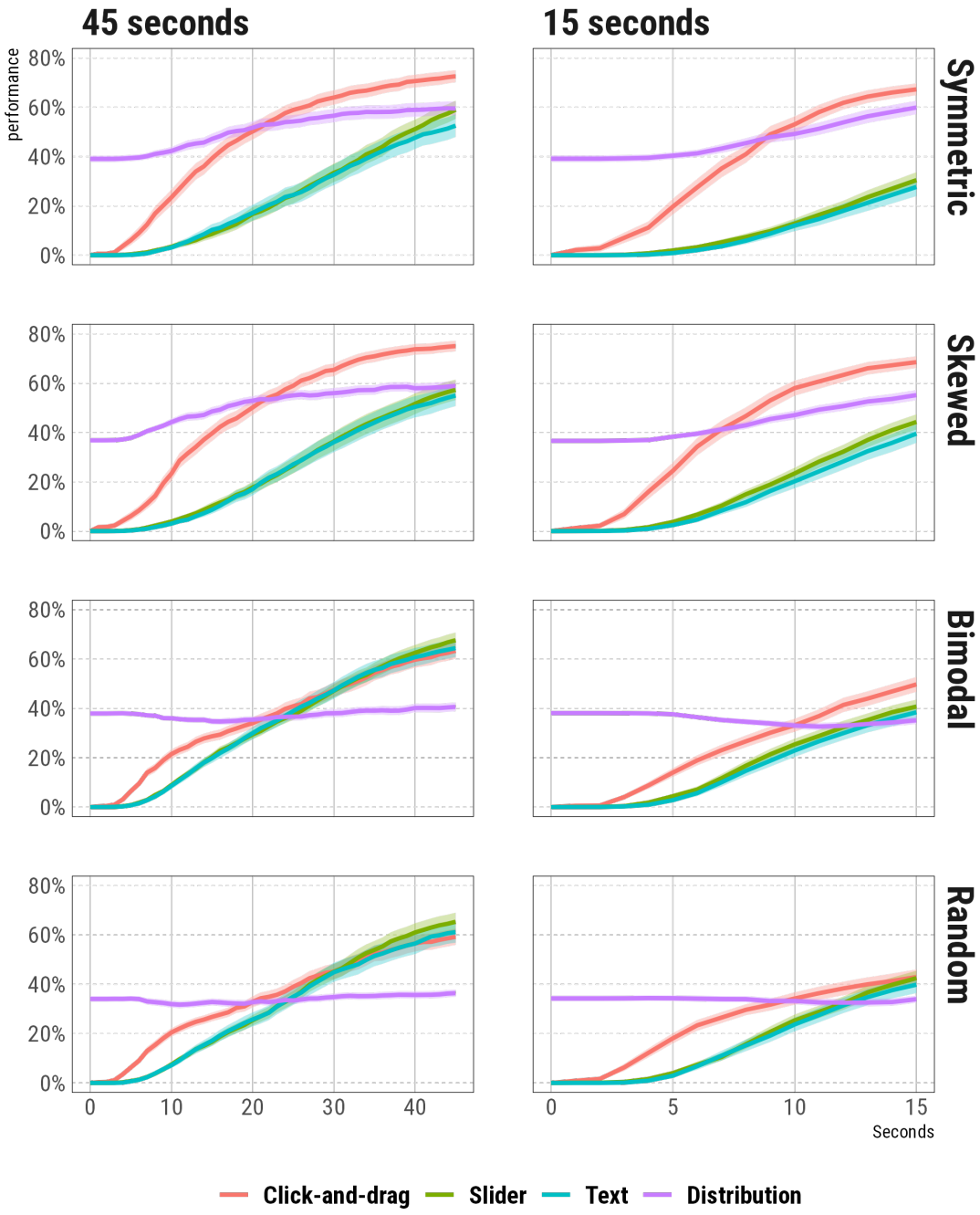


Figure A.5: Performance dynamics by interface – for different target shapes

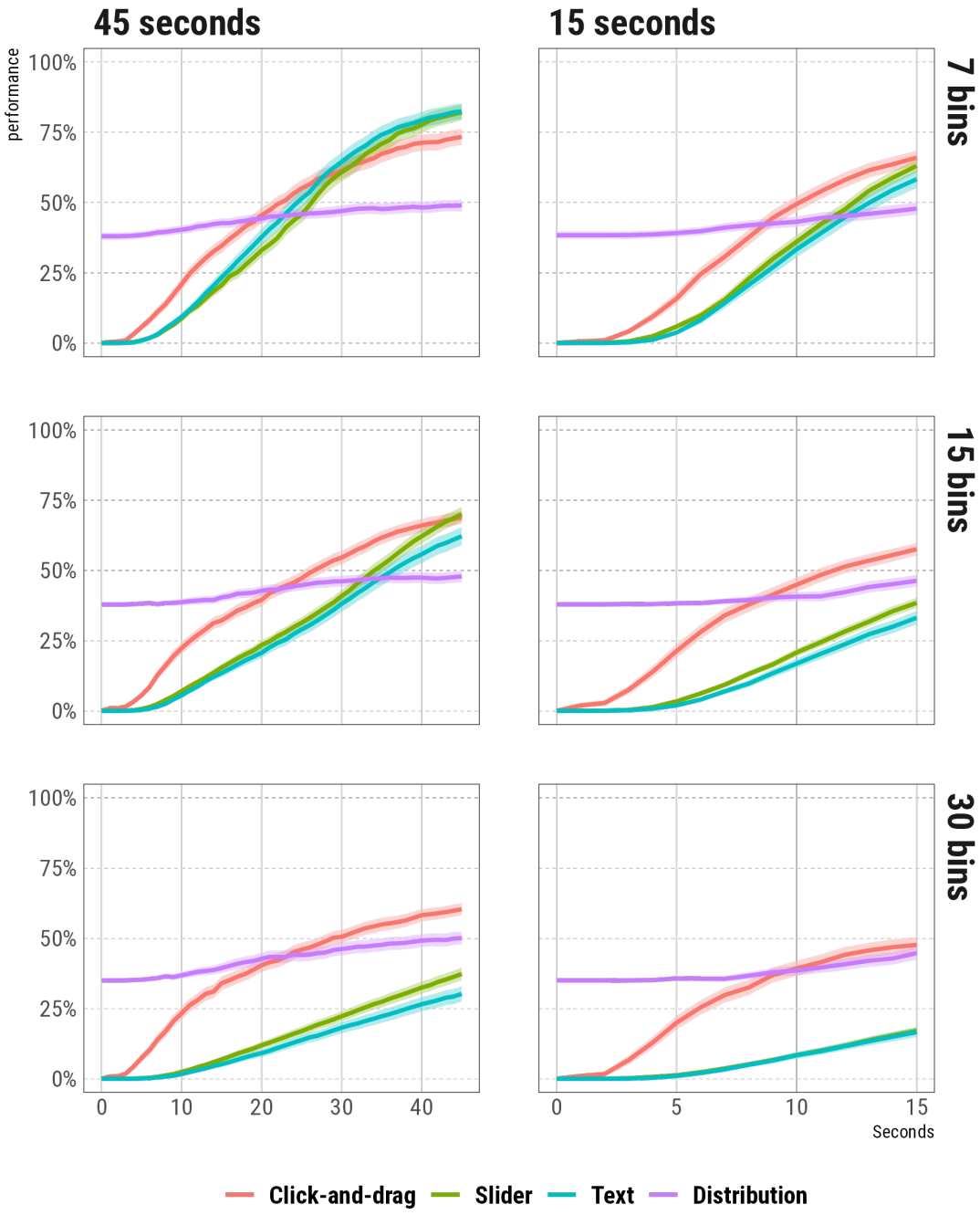


Figure A.6: Performance dynamics by interface – for different number of bins

## Appendix B. Screenshots


Figure B.7: Screenshot of the Text interface in action

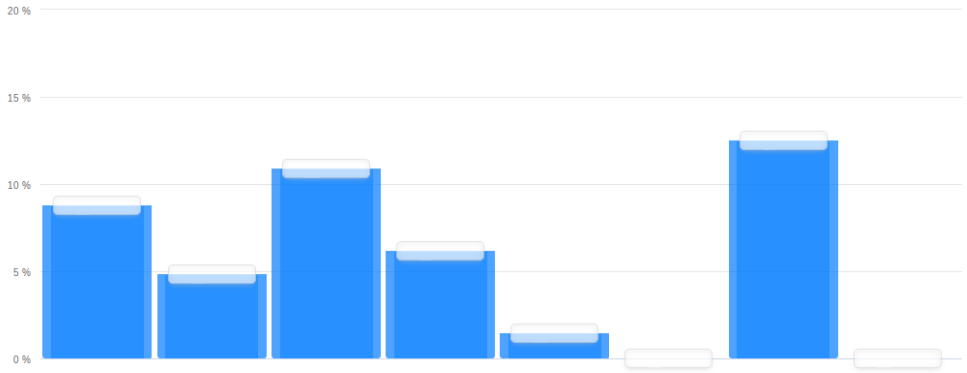


Figure B.8: Screenshot of the Slider interface in action

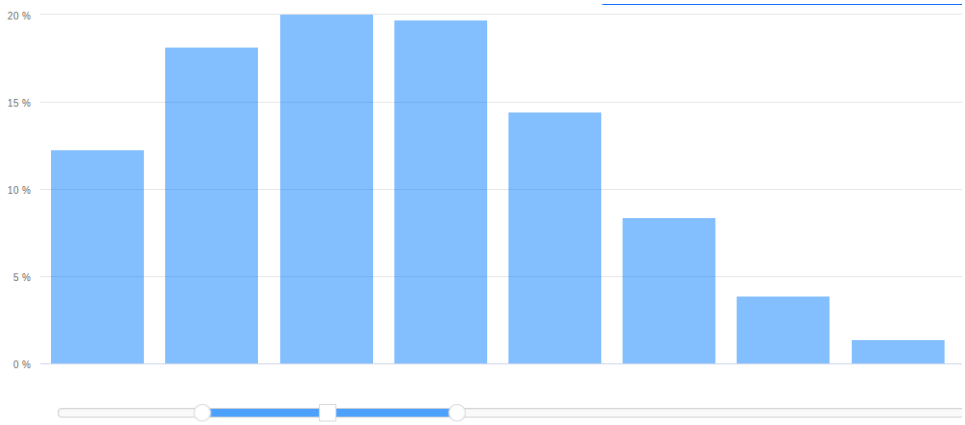


Figure B.9: Screenshot of the distDibution interface in action

## Appendix C. Target distributions

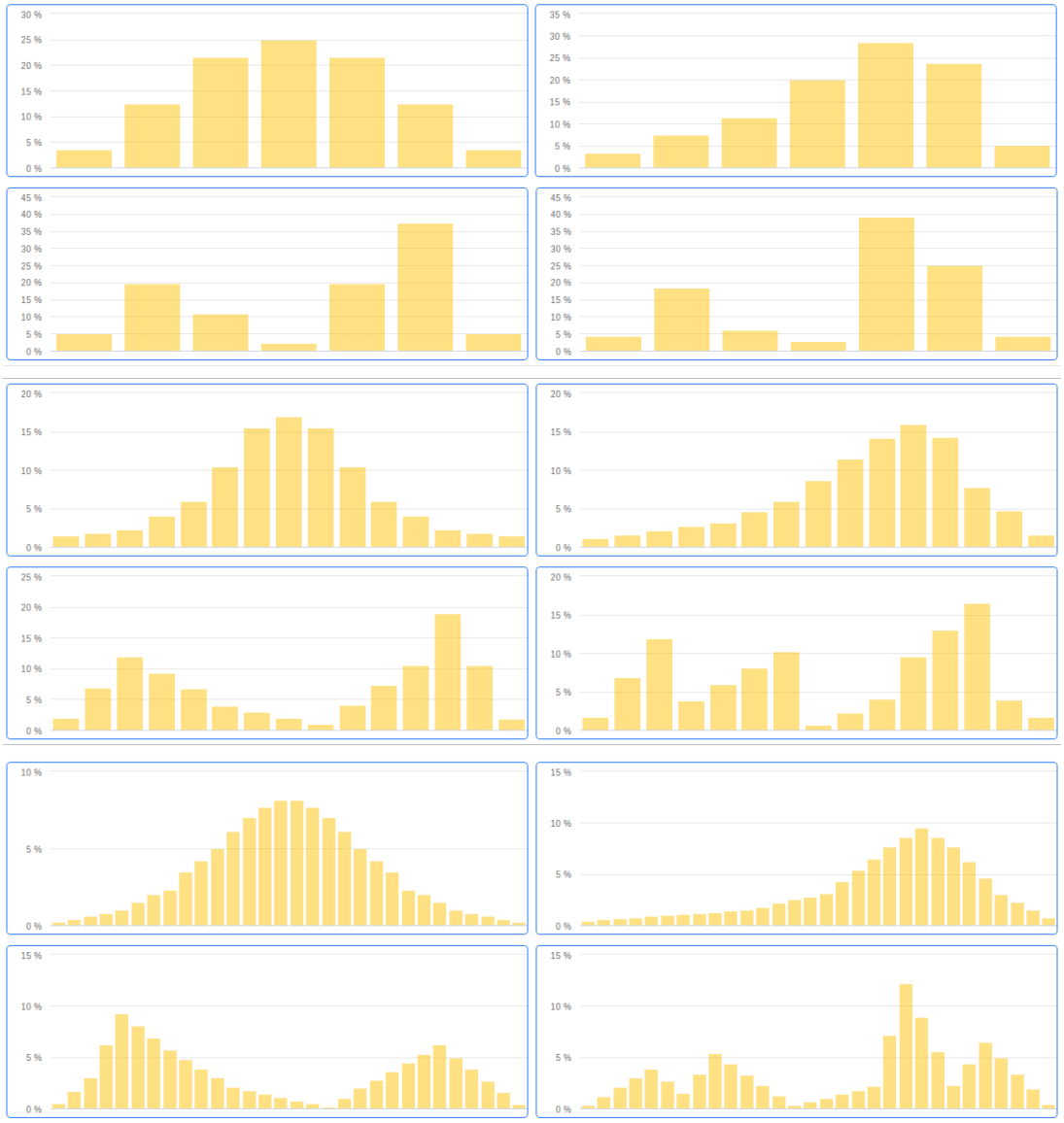

Figure C.10: The 12 target distributions

## Appendix D. Devices used by subjects

The experiemnt could technically be run by any browser, on any device. Nonetheless, for practical reasons linked to screen size, we required all participants to run the experiment on a PC – as opposed to tablet or phone. Still, subjects could do so using a variety of input devices – mouse, touchpad, keybard, touchscreen – and in principle the performance of the different interfaces is not orthogonal to the interface used to perform the task. Without a keyboard, the Text interface is unusable. With a touchpad (as opposed to a mouse), the Click-and-drag and Slider interfaces are harder to use. Table D.8 provides data on the system used by subjects and their input devices, by treatment.

| | Click-and-drag | Slider | Text | Distribution |
|---|---|---|---|---|
| Keyboard | 44.21% | 30.77% | 96.7% | 37.89% |
| Mouse | 98.95% | 94.51% | 82.42% | 97.89% |
| Touchpad | 9.47% | 10.99% | 13.19% | 12.63% |
| Touchscreen | 5.26% | 4.4% | 7.69% | 10.53% |

Table D.8: Input devices by treatment, share of subjects

Not surprisingly since 100% of the sample sat in front of a PC, most subjects used a mouse and a keyboard (when needed). The share of subjects that used *also* a touchpad or a touchscreen is low; but the share that used *just* those is so small (9 subjects using only a touchpad, 1 subject using only a touchscreen) as not to warrant a robustness analysis by type of input interface.

## Appendix E. Experimental instructions

*Common instructions*

Welcome to this research project! In this study, you will have the opportunity to earn money by working on a number of tasks.

*Procedures and Participation.* This study takes approximately 20 minutes and participation is voluntary. Please complete the task until the end. If you drop out of the task before finishing it you will get no reward. You are only allowed to participate in this study once.

*Confidentiality.* Information collected in this study is for academic purposes only and will be kept strictly anonymous.

*Payment.* If you complete this study, you will receive \$0.50 for your participation (HIT reward). According to your performance, you may earn additional money (bonus) during the study. The number of points for your bonus depends on how carefully you solve the tasks you will be asked to work on. You will be credited your Total reward (HIT reward + bonus reward) shortly after the completion of this study.

*Match the graph Task.* Read these instructions carefully. Your payment will depend on it. Moreover, You will later face control questions, and you will get no reward if you do not answer them correctly.

You will see in the top-right corner of your screen a small figure with a bar-graph in it. Your task is to recreate the same bar-graph, but then in a larger frame in the middle of the screen. You will receive money depending on how close the shape of your created bar-graph is to the target shown in the top-right corner of the screen.

How close your graphs is to the target picture is measured in percentage points, where you can attain a maximum score of 100%. Each screen is worth \$0.20. This means that if you reach 100%, you earn 20 USD cents; if 50%, 10 cents; if 10%, 2 cents. Each 5% increase in your score is worth 1 cent.

You will have to complete 24 graph-matching tasks. Your bonus is given by the sum of the amounts earned in each of the 24 tasks. A perfect score results in a bonus of 4.8 USD.

In the next screen, you can try the interface to familiarize yourself with the task. You can test the interface as you wish without affecting your score. Try to get the best score during the 90 seconds.

*Playground: familiarize yourself with the task (no bonus)*

*Click-and-Drag*

You can adjust the bar-graph by adding, moving or removing anchor-points: You can add anchor-points by clicking anywhere on the graph creates. You can move anchor-points around by dragging them. You can remove anchor-points by clicking on them.

*Slider*

You can adjust the bar-graph by dragging each bar up or down. Click on the top of the bar to drag it.

*Text*

You can adjust the bar-graph by entering a numerical bar height for each bar in the respective text field below the horizontal axis.

*Distribution*

You can adjust the bar-graph by adjusting the position of the horizontal slider buttons below the graph. You can add additional sliders to fine-tune the bar-graph.

## Appendix F. Control Questions

The four control questions were implemented as multiple choice questions. The available replies are in square brackets, in **bold** the correct answers. For control question 4, each answer was correct for one and only one interface, in this order: Click-and-drag, Slider, Text, Distribution.

1. How much is the fixed participation fee (HIT reward)? *[0.10$; 0.20$; **0.50$**; 1.00$]*
2. How much (bonus) can you earn on each of the 24 screens? *[0.10$; **0.20$**; 0.50$; 1.00$]*
3. The more your bar-graph matches the target picture, the lower your score. *[True, **False**]*
4. How do you adjust the bar-graph? *[by adding and moving anchor points; by dragging each individual bar up and down; by inputting the height of the bar in a text field; by moving the horizontal slider below the graph]*