Judgment and Decision Making (2023), **Vol. 18**:e27 1–25
doi:10.1017/jdm.2023.21

EMPIRICAL ARTICLE

Comparing input interfaces to elicit belief distributions

Paolo Crosetto [1] and Thomas de Haan [2]

[1]GAEL, Grenoble INP, CNRS, INRAE, Université Grenoble Alpes, Grenoble, France and [2]Department of Economics, University of Bergen, Bergen, Norway

Corresponding author: Paolo Crosetto; Email: paolo.crosetto@inrae.fr

Received: 3 October 2022; **Revised:** 21 May 2023; **Accepted:** 6 June 2023

Keywords: belief elicitation; forecasting; scoring rules; interfaces

Abstract

This paper introduces a new software interface to elicit belief distributions of any shape: *Click-and-Drag*. The interface was tested 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 by the most popular crowd-forecasting website. By means of a pre-registered experiment on Amazon Mechanical Turk, quantitative data on the accuracy of reported beliefs in a series of induced-value scenarios varying by granularity, shape, and time constraints, as well as subjective data on user experience were collected. Click-and-Drag outperformed all other interfaces by accuracy and speed, and was self-reported as being more intuitive and less frustrating, confirming the pre-registered hypothesis. Aside of the pre-registered results, Click-and-Drag generated the least drop-out rate from the task, and scored best in a sentiment analysis of an open-ended general question. Further, the interface was used to collect homegrown predictions on temperature in New York City in 2022 and 2042. Click-and-Drag elicited distributions were smoother with less idiosyncratic spikes. Free and open source, ready to use oTree, Qualtrics and Limesurvey plugins for Click-and-Drag, and all other tested interfaces are available at https://beliefelicitation.github.io/.

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. Connected to all these problems, is the challenge of how the elicitation interface facilitates or hinders 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 (Armantier et al., 2016; Kryvtsov and Petersen, 2021; 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 (Delavande and Rohwedder, 2008; Sandor and Wedel, 2001) and in political science (Leemann et al., 2021).

Belief elicitation and aggregation has also gained traction outside of academia. Increasingly popular websites ask professional forecasters, or 'crowds' for predictions (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 (Jain et al., 2013; Schlag and van der Weele, 2015), 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 with participants reporting modes when asked for means or medians.

There can be good reasons to choose to ask for a complete distribution when eliciting a forecast. A probability distribution provides a more complete picture of the uncertainty associated with a variable than a single point estimate such as the mean. It also includes information about the variance and skewness of a distribution, and whether there are one or multiple peaks of likelihood.

This could better enable a risk assessment, as this extra information a full distribution contains can be used to assess the consequences and level of risk associated with different decisions a policymaker might think of implementing. More generally, eliciting a full distribution fits better with Bayesian theory, which requires the specification of a full probability distribution for all uncertain variables. Eliciting a full prior and posterior distribution, for example, allows to investigate the Bayesian updating of an agent (Harrison et al., 2022).

Although most (experimental) applications of incentivized elicitation have been used to elicit beliefs over binary events, the theory on providing proper incentives has since early-on focused both on eliciting distributions with multiple discrete states (Savage, 1971), or even (discrete approximations) of full continuous distributions (Matheson and Winkler, 1976).

Another argument to favor eliciting complete distributions 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 with participants reporting modes when asked for means or medians.

The literature on the elicitation of beliefs is still young, and fundamental discussions on how best to ask participants about their predictions are ongoing. One central discussion focuses on whether and how to incentivize participants to give their best belief estimates (Danz et al., 2022; Trautmann and van de Kuilen, 2015). 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 2 contributions. First, we present a first systematic comparison of several elicitation interfaces, testing their performance. Second, we introduce a newly developed 'Click-and-Drag' interface and test how it compares with other methods 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 in another part. Participants are asked to reproduce as accurately as possible several distributions of varying shape in a very short to short time span of 15–45 seconds, and are paid according to the distance between the target and submitted distribution.

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

We ran 4 between-subjects treatments, corresponding to 4 different interfaces: *Click-and-Drag*, where the distribution is determined by support points which the participants can create and place with the mouse; *Slider*, where each bin has a slider which the participants can individually drag up or down using the mouse; *Text*, where participants can numerically fill in the height of each bin in the distribution; and *Distribution*, inspired by the interface used on the forecasting website metaculus.com, which 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, Click-and-Drag clearly outperforms all other interfaces in terms of both performance and speed of improvement of the distribution approximation. Furthermore, participants find Click-and-Drag more intuitive and less frustrating than the other interfaces, including Slider, which has been frequently used in experimental economics (Harrison et al., 2017; Harrison and Phillips, 2014).

The overall performance difference between Click-and-Drag and the competition is clear. Only in one limited case does this interface get outperformed by Distribution. 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, but performs poorly in all other instances, generating the highest level of frustration and drop-out among participants. Click-and-Drag can also easily be pre-set to start from a given distribution rather than from zero, but we think this could generate 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

Click-and-Drag was designed to fulfill 2 main goals. First, we wanted an interface which could be understood with minimal instructions and practice. Second, we wanted to make an interface that would scale well, that is, with which creating a distribution over 5 or 50 bins, up to a (near) continuous distribution would require nearly equal effort. One could imagine as an application for many bins an oil price forecast where small differences matter, yet the typical monthly price variance and hence uncertainty of the future oil price are large.

Our solution to these design challenges was to let subjects draw a curve that would then dictate the shape of the underlying bins distribution. *Drawing a curve* is an intuitive exercise, and it effortlessly scales to any different number of bins by simple discretization. To make this curve-drawing intuitive, we built on top of standard JavaScript libraries. With 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, Click-and-Drag is developed using the JavaScript libraries `Highcharts.js` and `jQuery`. 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.

We think the main advantage of Click-and-Drag is that it allows the user to quickly sketch the overall shape of a distribution with just a few clicks and drags, as opposed to manually move a number of bars or entering text, yet at the same time allowing the user to be as precise about the distribution as needed with further adjustments. This ‘sketch quickly, then refine’ strategy might be more similar to what people actually do when thinking about the probability of future events. A second advantage is to be mostly agnostic to the number of bins—entering a 10- or 100-bin distribution requires the same effort, without imposing any default, specific distributional shape, as opposed to, for example, an interface built on top of a given distribution.

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

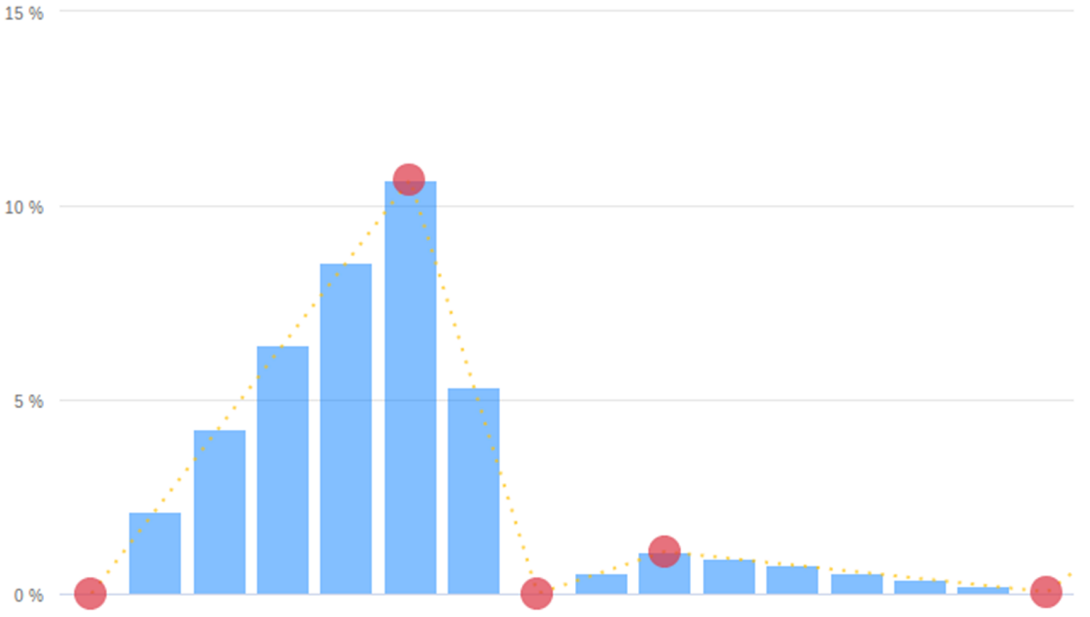

Figure 1. *Screenshot of the Click-and-Drag interface in action.*

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

3. Experimental test

3.1. Task

To test the different interfaces we set up a 'mimic-the-distribution' task. Subjects had to recreate a target discrete probability distribution, depicted as a series of bins of varying height, using 1 of the 4 belief elicitation interfaces. The nearer subjects got to the target distribution, the higher their payoffs. Figure 2 shows a screenshot of the task, for Click-and-Drag (screenshots for the other interfaces are shown in Figures B1–B3 in Appendix Appendix B).

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

To input their distribution, subjects could use 1 of the 4 interfaces.

Click-and-Drag As described in detail above.

Text This rudimentary elicitation interface has long been the norm in experimental economics, mainly thanks to its technical simplicity. 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). In practice, the interface is usually limited at 2–5 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 use outside of the laboratory is the U.S. Survey of Consumer Expectations Federal Reserve Bank of New York (FRBNY) (1999) where, inspired by Manski (De Bruin et al., 2011; Dominitz and Manski, 2004; Manski, 2004) the New York Federal Reserve asked experts the probability that future inflation would lie between 2 certain percentage levels. There have been critiques to this questionnaire method, for example, by Comerford (2022), who argues this type of density elicitation performs worse than a simpler inflation directional question.

Slider This is an incremental improvement over Text. Subjects can move a slider to increase or decrease the probability mass of each 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 Andreozzi et al. (2020); Fairley et al. (2019); Harrison et al. (2017, 2022). In our implementation, the normalization is carried out by the software.

Distribution The rise of online 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 3 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 received a limited reply that did not allow us to exactly replicate their interface. We hence created Distribution, based on a skew-normal distribution, defined by 3 parameters: ξ for the location, ω for the scale, and α for the shape. We translated these 3 parameters into the 3-handle interface proposed by Metaculus assigning ξ to the middle (location) handle, ω as the (normalized) distance between the left and right handles, and α 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.

For all interfaces but Distribution, the task started with the interface set at 0 for each bin. Distribution started with a wide, centered normal-looking distribution; this was necessary as there was no practical way in which we could have started from 0 but for leaving the subjects with no initial slider, and still show subjects what the interface could do.

3.2. Treatments

We tested the 4 interfaces between subjects. To assess the robustness of the interfaces to scale, difficulty, and time constraints we added 3 within-subjects variations.

Target distribution shape. We chose 4 different continuous shapes increasing in complexity. A truncated normal distribution, symmetric and single-peaked, is the baseline

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

shape. We then added skew; made it bimodal; and added random noise to the asymmetric, bimodal distribution. Some shapes are hence harder than others to reproduce.

Level of discretization. The 4 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—that is, they are not allowed to submit the distribution before the time is up. We do this to measure how the different interface's performance scales under time constraints.

Subjects faced 24 screens (4 shapes × 3 discretizations × 2 times constraints). 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 Figure C1 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 2 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. More in detail, let $p_i^{elicited}$ be the probability allocated to bin i by a participant and p_i^{target} the (normalized) probability allocated to bin i in the target distribution. The Accuracy score, given n bins in total, is then defined as

$$Accuracy = 100 \times \left(1 - \sum_{i=1}^{n} |p_i^{elicited} - p_i^{target}|\right).$$

Given this measure, a starting distribution of 0 probability in each bin gave an accuracy score of 0%. Better interfaces have higher accuracy and entering exactly the target interface would give a score of 100%.

Adjustment path. We record the accuracy at each moment subjects interacted with the interface.[1] 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 3 questions focusing on ease of use, frustration, time needed to understand the interface (on 1–7 Likert scales), and an open question asking for general comments. We further asked which device was used for input (keyboard, mouse, touchpad, and touchscreen) as the interface performance can vary with input methods.

Finally, as a first test of the interfaces to elicit homegrown beliefs rather than induced values, we asked subjects to report their beliefs about the maximum daily temperature in New York City for July 4, 2022—a date that was in the near future for the experimental subjects—and for July 4, 2042, 20 years in the future. We do this using temperatures, for which all subject should have at least a fuzzy distribution

[1] 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.

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

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 implemented the experiment using oTree (Chen et al., 2016), and ran it on Amazon Mechanical Turk. The oTree application was developed by Mu Numérique SAS, Grenoble, and is freely available at https://github.com/beliefelicitation/mturk. We pre-registered and aimed for 360 Mturkers, 90 per interface.[2] 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 8,640 completed 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 (Appendix F), subjects faced a playground to practice the interface and the task for up to 90 seconds. In order to weed out bots, we then asked 4 control questions (Appendix G). Subjects had up to 3 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 3 times the control questions were excluded from the experiment, with no bonus. In the case bots were able to pass the control question screen, they were usually unable to interact with the elicitation interfaces, since those are coded in JavaScript and as such much harder for bots usually designed for HTML fields. We labeled as bots and dropped from the sample all subjects not interacting at all on any screen, and thus earning 0.

Subjects got payed for every elicitation they completed. 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 50% fixed bonus. This makes our experiment reasonably well paid for Mechanical Turk.

We ran the experiment over 3 sessions. A first session with 40 subjects to test for eventual bugs and problems (there were none); a second one with the bulk of subjects; and a last session to fill treatments after weeding out bots. Sessions ran on June 20–22, 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 Click-and-Drag would outperform the other interfaces in terms of accuracy, robustness to increased number of bins, different distribution shapes, stricter time constraint, convergence in time, and that it would be self-reported as more intuitive and less frustrating.[3][4]

[2] Since no other paper ran a between-interface comparison, let alone with our task or with our new Click-and-Drag interface, we had no real benchmark to use to run sensible power computations. These would have been little more than wild guess, and we refrained from making any *ex ante*. A post hoc power analysis carried out using G*Power (Faul et al., 2007) shows that we have a 92% power of detecting a medium between-subjects effect and 99% power for a medium within-subjects effect.

[3] 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. Unfortunately, this hypothesis was based on a previous iteration of the experimental design, with 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 failed to update the pre-registration. We do not have the data to test this hypothesis.

[4] We further pre-registered a robustness check by input device (mouse vs. touchpad vs. touchscreen), but the share of subjects using other input devices than mouse + keyboard is too low to allow such an analysis. See Appendix D for details.

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

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

	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%

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 significantly different across treatments by gender ($\chi^2(3) = 2.58, p = 0.461$), and by age (ANOVA, $F(3, 368) = 2.61, p = 0.051$). When unpacking by treatment, Slider involved significantly older participants than all others ($t(179) = 2.53, p = 0.012$ and $d = 0.37$ *vs.* Click-and-Drag, $t(180) = 2.16, p = 0.032$ and $d = 0.32$ *vs.* Text, $t(184) = 2.13, p = 0.34$ and $d = 0.31$ *vs.* Distribution), that are in turn not statistically different from each other.

Subjects had to clear control questions to move on to the main task. They had 3 trials. After the first trial, they were given the correct answers. Between 40% and 50% of subjects cleared the control questions screen on their first trial. This share, a proxy for the average understanding in a treatment, is not significantly different across treatments ($\chi^2(3) = 3.27, p = 0.352$).

Subjects earned 2–3 Euro on average, depending on the treatment; the payoffs are statistically different across treatments (ANOVA, $F(3, 368) = 22.6, p < 0.001$). The difference is driven by Click-and-Drag, which generates significantly higher payoffs than all other interfaces ($t(149) = 7.89$, $p < 0.001$, $d = 1.14$ *vs* Distribution, $t(176) = 6.26$, $p < 0.001$, $d = 0.92$ *vs* Text, $t(184) = 5.44$, $p < 0.001$, $d = 0.80$ *vs* Slider). No other pairwise treatment comparison is significant.

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 target distributions.

As pre-registered, Click-and-Drag shows a better overall accuracy than all other interfaces ($t(4100) = 18.4$, $p < 0.001$, $d = 0.56$ *vs* Distribution, $t(3885) = 19.2$, $p < 0.001$, $d = 0.59$ *vs* Text, $t(4112) = 13.2$, $p < 0.001$, $d = 0.41$ *vs* Slider). It is also robust to most variations. When moving from 45 to 15 allotted seconds, it loses less accuracy when compared to Slider ($t(2043) = 9.71$, $p < 0.001$, $d = 0.43$) and Text ($t(1598) = 13.5$, $p < 0.001$, $d = 0.29$), but more than Distribution, that is the only interface starting above 0, and yielding 'average' results with minimal (or no) effort ($t(1920) = 7.47$, $p < 0.001$, $d = 0.32$). Click-and-Drag is also robust to increasing the number of bins. It shows the lowest loss of performance when moving from 7 to 15 bins (against Slider, $t(1320) = 7.17$, $p \leq 0.001$, $d = 0.39$; against Text, $t(1256) = 9.30$, $p \leq 0.001$, $d = 0.51$), and from 15 to 30 bins (against Slider, $t(1370) = 13.1$, $p \leq 0.001$, $d = 0.70$; against Text, $t(1297) = 9.11$, $p \leq 0.001$, $d = 0.50$). This is again with the exclusion of Distribution, that loses less than Click-and-Drag (7–15 bins, $t(1294) = 4.91$, $p \leq 0.001$, $d = 0.26$; 15–30 bins, $t(1242) = 8.9$, $p0.001$, $d = 0.47$). When looking at the shapes, Click-and-Drag outperforms all others for Symmetric (against Slider, $t(8.56) = 15.3$, $p < 0.001$, $d = 0.96$; against Text, $t(780) = 16.3$, $p < 0.001$, $d = 1.02$; against Distribution, $t(1031) = 7.80$, $p < 0.001$, $d = 0.48$), Skewed (against Slider, $t(928) = 12.3$, $p < 0.001$, $d = 0.76$; against Text, $t(880) = 15.9$, $p < 0.001$, $d = 0.98$; against Distribution, $t(1023) = 9.92$, $p < 0.001$, $d = 0.61$) and Bimodal (against Slider, $t(1067) = 2.07$, $p = 0.038$, $d = 0.12$; against Text, $t(1037) = 5.73$, $p < 0.001$, $d = 0.35$; against Distribution, $t(760) = 11.5$, $p < 0.001$, $d = 0.68$) target distribution shapes; it is not statistically

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

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

	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]

different from Slider for Random shapes, with a negligible effect size ($t(1063) = -1.38$, $p = 0.17$, $d = 0.08$), but still better than Text ($t(1001) = 1.97$, $p = 0.049$, $d = 0.13$) and Distribution ($t(760) = 11.5$, $p < 0.01$, $d = 0.68$).

The above results assume independence across trials, even within a single subject. This is likely a strong assumption. Testing results by first averaging over each subject and then running the tests does not change the picture. The results of those tests are reported in Table E1 in Appendix Appendix E. All results stay qualitatively the same.

These results are driven mainly by the variation of the number of bins, where the performance drop of Slider and Text is most notable. Table A1 in Appendix Appendix A gives detailed results for each of the 24 screens.

4.3. Adjustment path

Given the increasing importance of response times and *choice processes* in experimental economics (Spiliopoulos and Ortmann, 2017), we recorded the state of the distribution after each interaction with the interface. 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 faster than the competitors, in the sense of allowing subject to quickly cover most of the ground to the final, submitted distribution. 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. 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 enjoys a mechanical advantage, since the starting accuracy is *not* 0; still, its slope is the shallower of all the interfaces, indicating a slow advance toward the final submission. The result is even clearer in the 15 seconds condition, where after 5 seconds Click-and-drag subjects reach 20% accuracy, while Slider and Text linger around 3%.

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

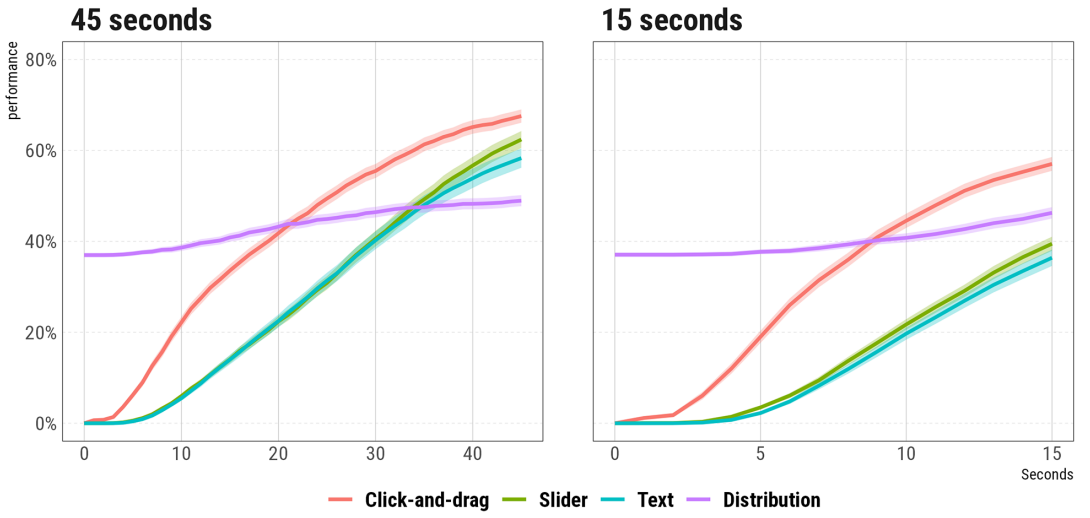

Figure 3. *Performance dynamics by interface.*

Table 3. *Likert scale (1–7) self-reported interface assessment, mean, and 95% confidence interval.*

	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]

The main result—Click-and-Drag allows faster convergence—is replicated also when looking separately by number of bins and shape (Figures A1 and A2 in Appendix Appendix A).

4.4. Self-reported assessment

Table 3 reports the mean and 95% confidence interval of the subjects self-reported assessment (Likert scales, 1–7) on ease of use, frustration, and ability to quickly understand interfaces.

Click-and-Drag ranks first in ease of use and in generating least frustration, and third in the speed of understanding. Most results are not significant, though. Click-and-Drag is perceived as less frustrating than Text ($t(184) = 2.51$, $p = 0.013$, $d = 0.37$) and beats Distribution in all dimensions (ease of use, $t(188) = 2.04$, $p = 0.003$, $d = 0.44$; frustration, $t(187) = 2.73$, $p = 0.007$, $d = 0.39$; understanding, $t(185) = 2.50$, $p = 0.013$, $d = 0.36$), other differences not being significant.

5. Exploratory results

This section includes analyses that were *not* pre-registered, as some *ex post* obviously important analyses did not occur to us *ex ante*.

5.1. Slackers

Elicitation interfaces can be frustrating. This can lead subjects to drop out—that is, not to finish the task, get distracted, or just let the time pass without collecting payoffs. Especially in the case of Mechanical

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

Table 4. *Mean number of slacked screens and distribution of slackers types by treatment.*

	Mean slack	No	Distribution of slackers by type			
			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%

Turk subjects in an online, unsupervised setting, subjects could just leave their browser tab open and tend to other tasks.

The number of persons dropping out on one or more screens—we call them *slackers*—can 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.

Table 4 reports the mean number of screens slacked, and the distribution of the number of screens (out of 24) on which a subject slacked, by treatment. The number of slackers and the severity of their disengagement from the task is severely affected by the treatments. Click-and-Drag proves to be the most engaging, with nearly half the subjects never slacking and most of the others slacking on 1–3 screens. Slider shows significantly more slacking, and Text even more so. For both interfaces, it proved quite frustrating to complete the task for the 15 and 30 bins targets, especially so for Text. Distribution shows a different pattern. In this interface, subjects start from a wide normal distribution, and as a consequence get positive payoffs from the start. Still, using the handles to mimic the target distribution is sometimes hard, and subjects might have decided for a satisficing strategy of keeping the initial payoffs and exert no or little effort. This might explain why *no* subject exerted effort on *all* screens, and the majority of subjects is classified as a serious slacker for Distribution.

Click-and-Drag features the lowest mean amount of screens with no interactions, significantly different from Text ($t(121) = 8/06$, $p < 0.001$, $d = 1.20$), Distribution ($t(115) = 22.1$, $p < 0.001$, $d = 3.21$), and Slider ($t(182) = 2.56$, $p = 0.01$, $d = 0.38$), 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 Click-and-Drag is limited to avoiding slackers. Slackers perform rather badly, and bias the mean performance of the affected interfaces downward. 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 24.

When excluding slackers, Click-and-Drag loses a bit of its edge. It still outperforms Slider and Text in overall accuracy (against Slider: $t(3778) = 12.6$, $p < 0.001$, $d = 0.40$; against Text: $t(1658) = 7.82$, $p < 0.001$, $d = 0.32$), but is not statistically different from Distribution ($t(26) = 1.40$, $p = 0.173$, $d = 0.21$). Note, however, that only two subjects are included for Distribution. This shows how hard it was for Distribution to be used consistently—98% of subjects slacked on *more* than 3 screens, thus voiding of all meaning any statistical test. Nonetheless, the result is also telling: the 2 heavily self-selected

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

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

	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]

subjects working hard on all screens had a performance non distinguishable from the mean subject using Click-and-Drag. It takes motivation and dedication to reach good results with Distribution, and the frustration it generates discourages 98% of subjects from doing so.[5]

5.3. Sentiment analysis of the open-ended comments

Subjects had the possibility of leaving a non-compulsory, open-ended comment on their experience. One hundred forty-nine subjects out of 372 did. The modal reply was a variation of 'no problem', 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.[6] A positive mean sentiment means that the messages were more positive than negative.

Overall, sentiment over our experiment was positive. The ranking of the sentiment analysis by treatment confirms our main results. Click-and-Drag had the highest mean sentiment, at 0.58 (SD 0.59); followed by Slider (0.43, SD 0.5), Text (0.4, SD 0.41), and Distribution (0.36, SD 0.47). Differences were not significant (ANOVA, $F(3, 149) = 1.41$, $p = 0.242$).

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

Additionally to the pre-registered main task, we asked 2 *non-incentivized* direct belief elicitation questions.

This was a first attempt at observing whether the interface impacts reported beliefs. We asked 2 questions related to maximum temperature in New York City on July 4 of 2 given years—2022, the year the experiment took place, and 2042, 20 years in the future. We chose a highly discussed topic

[5]When dropping the independence assumption and averaging across subjects before running the tests, results do not change. Significant against Slider: $t(171) = 6.64$, $p < 0.001$, $d = 1.01$; against Text, $t(72.9) = 4.04$, $p < 0.001$, $d = 0.8$; not significant against Distribution, but with just 2 observations $t(1) = 0.52$, $p = 0.694$, $d = 0.72$.

[6]Sentiment analysis is a text-mining technique that uses dictionaries that associate a valence with any word in a dictionary, and then applies this valence to sentences. The sentiment analysis was run using the `R` package `syuzhet` (Jockers, 2015), which works by assigning an emotional value to each word in a text, in the form of a positive or negative numerical value to indicate the intensity of the emotion.

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

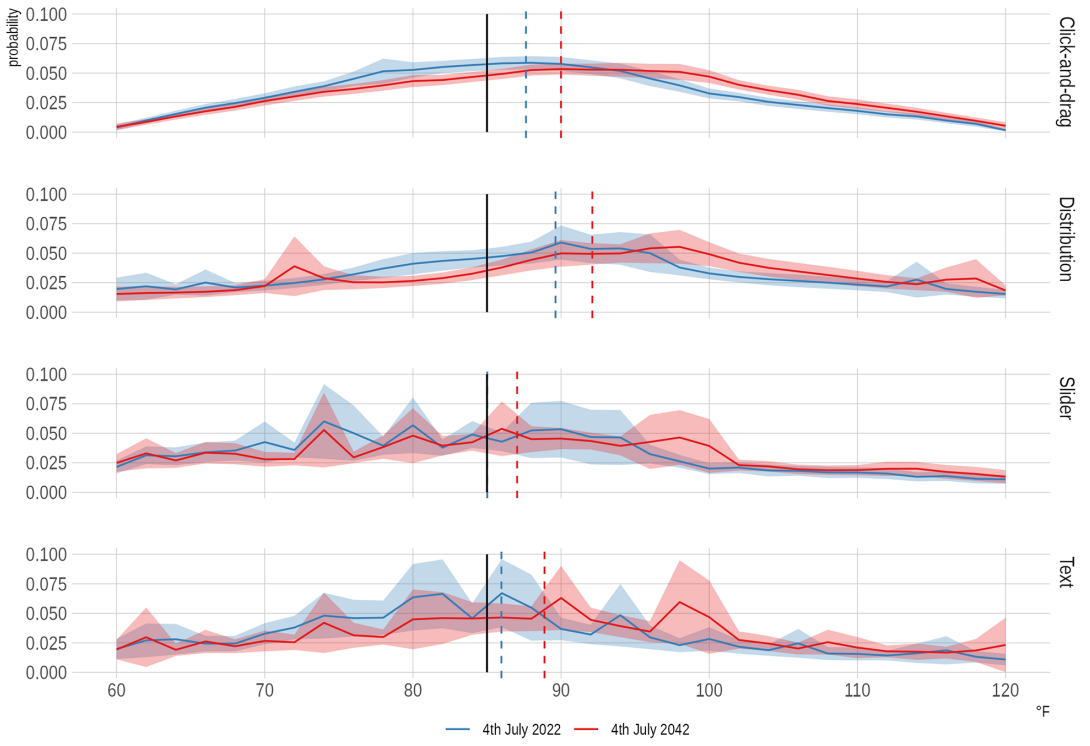

Figure 4. *Distribution of elicited beliefs on maximum temperature on July 4, 2022 and 2042 in New York City, by treatment. Mean beliefs in color; true realization for July 4, 2022, in black.*

ripe with uncertainty—climate change—with which most people are familiar and on which reasonable information is available. We expected each 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 2 questions, subjects entered 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. The questions were not incentivized. This resulted in some subjects slacking—that is, not interacting with the screen in any way. We identify slackers as subjects having not moved the interface at all. Confirming the slackers results above, only 3.16% of subjects slacked with Click-and-Drag, 4.4% with Slider, 14.1% with Text, and 19.6% with Distribution. In the following, we report results *excluding* slackers.

Figure 4 shows the aggregate distributions by treatment for both temperature questions,[7] and comparing it to the actual temperature obtained on July 4, 2022 (85°F).

Table 6 reports the means and confidence intervals of the plotted distributions. Perceived temperatures were overall slightly above the realized temperature of July 4, 2022—with Slider getting the 2022 temperature on average right and Text not far away. Expected warming over 20 years is on average 2.4°F. The 95% confidence interval for the mean prediction is smaller for Click-and-Drag (23.2 degrees for 2022 and 22.8 for 2024) than for any other interface. Distribution has an interval about twice as large as Click-and-Drag (41.6 degrees for 2022 and 45.6 for 2042), Slider more than twice as large (55.8 and 55.7), and Text nearly three times as large (65.9 and 72.5).

[7]These are obtained aggregating data by treatment, computing the mean and confidence interval for each of the 31 bins, for each question.

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

Table 6. *Mean and 95% confidence interval of the elicited beliefs—maximum temperature on July 4.*

	Click-and-Drag	Distribution	Slider	Text
July 4, 2022	87.63 [76.01,99.25]	89.62 [68.82,110.43]	85.02 [57.1,112.93]	85.97 [53.02,118.93]
July 4, 2042	89.99 [78.6,101.39]	92.11 [69.29,114.93]	87.03 [59.19,114.88]	88.88 [52.65,125.11]

Table 7. *Mean (SD) of the scores obtained by subjects for the July 4, 2022 prediction, by interface.*

	Quadratic	Spherical
Click-and-Drag	0.53 (0.034)	0.246 (0.089)
Distribution	0.507 (0.058)	0.193 (0.1)
Slider	0.472 (0.163)	0.188 (0.156)
Text	0.464 (0.145)	0.164 (0.146)

Evaluating the results of this task is less straightforward than with the mimic-the-distribution tasks, as we cannot know the real beliefs of subjects. Nonetheless, the data do give us some suggestions on what role, if any, the interfaces can play when subjects use them to enter their beliefs.

Visually, distributions vary across elicitation interfaces. The aggregate distribution generated by Click-and-Drag subjects is smoother than for all other interfaces, with many less peaks, while Slider and Text feature many extreme peaks. The confidence interval around the mean distribution by treatment is also much larger in several points for Slider, and especially Text. This is likely the result of subjects concentrating beliefs much more with tedious interfaces. In Slider, 8.5% of all subjects concentrated all mass into one bin; in Text, 6.7% did. No subjects in Click-and-Drag and Distribution, where it is easier to introduce smooth distributions, did.[8]

Another way to compare distributions is to use proper scoring rules, to see how participants would have scored if their beliefs were incentivized. This is possible for 2022 only, since we know the realized maximum temperature, 85°F. Table 7 reports the mean (SD) of the scores obtained by applying the Quadratic and the Spherical Scoring Rules to each participant's beliefs.[9]

The mean score obtained by Click-and-Drag is significantly higher than Text ($t(85) = 3.95$, $p \leq 0.001$ $d = 0.65$), Slider ($t(93) = 3.25$, $p = 0.002$ $d = 0.50$), and Distribution ($t(118) = 3.03$, $p = 0.003$ $d = 0.49$) for the QSR; and the same goes for the SSR (against Text, $t(124) = 4.34$, $p \leq 0.001$, $d = 0.69$; against Slider, $t(135) = 3.01$, $p = 0.003$, $d = 0.46$; against Distribution, $t(156) = 3.63$, $p \leq 0.001$, $d = 0.56$). This result might seem striking when observing that Slider ($t(86) = 0.022$, $p = 0.983$) and Text ($t(78) = 0.987$, $p = 0.327$) subjects got the 2022 prediction *on average* right while Click-and-Drag ($t(91) = 5.96$, $p \leq 0.001$) and Distribution ($t(77) = 5.78$, $p \leq 0.001$) overshot. But when looking more closely, Click-and-Drag subjects put the higher mass on the correct bin (bin 84-5, Click-and-Drag 5.7%, Slider, 4.9%, Text, 4.6%, and Distribution, 4.5%), and distributed less concentrated mass on the wrong bins; Slider and Text seem to do well on average *despite* having subjects placing a lot of mass in wrong bins, just because these large mistakes on average cancel each other.

So all in all we see that Click-and-Drag gives a smoother aggregate distribution that scores best on proper scoring rules but overshoots the realized temperature for 2022. Slider and Text were closest to the actual measured temperature, but mostly *despite* themselves: the scattered peaks induced by both interfaces, together with lower mass in the tails, combine to give a 'wisdom of the crowd' effect,

[8]Detailed plots of beliefs entered by *each* participant are given in the online data repository on GitHub.

[9]$Score_{QSR} = 0.5 + p_i - \sum_{j=1}^{n-1} 0.5 \cdot p_j^2$; $Score_{SSR} = \frac{p_i}{\sqrt{\sum_{j=1}^{n} p_j^2}}$, where p_i represents the probability mass allocated to the true outcome.

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

whereby while most subjects are wrong, their aggregate prediction is accurate—though with a large confidence interval.

7. Limitations

We believe our study constitutes a fair and direct testing of Click-and-Drag against other popular belief elicitation interfaces. But, to evaluate our results, we also highlight here the limitations of our work.

First, we rely on a *visual* representation of distributions; subjects have to mimic a target that is provided graphically, in the form of probability density functions. This is just one possible way of conveying uncertainty and probability distributions. It is certainly the most widespread in statistics, but it might be different from what subject have in mind when they think of probability, uncertainty, beliefs, and forecasts. Moreover, the choice of distributions to mimic could also be extended. Our results hold only on the tested shapes, and other shapes could arguably be more suited to some interfaces—for instance, all probability mass restricted to 2 specific bins would make Slider and even Text shine with respect to Distribution, and probably also Click-and-Drag. It is straightforward, however, to run further tests with different distributions and different ways to convey uncertainty.

Second, we could not include all potential alternative belief interfaces. One crucial missing competitor is the Distribution Builder interface, introduced by Sharpe and Goldstein (2000), used among others by Goldstein et al. (2008), and refined in Goldstein and Rothschild (2014), a frequency-based method that uses discrete units of probability, with an intuitive interface where subjects press on a '+/-' button to visually add/subtract probability mass to any of a series of bins (code on GitHub).[10] Other interfaces might be used in the industry, on prediction websites, or in academia of which we are not aware. A straightforward extension of the current work would be to run a second horse race with these interfaces.

Third, we did not fully appreciate the *slacker* problem ex ante—the fact that Mechanical Turk subjects would drop off the task. Yet we think that this is more a blessing than a curse. The presence of slackers was clearly lower with Click-and-Drag, showing that other interfaces generate more boredom, frustration, and drop-outs; and results are robust to eliminating slackers. Since drop-outs are bound to exist, knowing which interface is less likely to generate them and is more robust to their exclusion is important, especially for online settings.

Fourth, Mechanical Turk studies can suffer from poor subject pool problems, and our data are thus noisier than if we had ran the experiment in a laboratory or using other platforms such as Prolific. This also explains why accuracy was in absolute terms rather low for what can be seen as a rather simple task—accuracy on individual screens never exceeded about 80% for the simpler screens (Appendix Appendix A). While lower than what an academic reader could expect, we think that such a performance was actually *good* given the sample. Extensions to other subject pools is straightforward, though; and we deliberately chose to go for the noisier subject pool, since we wanted to see how complex belief elicitation could be made palatable and easy for *anyone* with minimal instructions and the proper interface.

8 Conclusion

We introduce Click-and-Drag, a new belief elicitation interface, and test its performance against 3 other interfaces used in the experimental literature or by a crowd-prediction website. We find considerable variance in performance of different interfaces across different task characteristics, such

[10]To be more precise, this Distribution Builder allows subjects to build a discrete probability distribution using a 'click-and-drag' mechanism of its own. A distribution is divided in 64 blocks (containing a probability mass of 1/64), displayed as icons of a person, and subjects can place the 64 blocks one by one to build a discrete distribution displayed as a histogram on the screen. The idea behind this interface is that each block represents a potential scenario for the person (hence the person icon) and allows the user to develop a intuition for the histogram and its implied likelihoods.

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

as allotted time, number of bins, and shape of a target distribution. The elicitation interface appears to clearly matter.

We find that Click-and-Drag overall outperforms the competition, 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 allowed 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 clearly outperformed especially the Slider and Text input interfaces.

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

Our results suggest that researchers who plan to elicit belief distributions will likely benefit from adopting Click-and-Drag. Moreover, with an improved interface to elicit entire distributions in a quick and intuitive way, the choice to elicit a distribution, rather than just a mean or a mode, might become more attractive for both researchers and practitioners 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 empirical data what we can reasonably expect, and this should make it easier for fellow researchers—or, indeed for the crowd-prediction industry—to think of adopting it for their belief elicitation studies.

Data availability statement. For further information regarding the data for this paper, we refer you to https://osf.io/83asz/. Here one can find the original data, and scripts to replicate the data analysis used in the paper.

Acknowledgements. 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.

Funding statement. 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. https://doi.org/10.1016/j.jbef.2015.12.001

Comerford, D. (2022). *Response bias in survey measures of expectations: Evidence from the survey of consumer expectations' inflation module. Journal of Money, Credit and Banking*.

Danz, D., Vesterlund, L., & Wilson, A. J. (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review*, *112*(9), 2851–2883.

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. https://doi.org/10.1093/poq/nfn062

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. https://doi.org/10.1016/j.joep.2019.04.005

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Federal Reserve Bank of New York (FRBNY). (1999). *Survey of consumer expectations* [Technical report]. 2013–2020 Federal Reserve Bank of New York (FRBNY). http://www.newyorkfed.org/microeconomics/sce

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

Goldstein, D. G., Johnson, E. J., & Sharpe, W. F. (2008). Choosing outcomes versus choosing products: Consumer-focused retirement investment advice. *Journal of Consumer Research*, *35*, 440–456.
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*, *25*, 795–823.
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* (pp. 163–202). New York: Palgrave Macmillan.
Haruvy, E., Lahav, Y., & Noussair, C. N. (2007). Traders' expectations in asset markets: Experimental evidence. *American Economic Review*, *97*, 1901–1920. https://doi.org/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*. https://github.com/mjockers/syuzhet
Kröger, S., & Pierrot, T. (2019). *What point of a distribution summarises point predictions?* [WZB discussion paper]. Technical Report.
Kryvtsov, O., & Petersen, L. (2021). Central bank communication that works: Lessons from lab experiments. *Journal of Monetary Economics*, *117*, 760–780. 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. https://doi.org/10.1017/pan.2020.42
Manski, C. F. (2004). Measuring expectations. *Econometrica*, *72*, 1329–1376.
Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*, 1087–1096. http://www.jstor.org/stable/2629907
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. 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. https://doi.org/10.1509/jmkr.38.4.430.18904
Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*, 783–801.
Schlag, K. H., & van der Weele, J. J. (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. *Annual Review of Economics*, *6*, 103–128.
Sharpe, W. F., & Goldstein, D. G. (2000). *The distribution builder: A tool for inferring investor preferences*.
Spiliopoulos, L., & Ortmann, A. (2017). The BCD of response time analysis in experimental economics. *Experimental Economics*, *21*, 383–433: https://doi.org/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.

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

Appendix A. Detailed results by screen type

Table A1. *Final performance, mean, and 95% confidence interval, by condition for all interfaces.*

	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]

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

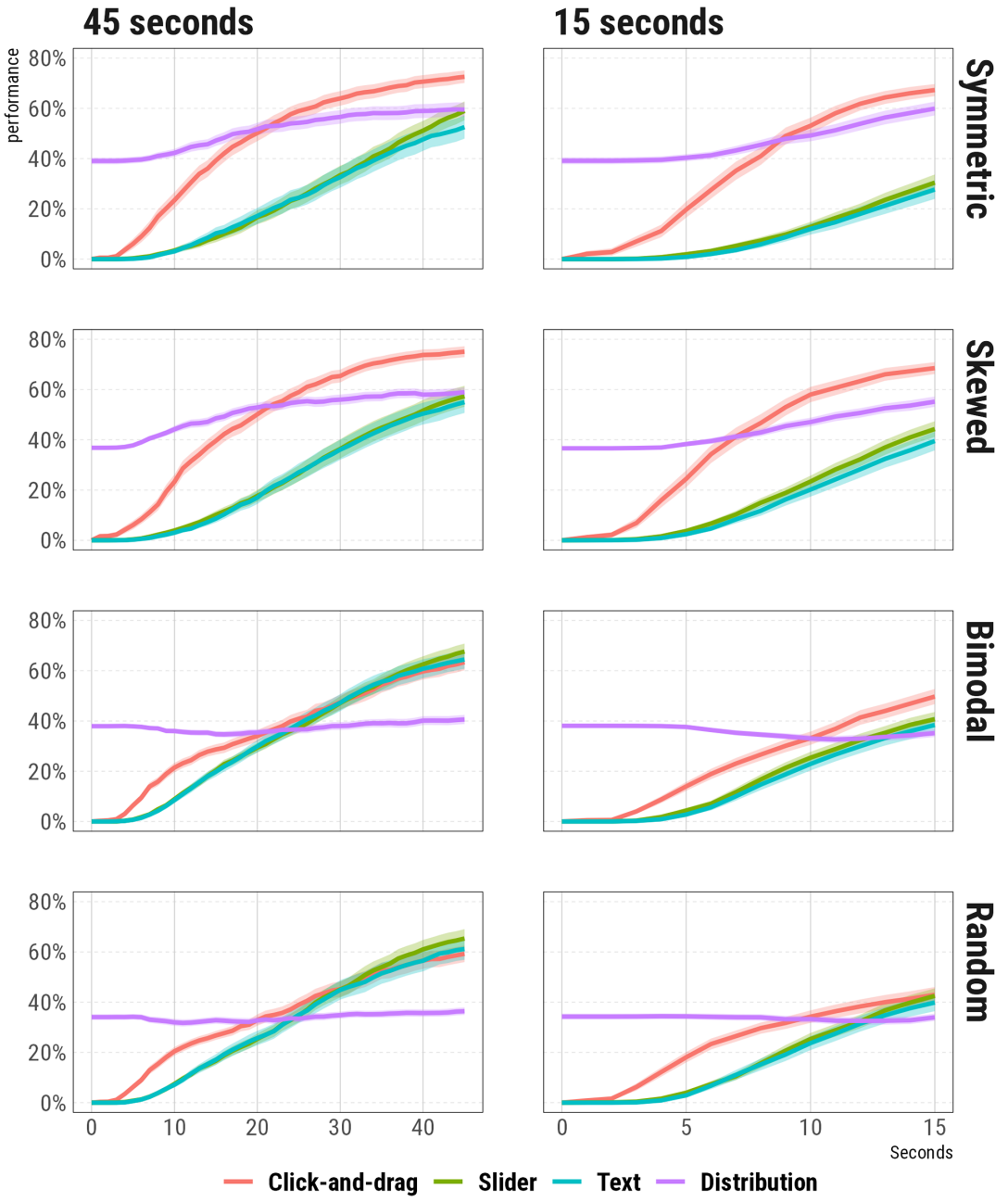

Figure A1. *Performance dynamics by interface—for different target shapes.*

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

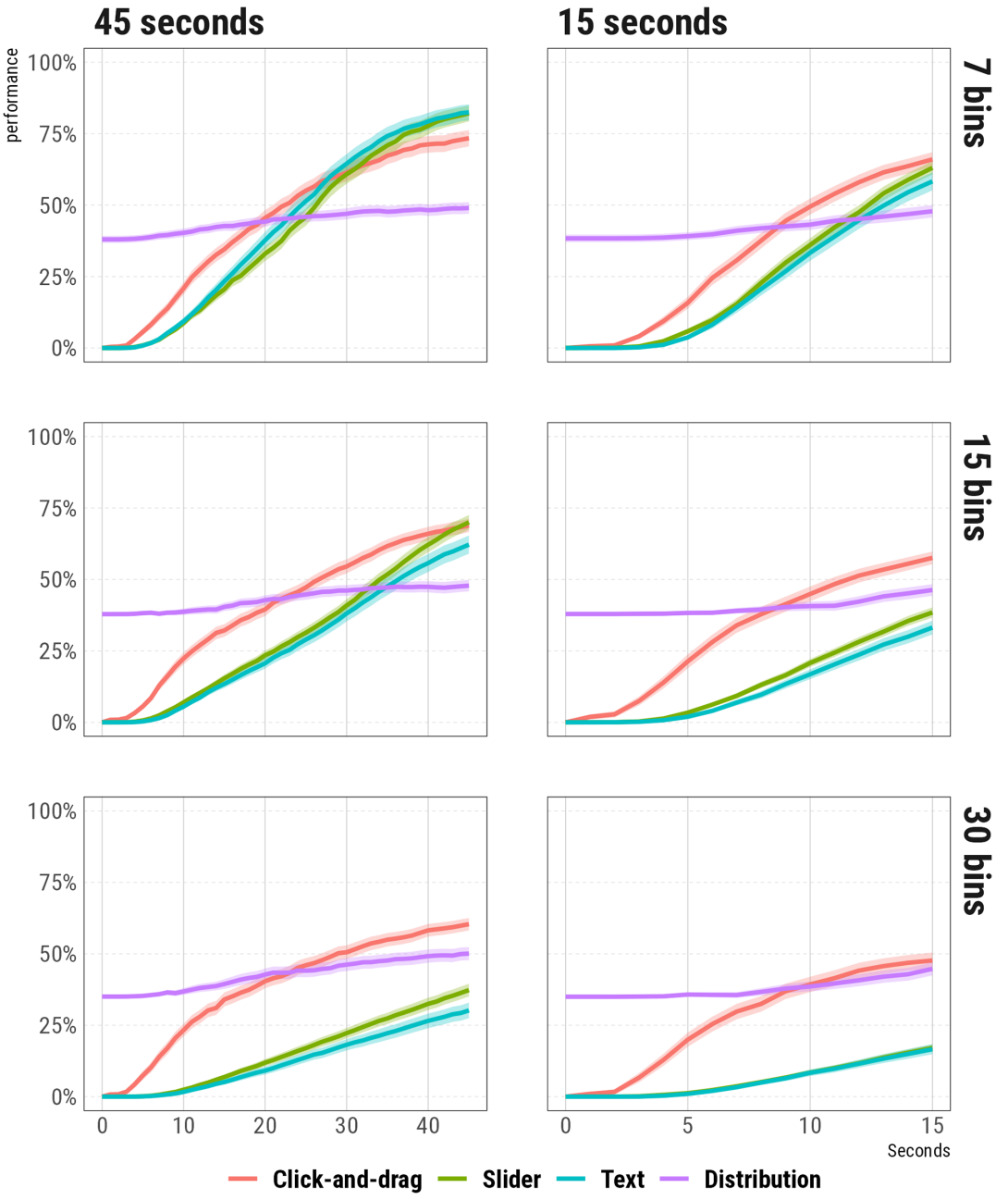

Figure A2. *Performance dynamics by interface—for different number of bins.*

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

Appendix B. Screenshots

Figure B1. *Screenshot of the Text interface in action.*

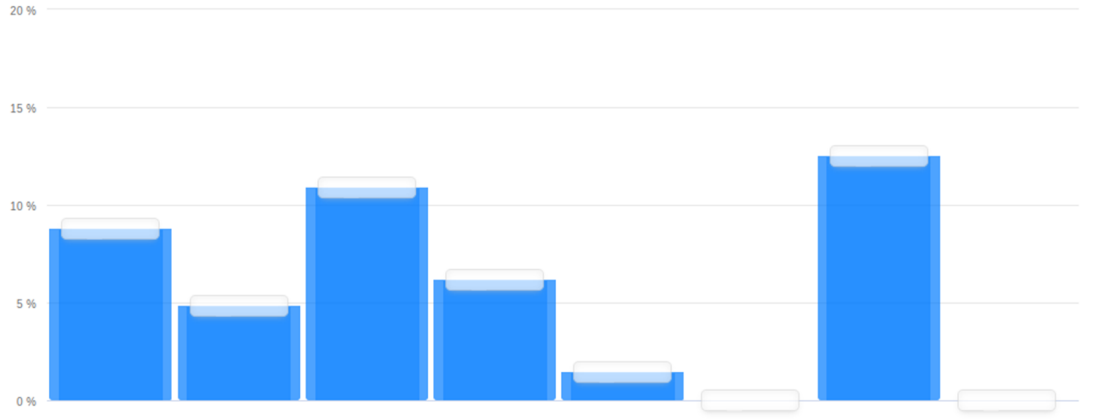

Figure B2. *Screenshot of the Slider interface in action.*

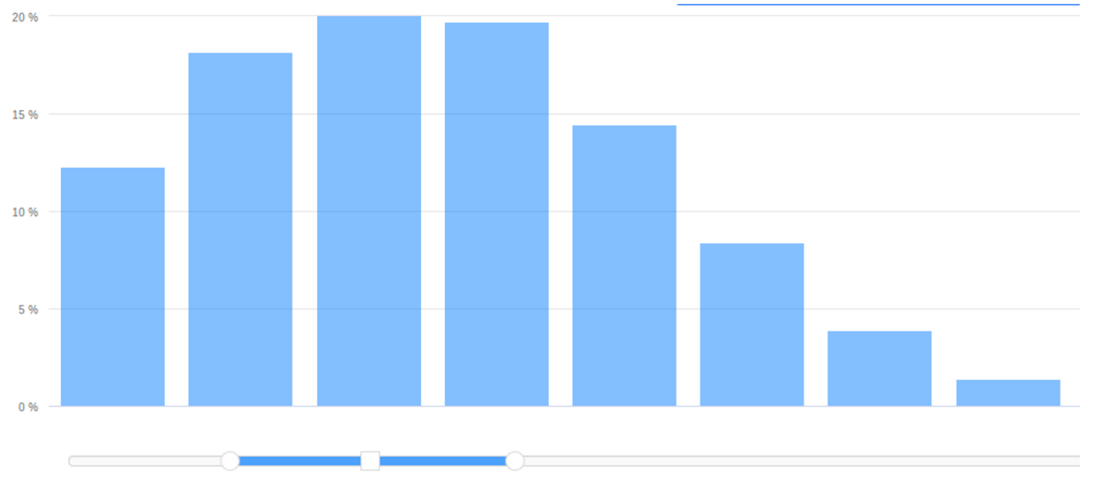

Figure B3. *Screenshot of the Distribution interface in action.*

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

Appendix C. Target distributions

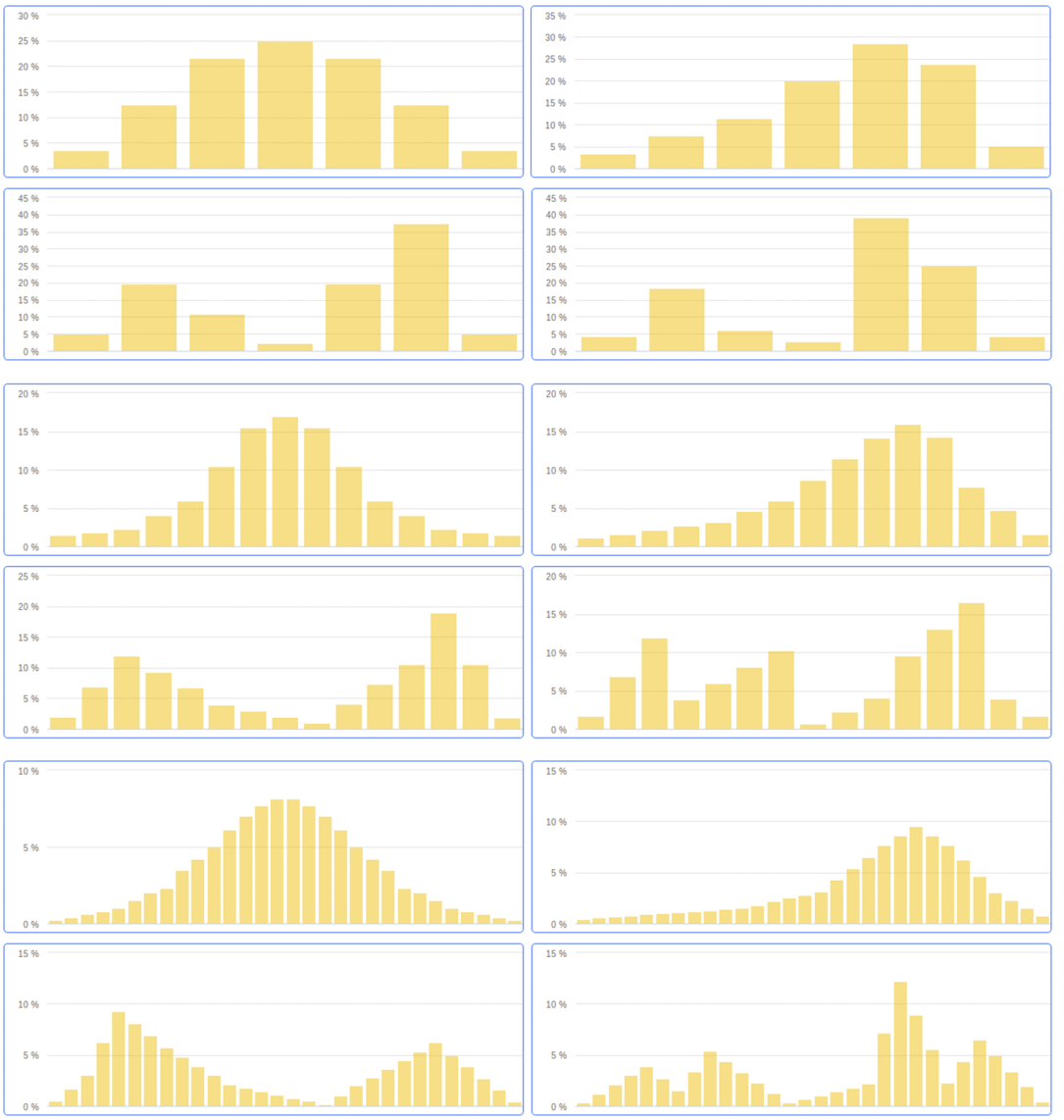

Figure C1. *The 12 target distributions.*

Appendix D. Devices used by subjects

The experiment 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, and 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 D1 provides data on the system used by subjects and their input devices, by treatment.

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

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

Table D1. *Input devices by treatment, share of subjects.*

	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%

the share that used *just* those is so small (9 subjects using only a touchpad and 1 subject using only a touchscreen) as not to warrant a robustness analysis by type of input device.

Appendix E. Additional tests

Table E1. *Accuracy t-test results after averaging by subject, Click-and-Drag against noted interface, all dimensions.*

	DoF	Stat	*p*	*d*
Overall				
Distribution	153.0567	7.844287	0.000	−1.1381687
Text	172.4061	7.760803	0.000	−1.1437196
Slider	182.6839	5.928641	0.000	−0.8672193
Difference 45 vs. 15 seconds				
Slider	181.1811	−6.397477	0.000	0.9439647
Distribution	150.6279	5.390289	0.000	−0.7864023
Text	167.5060	−4.805294	0.000	0.7177108
Difference 7 vs. 15 bins				
Text	173.7942	−7.170793	0.000	1.0699076
Slider	179.6188	−5.672178	0.000	0.8372205
Distribution	165.8082	3.550186	0.001	−0.5225658
Difference 15 vs. 30 bins				
Slider	181.4142	−10.771975	0.000	1.5889422
Distribution	151.7424	7.811113	0.000	−1.1462840
Text	161.8690	−7.468504	0.000	1.1191032
Shape: Symmetric				
Text	157.2231	11.766471	0.000	−1.7398753
Slider	175.6921	11.368250	0.000	−1.6738247
Distribution	178.0894	4.739753	0.000	−0.6877157
Shape: Skewed				
Text	164.0326	11.861943	0.000	−1.7515385
Slider	180.9572	9.935213	0.000	−1.4600386
Distribution	183.9978	6.768136	0.000	−0.9839087
Shape: Bimodal				
Distribution	127.9885	7.158407	0.000	−1.0376100
Text	182.5825	3.721195	0.000	−0.5458610
Slider	168.0176	1.467385	0.144	−0.2136190
Shape: Random				
Distribution	116.1297	8.103312	0.000	−1.1702483
Text	177.8366	1.291713	0.198	−0.1906088
Slider	176.2867	−1.190962	0.235	0.1737456

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

Appendix F. Experimental instructions

Appendix F.1 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; and 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.

Appendix F.2 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.

https://doi.org/10.1017/jdm.2023.21 Published online by Cambridge University Press

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 G. Control questions

The 4 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, and 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]*

Cite this article: Crosetto, P. and de Haan, T. (2023). Comparing input interfaces to elicit belief distributions. *Judgment and Decision Making*, e27. https://doi.org/10.1017/jdm.2023.21