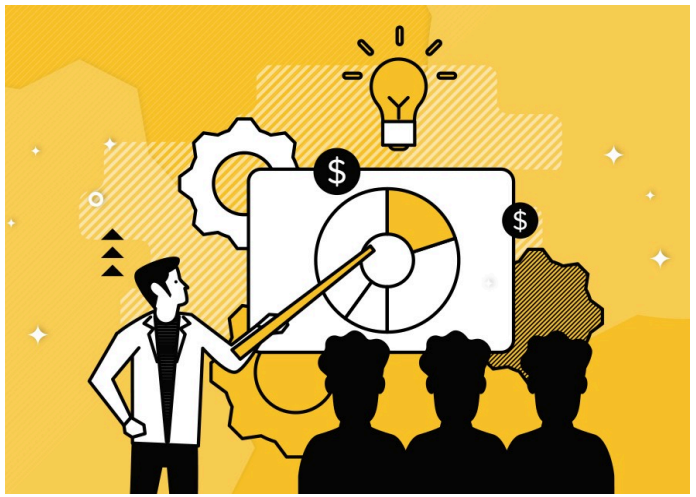
How to elicit beliefs when the distribution matters

Paolo Crosetto (INRAE) & Thomas De Haan (UiB)

SEET Meeting, Valencia, February, 3rd 2023

This talk

A sales pitch



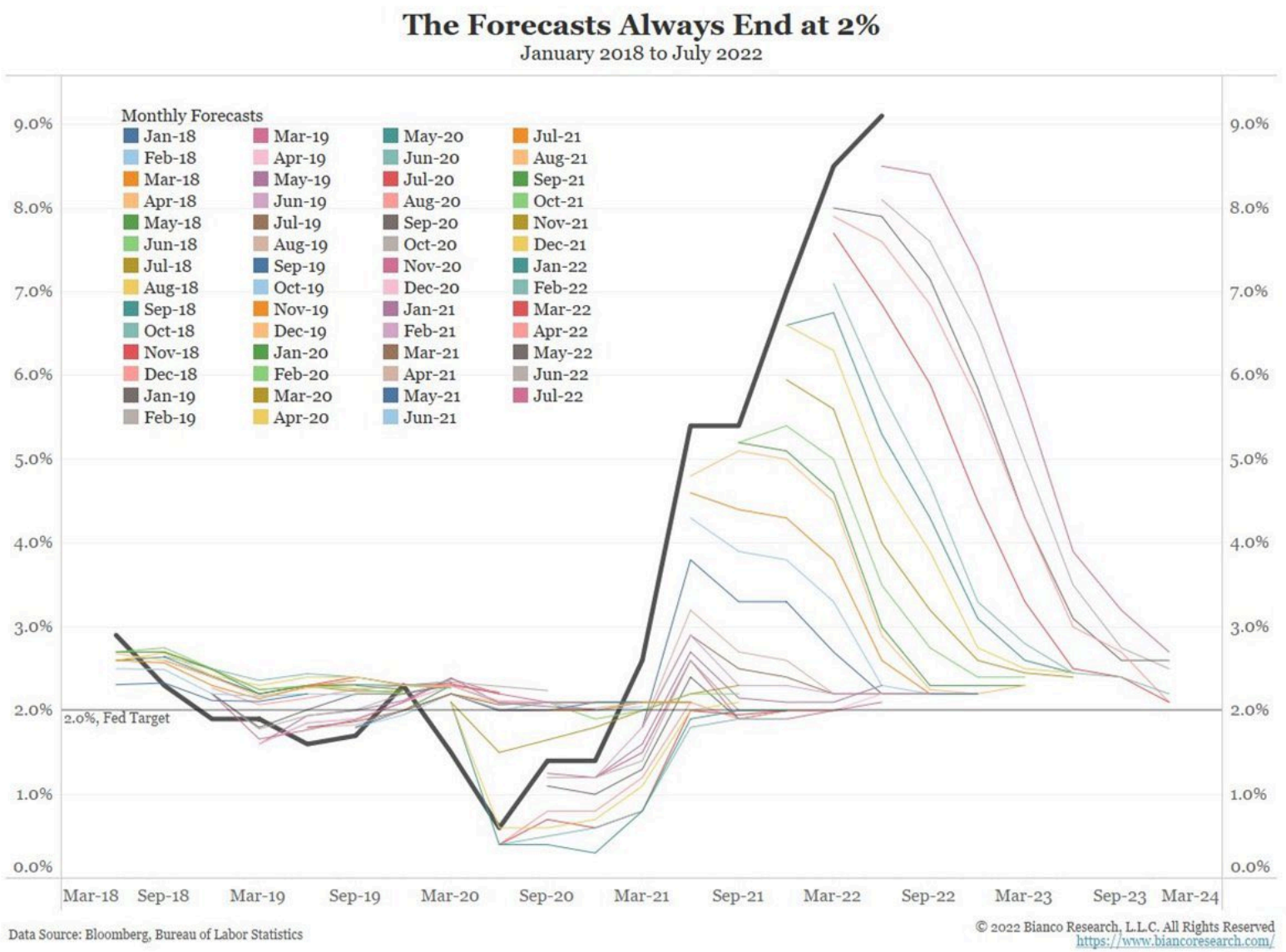
A (nascent) research program

It’s tough to make predictions

It’s tough to make predictions

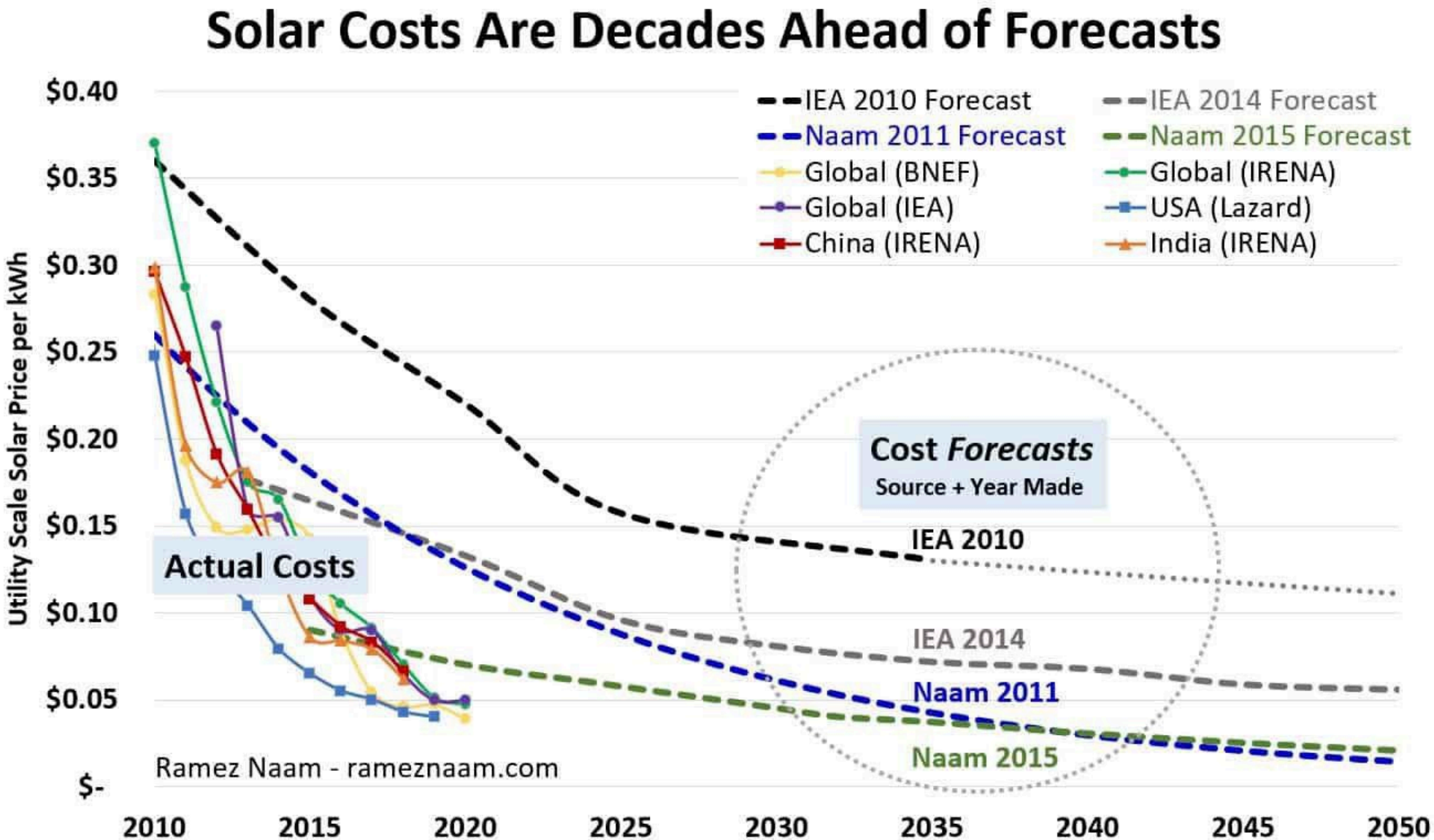
(especially about the future)

Inflation



Data Source: Bloomberg, Bureau of Labor Statistics

Solar


Solar Costs Are Decades Ahead of Forecasts
Utility Scale Solar Price per kWh
$0.40
$0.35
$0.30
$0.25
$0.20
$0.15
$0.10
$0.05
$-
2010
2015
2020
2025
2030
2035
2040
2045
2050
IEA 2010 Forecast
IEA 2014 Forecast
Naam 2011 Forecast
Naam 2015 Forecast
Global (BNEF)
Global (IRENA)
Global (IEA)
USA (Lazard)
China (IRENA)
India (IRENA)
Cost Forecasts
Source + Year Made
IEA 2010
IEA 2014
Naam 2011
Naam 2015
Actual Costs
Ramez Naam - rameznaam.com

Are they just *bad*

or do they just don’t face the right incentives?

Research program: how best to incentivize forecasters?

The problem

- usually: forecasters do **not** face incentive compatible contracts
  - i.e. *they don't pay for their mistakes*
  - market signals are distorted
  - opaque, small market
- solutions:
  - reputation mechanism (of some sort)
  - prediction markets
  - **write incentive-compatible contracts**

The tools & the roadmap

> we have the right tool: **scoring rules**

BUT:

- we *don't know* how good they are
- we *don't know* if they'd work in this context
- we **do** know that subjects do not understand them
- to the point that it might be better **not to use them** (Danz et al. AER 2022)

The (nascent) research program

1. How to **evaluate** scoring rules?

   Develop a paradigm to evaluate scoring rules in the context of forecasts

2. What is **the best** scoring rule?

   Horse race of scoring rules within the new paradigm

3. Would it **work** in the field?

   Field experiments in prediction website / markets

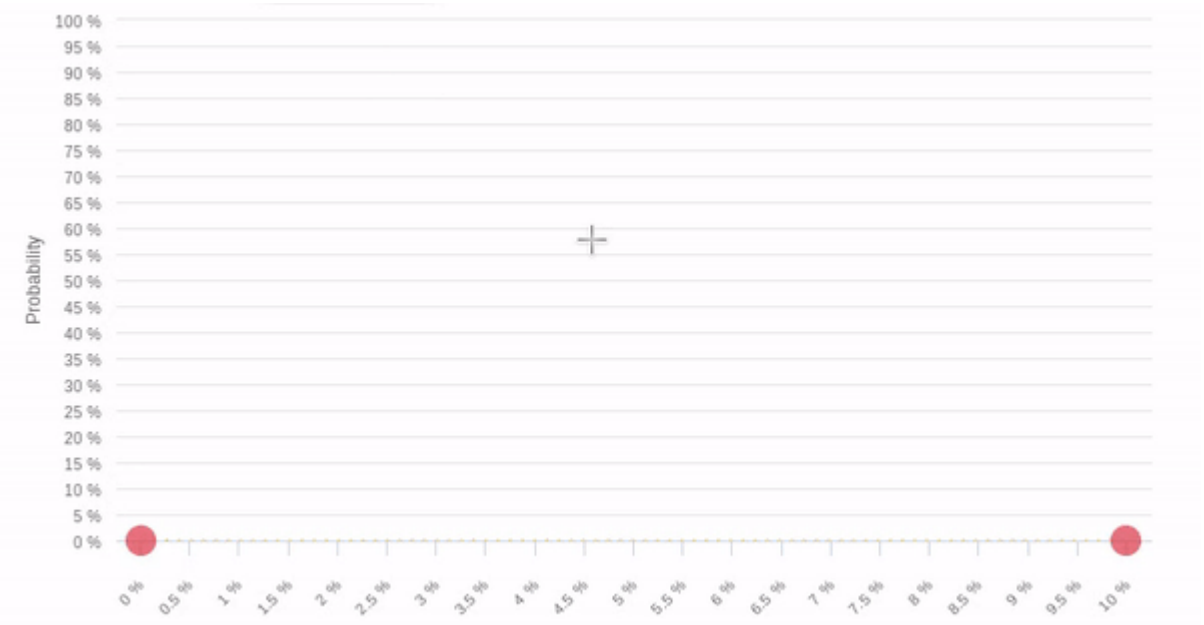
The (nascent) research program

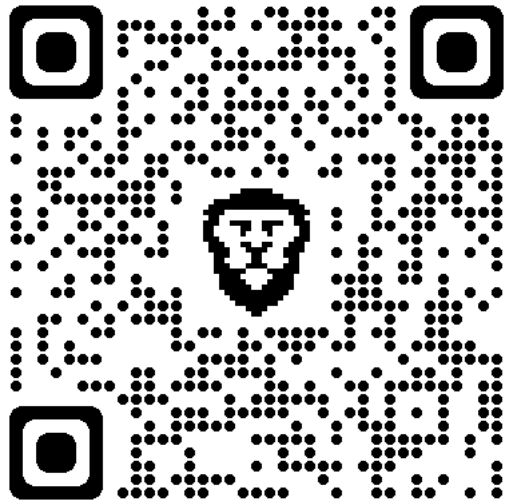
0. How best to **input** beliefs and forecasts? Does it matter?

   *Horse race of belief/forecast elicitation **interfaces***

Sales pitch: the **best** tool to elicit beliefs

One-slide version of the talk

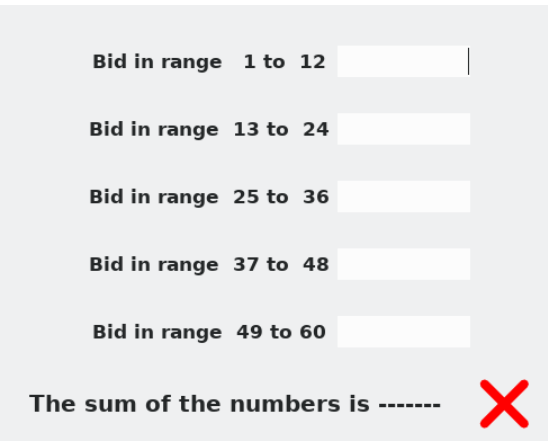




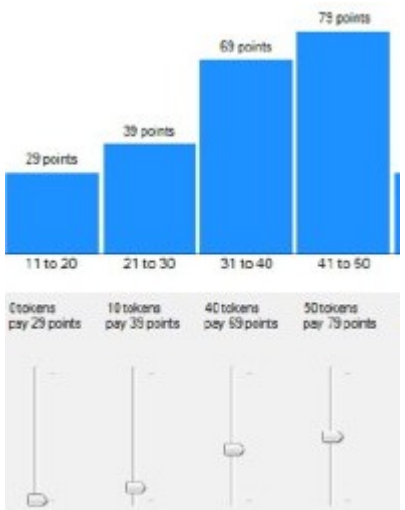
The sale pitch

- we make a case for **precise belief elicitation** as a **key lab** tool
- we **introduce** and **test** against others the **Click-and-Drag** interface

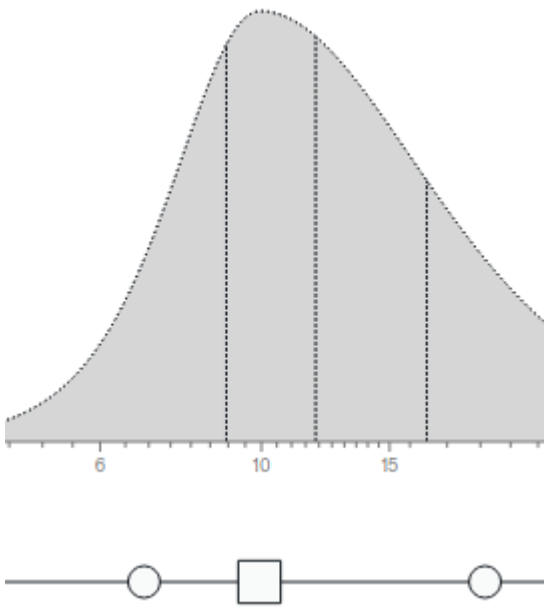
Text-based



Slider-based



Metaculus



Why eliciting belief distributions?

Beliefs matter

In the lab

- *perception* of stimuli is increasingly important
- experiments yield *fuzzy measures* – e.g. cognitive uncertainty

In the real world

- increasing importance of *forecasting* and *prediction markets*
- perception of *risks*, *costs* & *benefits* of **policies** crucial

How you elicit beliefs matters

Belief elicitation is usually **unintuitive** to subjects

- the tool used is **not neutral**:
  - forcing point estimates? **bias**!
  - asking for distributions? **hard**!
  - confidence intervals? **what**?

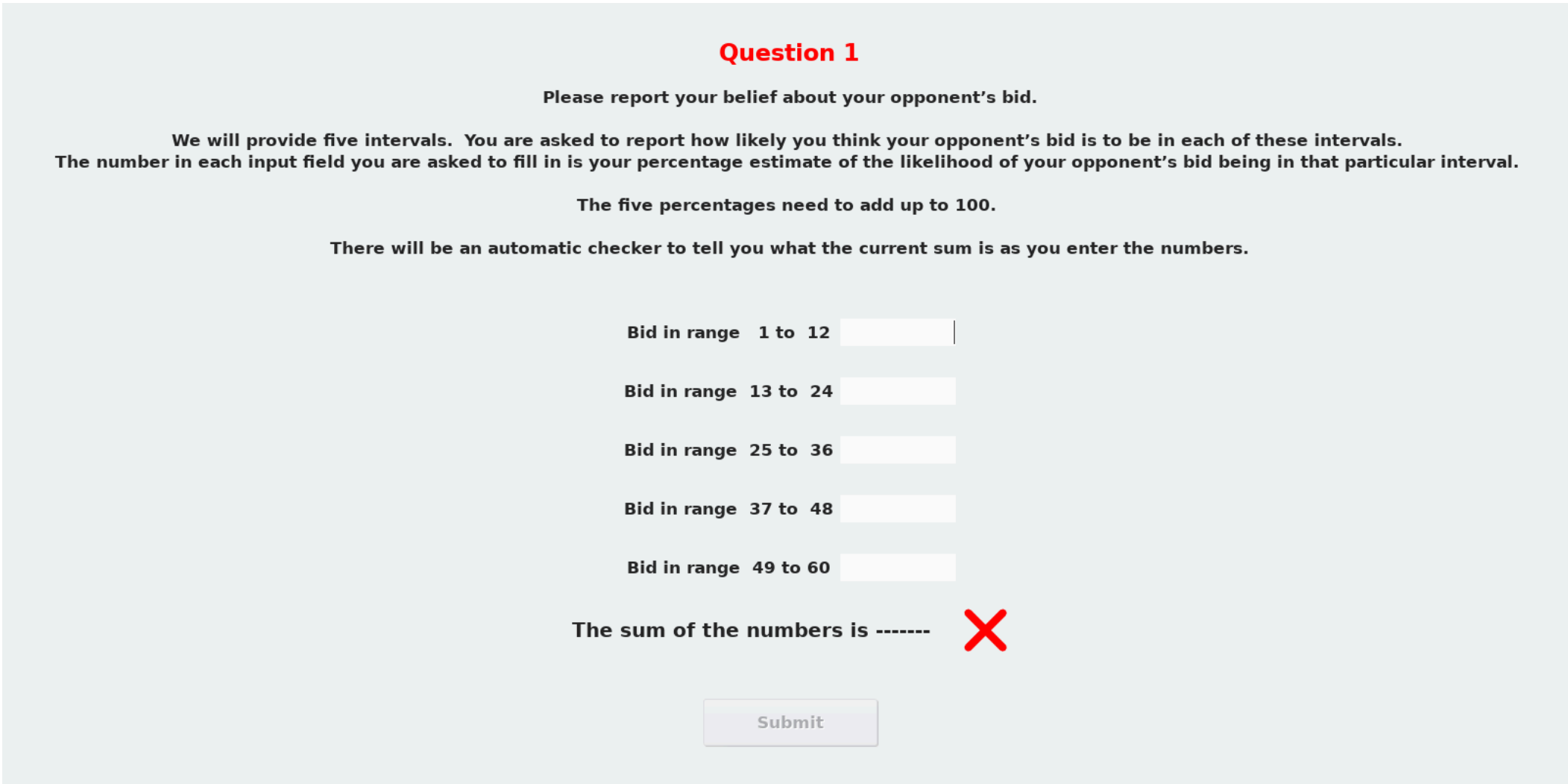
The state of the art

What do we aim for

A good interface *does not get in the way* of subjects. It should:

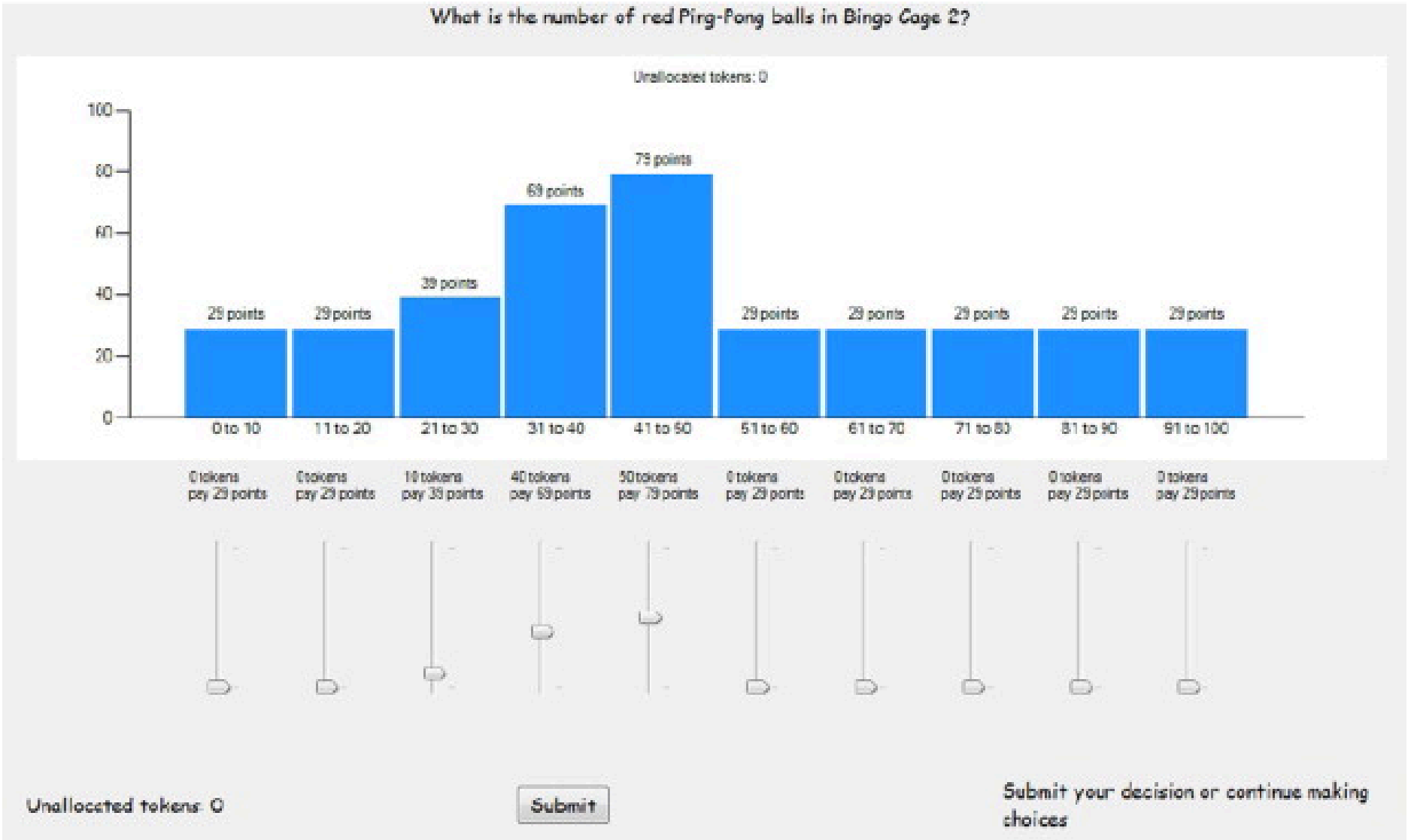
- make it easy to input **distributions** but also allow for point estimates
- allow for a **fast** sketch of your belief
- allow to be **accurate**
- **scale** to a lower/higher number of bins
- allow for non-standard, skewed, bimodal, **whatever** distributions

Text-input interfaces



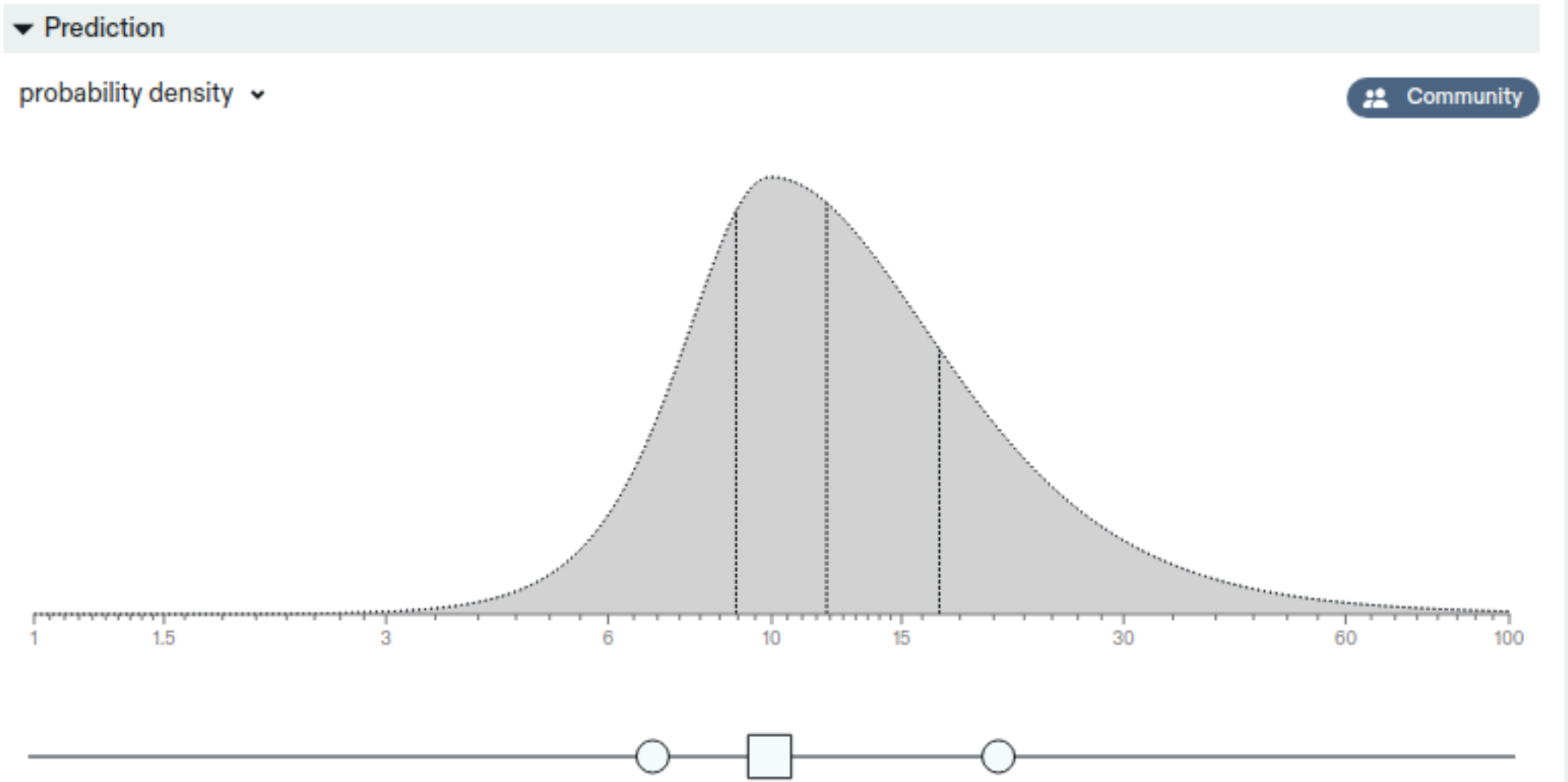
Crosetto et al. (JEPsy, 2020): *slow, scales badly, sum to 100; but precise*

Slider interfaces



Harrison et al (EL 2015): *better, scales OKish, sum to 100; but intuitive*

Distribution interfaces

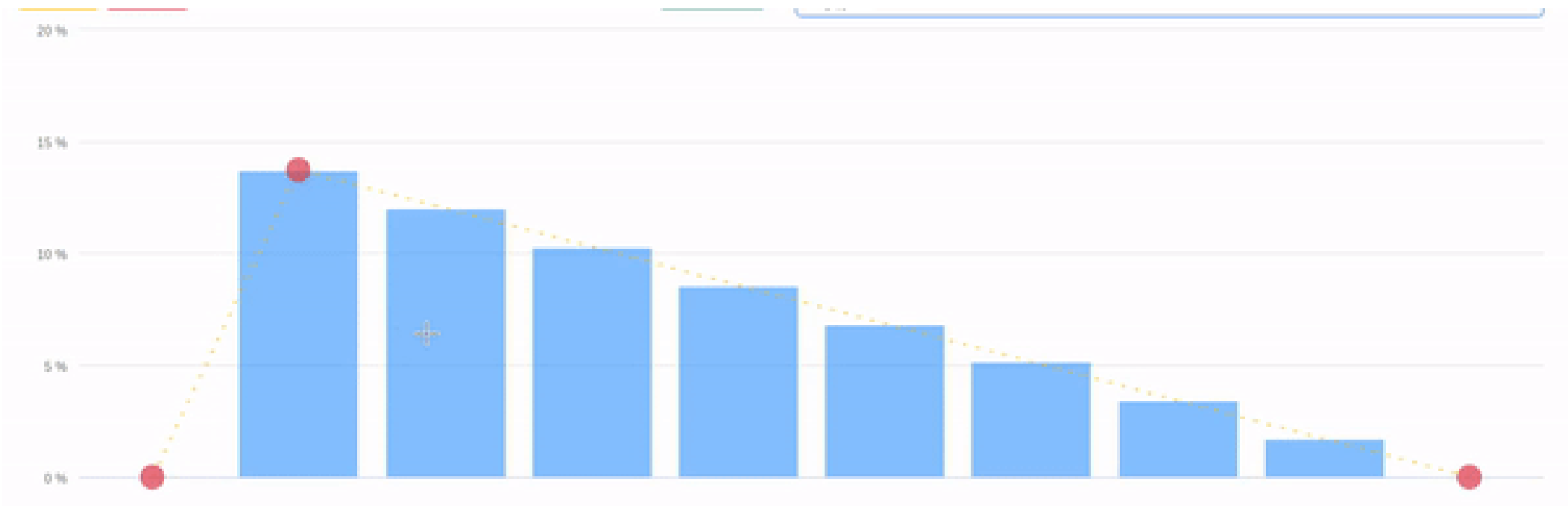


Metaculus: *fast, continuous, no sum to 100; but: limited in shapes, requires stat*

Click-and-Drag

Say hello to Click-and-drag

- adjust the graph by **adding, moving or removing anchor-points**
- **add** anchor-points by **clicking anywhere** on the graph.
- **move** anchor-points around by **dragging them**.
- **remove** anchor-points by **clicking on them**.



Experimental design

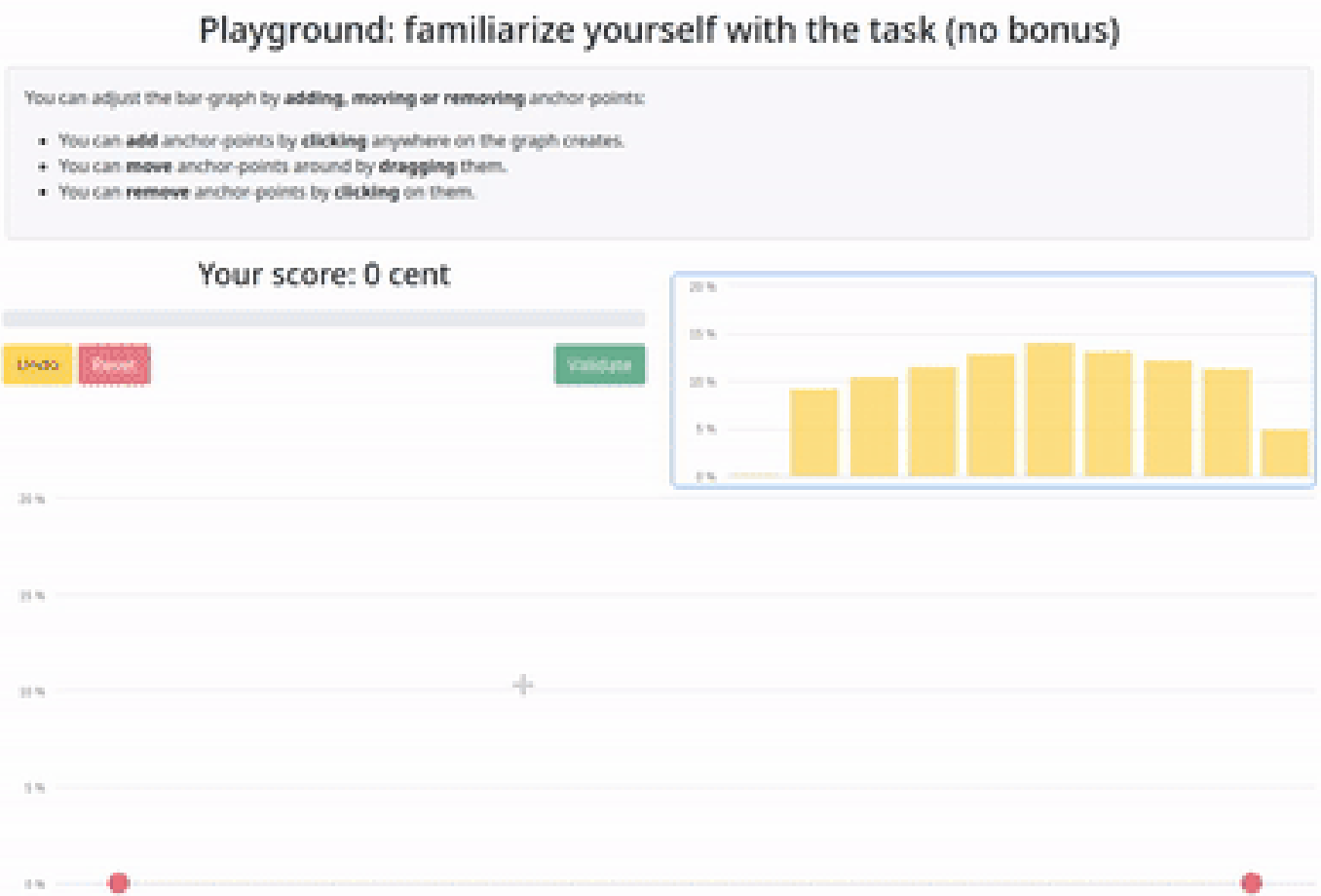
Task

- subjects need to **match a given distribution**

- within a certain **time constraint**

- the **closer** the match, the **higher** the payoff

Task - live!


Playground: familiarize yourself with the task (no bonus)
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.
Your score: 0 cent

Dimensions

- **Interfaces**: Click-and-Drag, Text, Slider, Distribution

- **Time** constraint: **15** and **45** seconds

- **Shape**: normal, skewed, skewed bimodal, complex

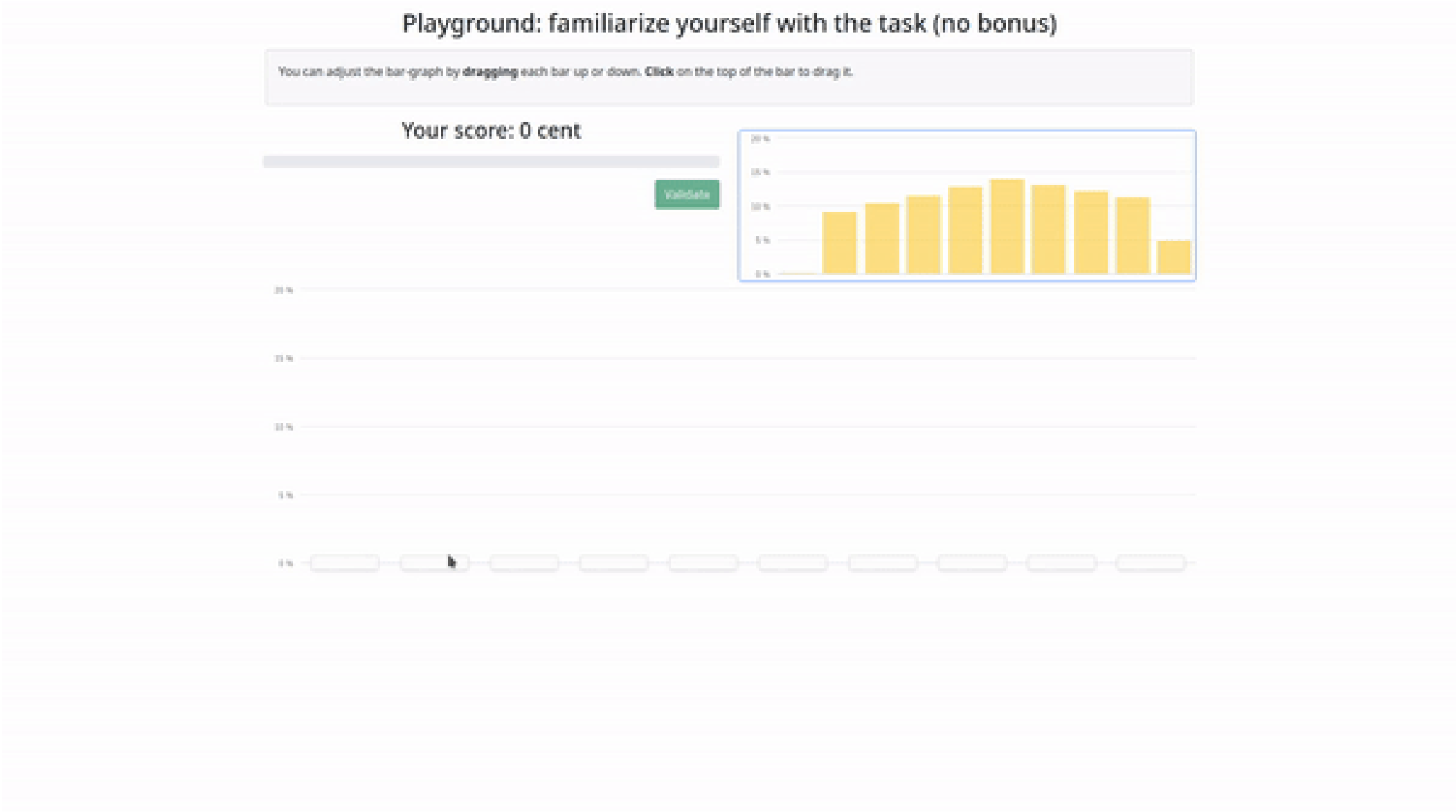
- **Scale**: 7, 15 or 30 bins

Treatments

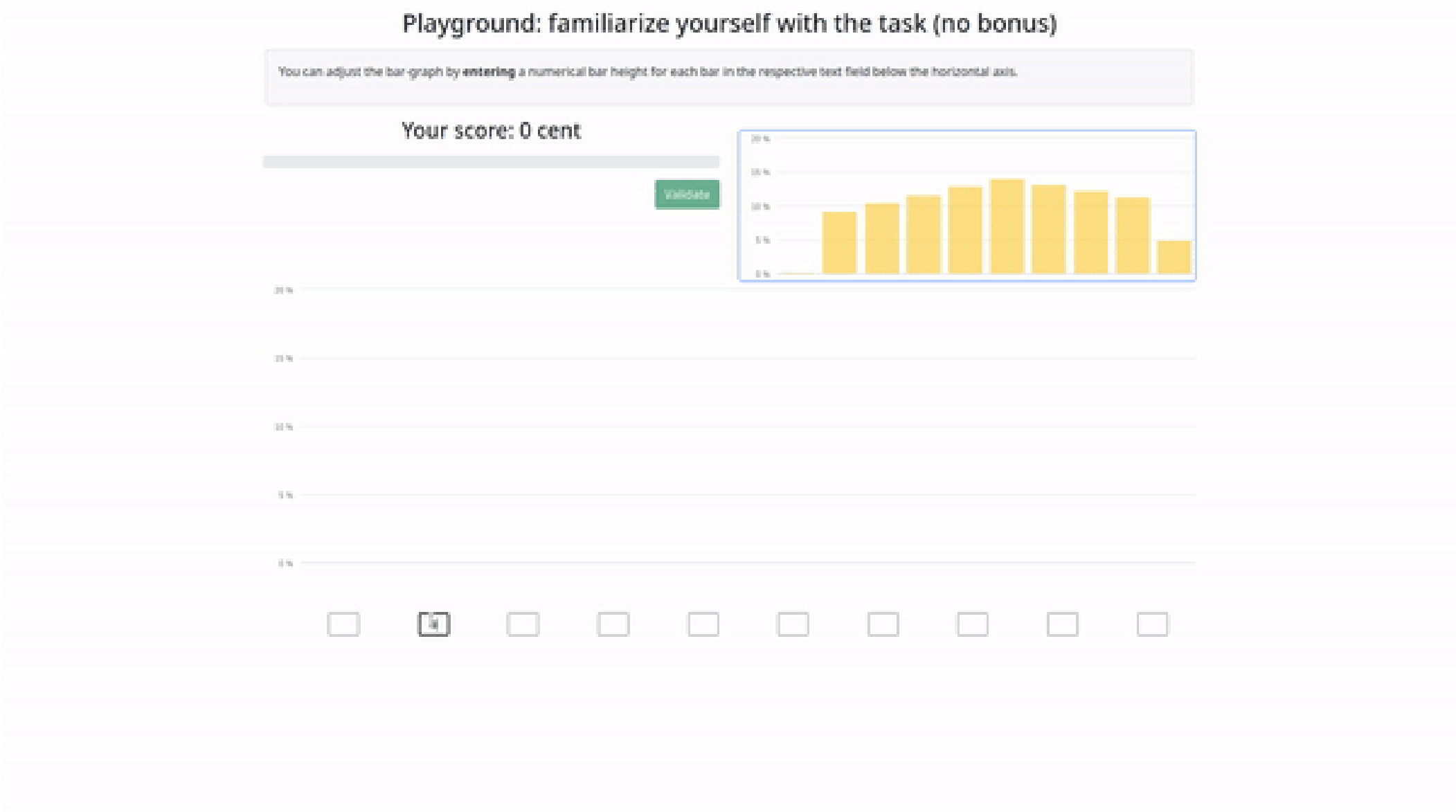
- **between-subjects:** interface

- **within-subjects**: time, difficulty, scale

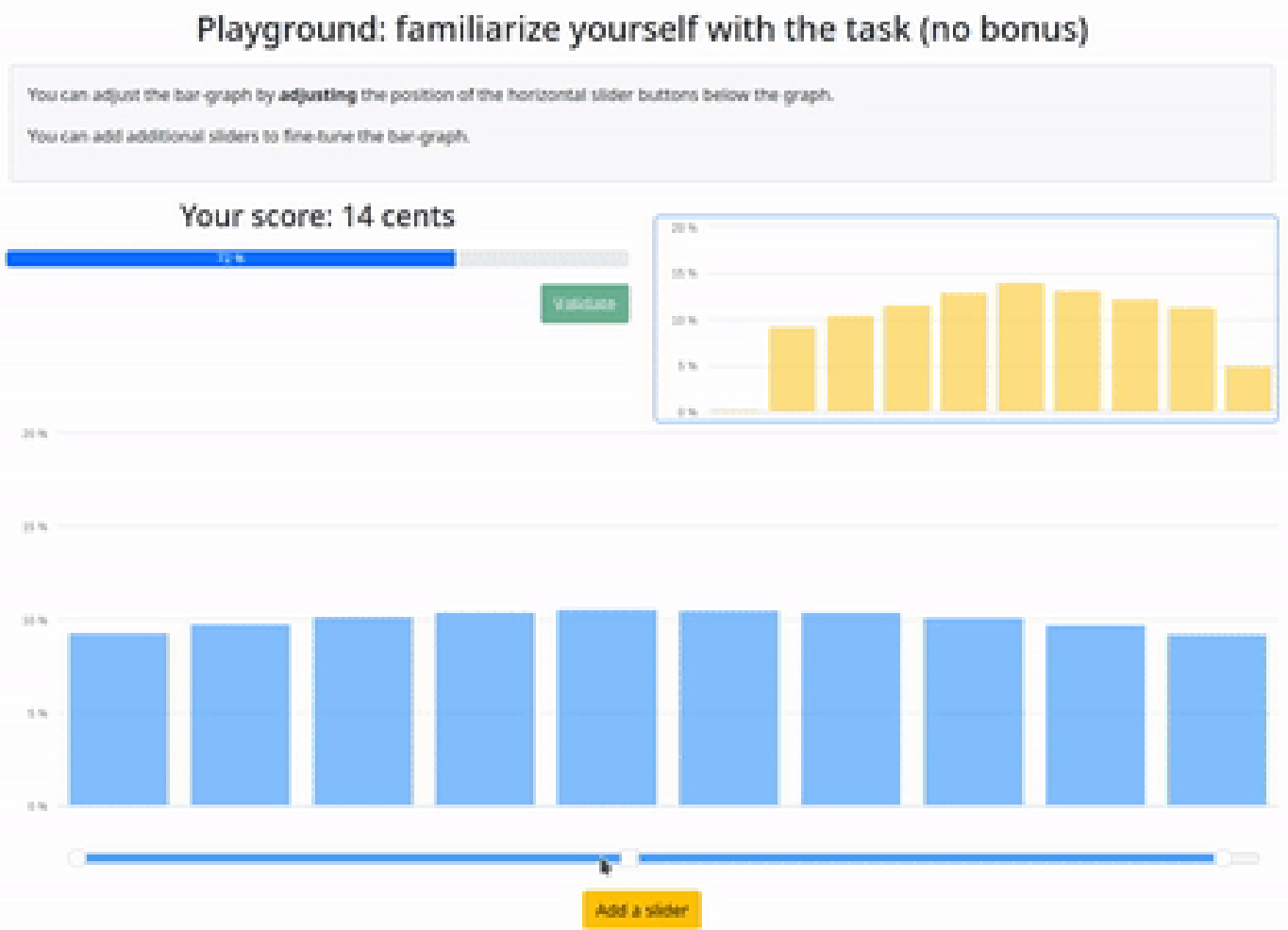
Interfaces/2: Slider



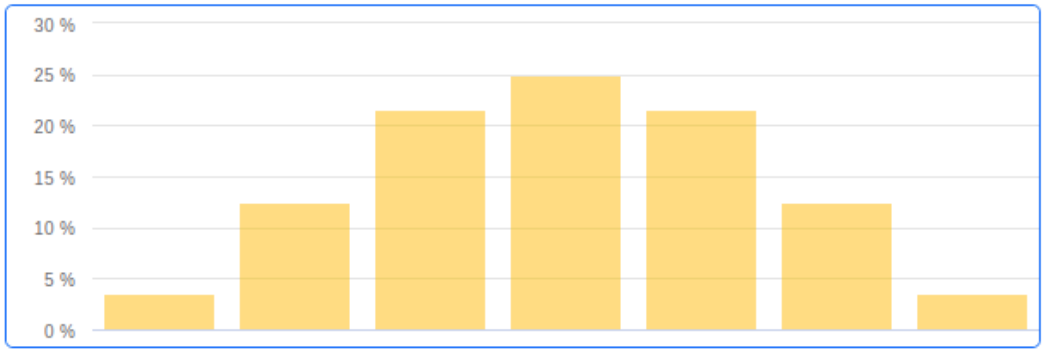
Interfaces/3: Text

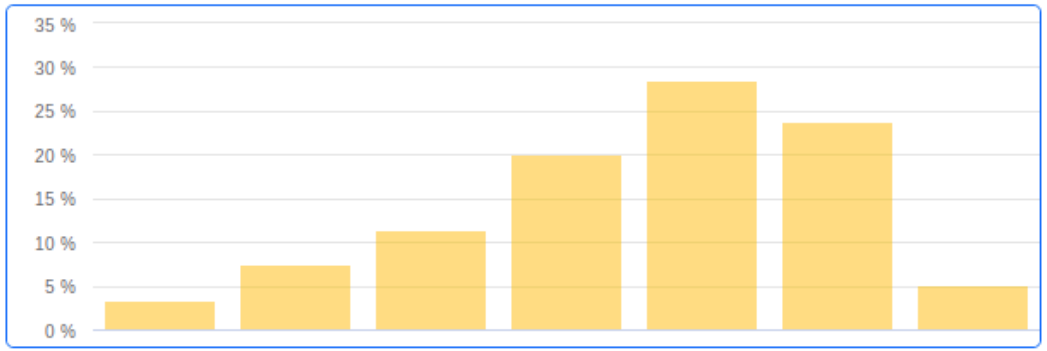


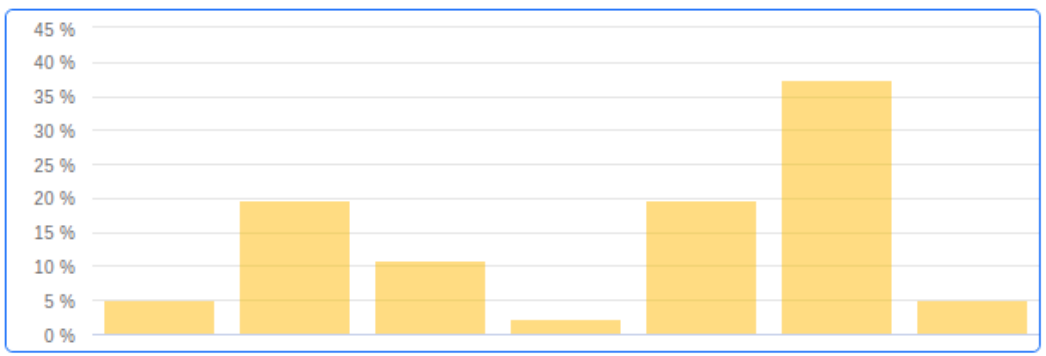
Interfaces/4: Distribution

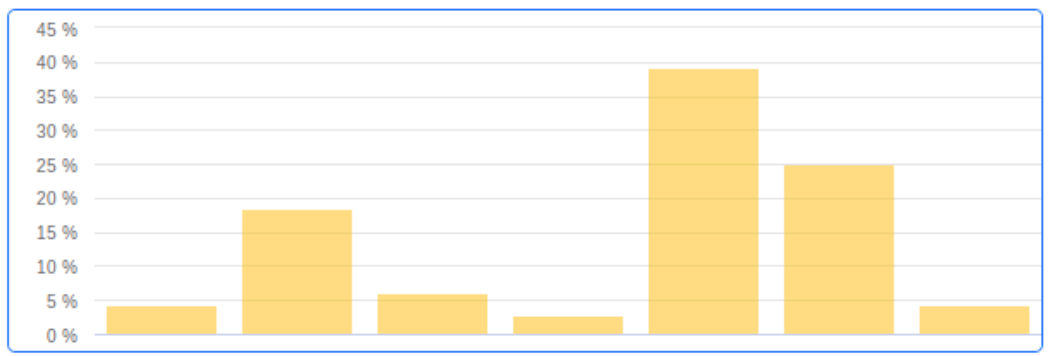


Distributions to mimic: 7 bins

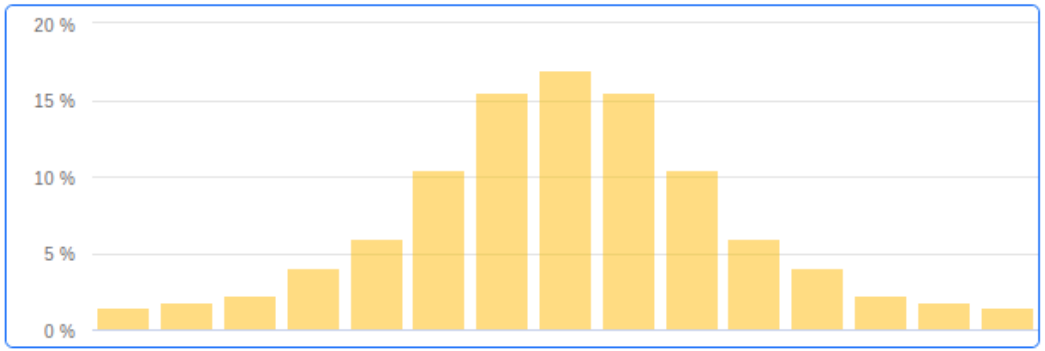


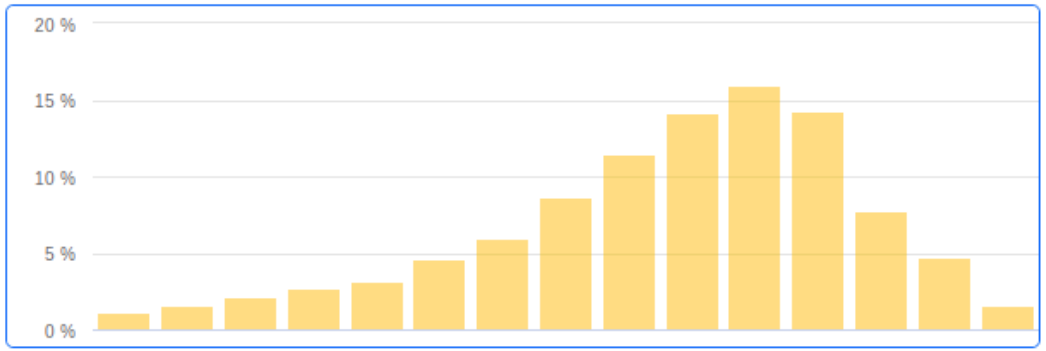


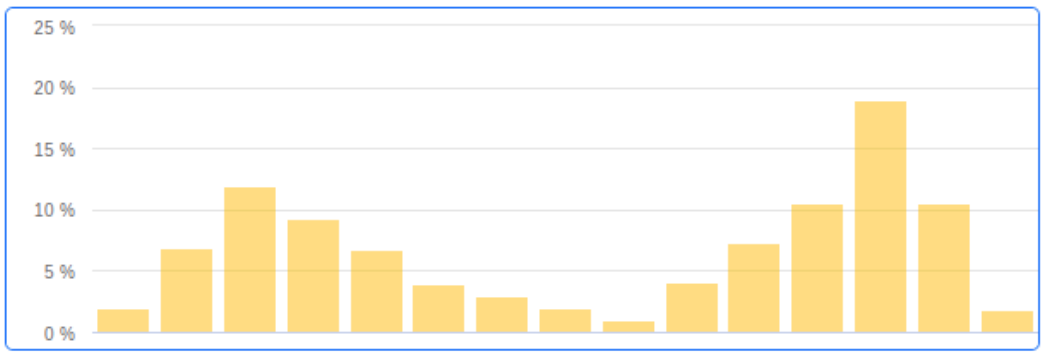


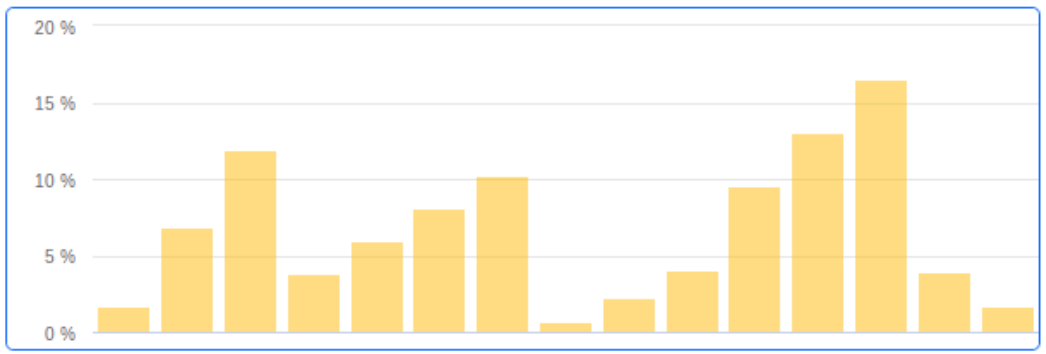


Distributions to mimic: 15 bins

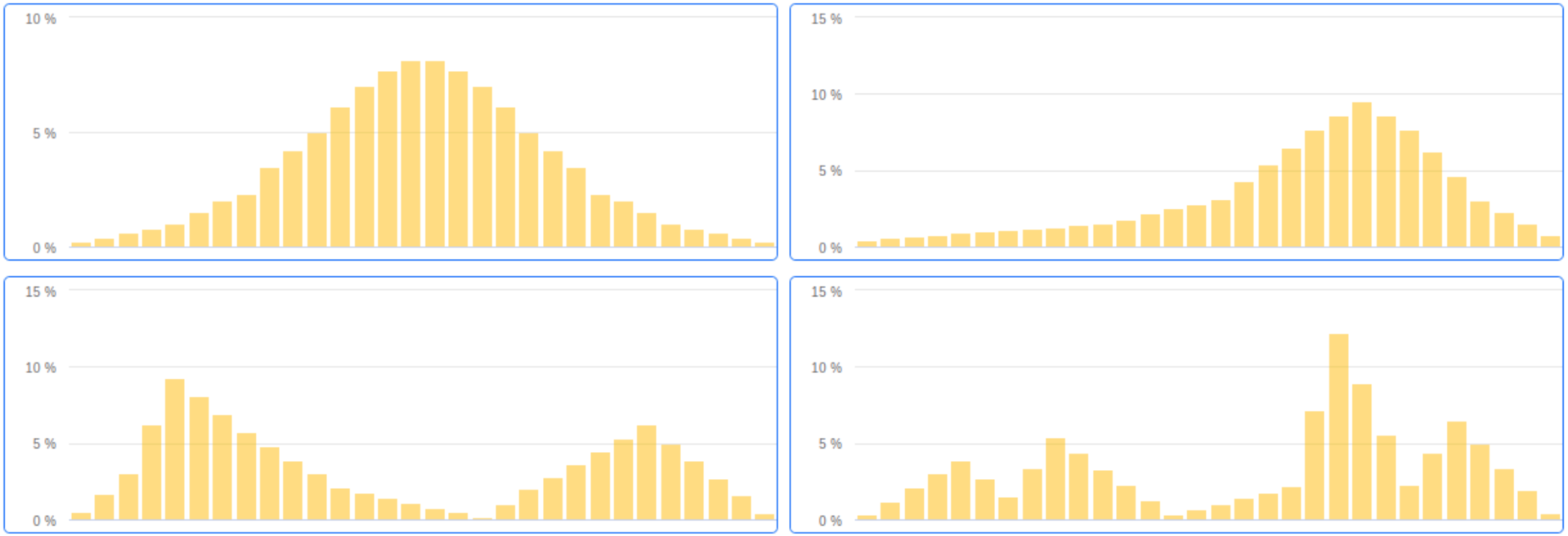








Distributions to mimic: 30 bins



Details

- **~360** Mturkers (90/arm)

- **24** screens: 3 nbins x 4 difficulty patterns x 2 time constraints

- **fixed** order: from less to more difficult / hurried

- fixed **time**: subjects cannot speed through, *must* spend 45(15) seconds

Data collected

For each screen, each subject:

- submitted **distribution** - final **distance** to target
- **path** to submission (time/accuracy of each click -> speed of convergence)

For each subject:

- age, gender
- self-reported assessment (easy? frustrating? intuitive?)

Pre-registered analysis

- **submitted distance** by screen type, shape, and time constraint

- **path of the distance** by screen type, shape, and time constraint

- **self-reported** assessments by interface

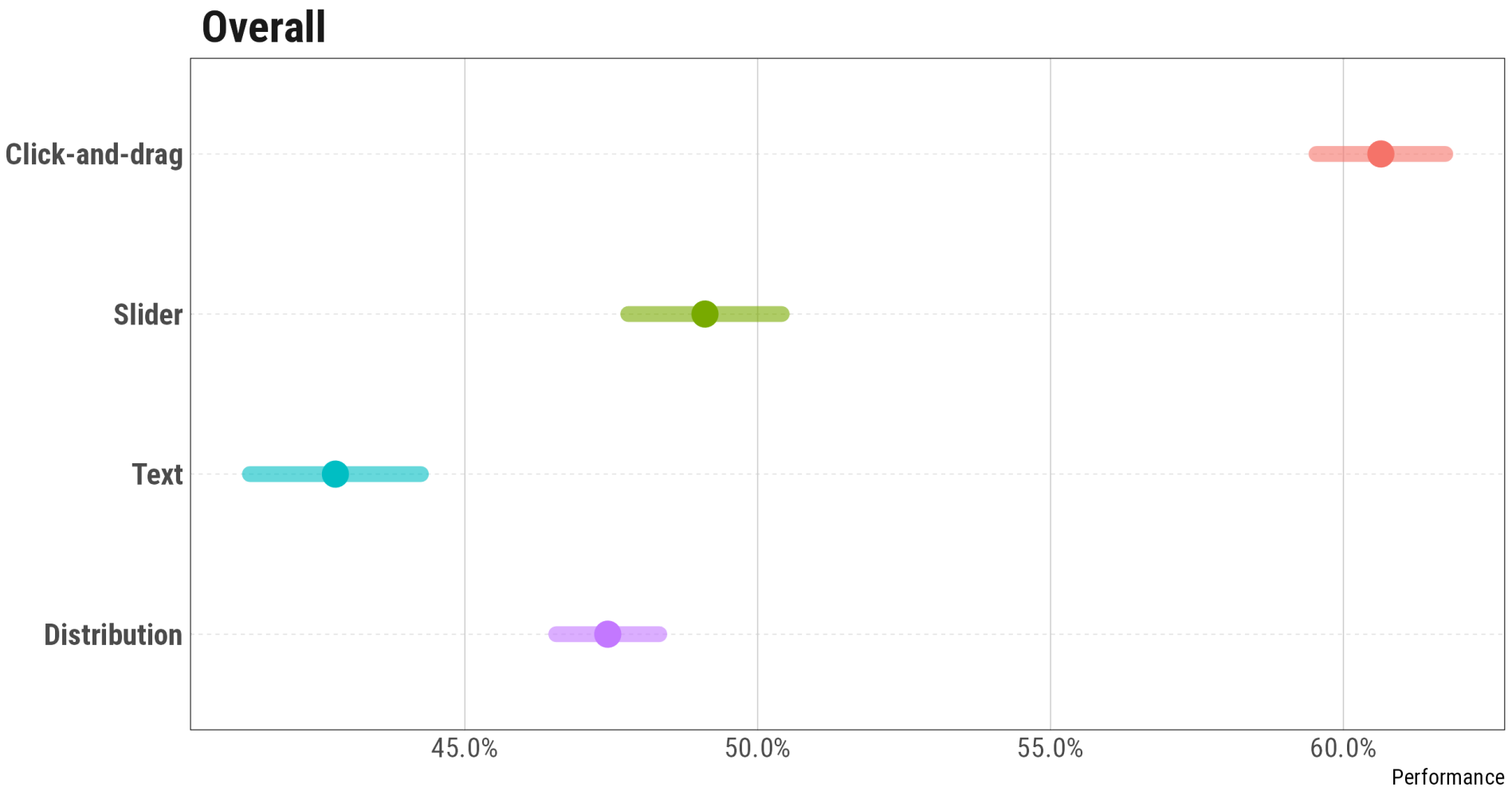
Pre-registered hypotheses

Pre-registered hypotheses

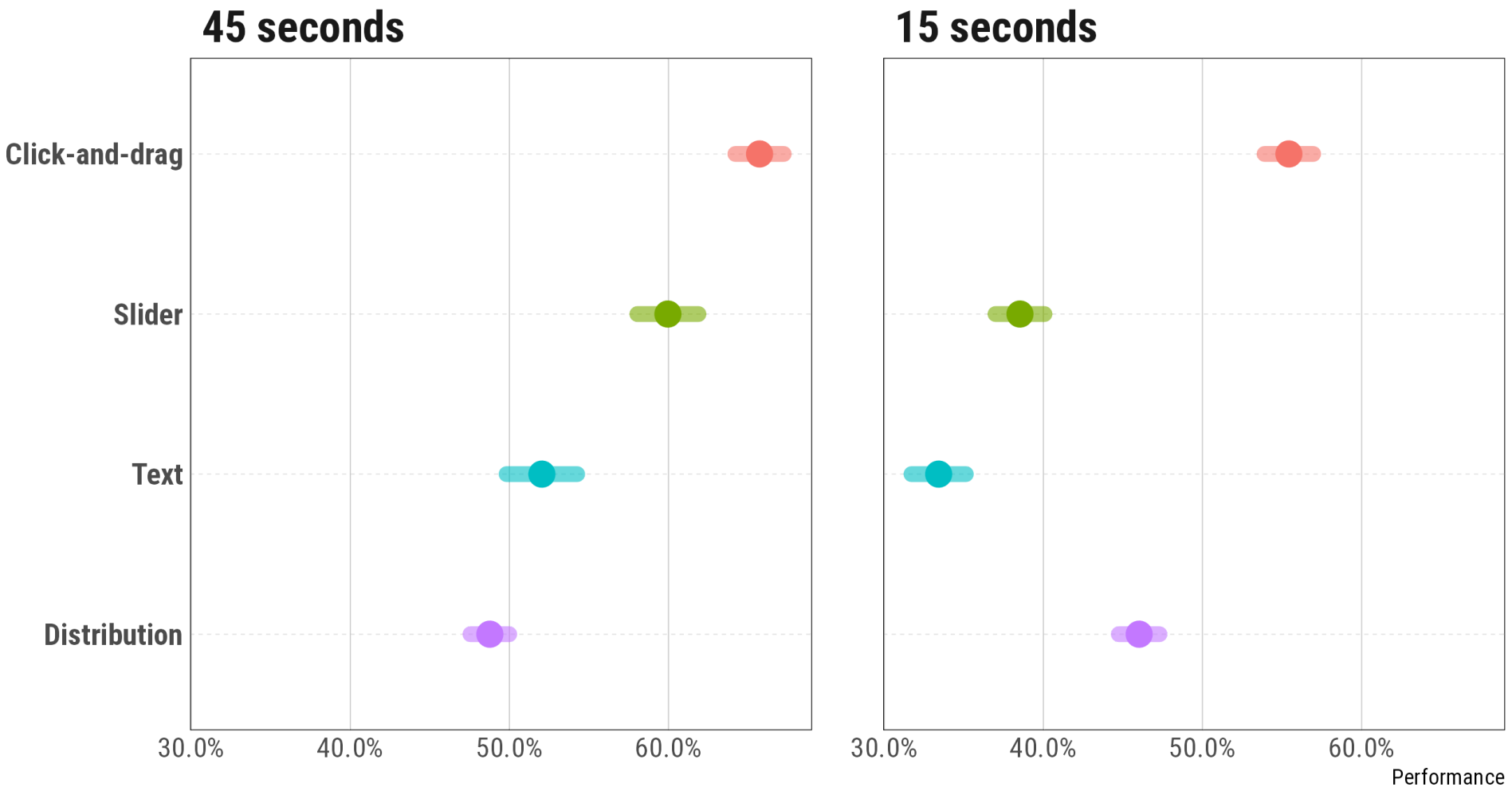
That’s simple: our interface wins in all dimensions.

Results

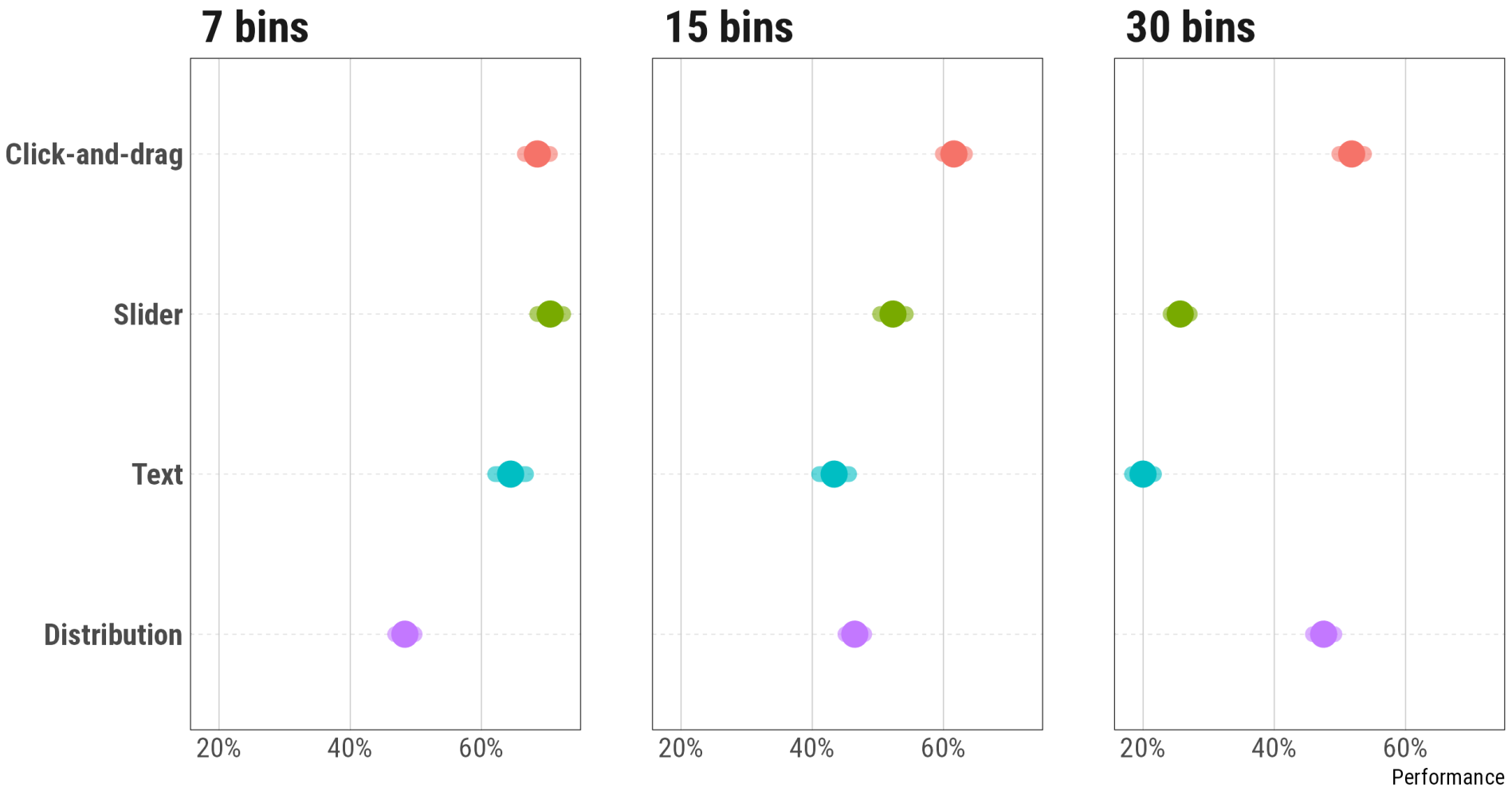
Accuracy – all screens



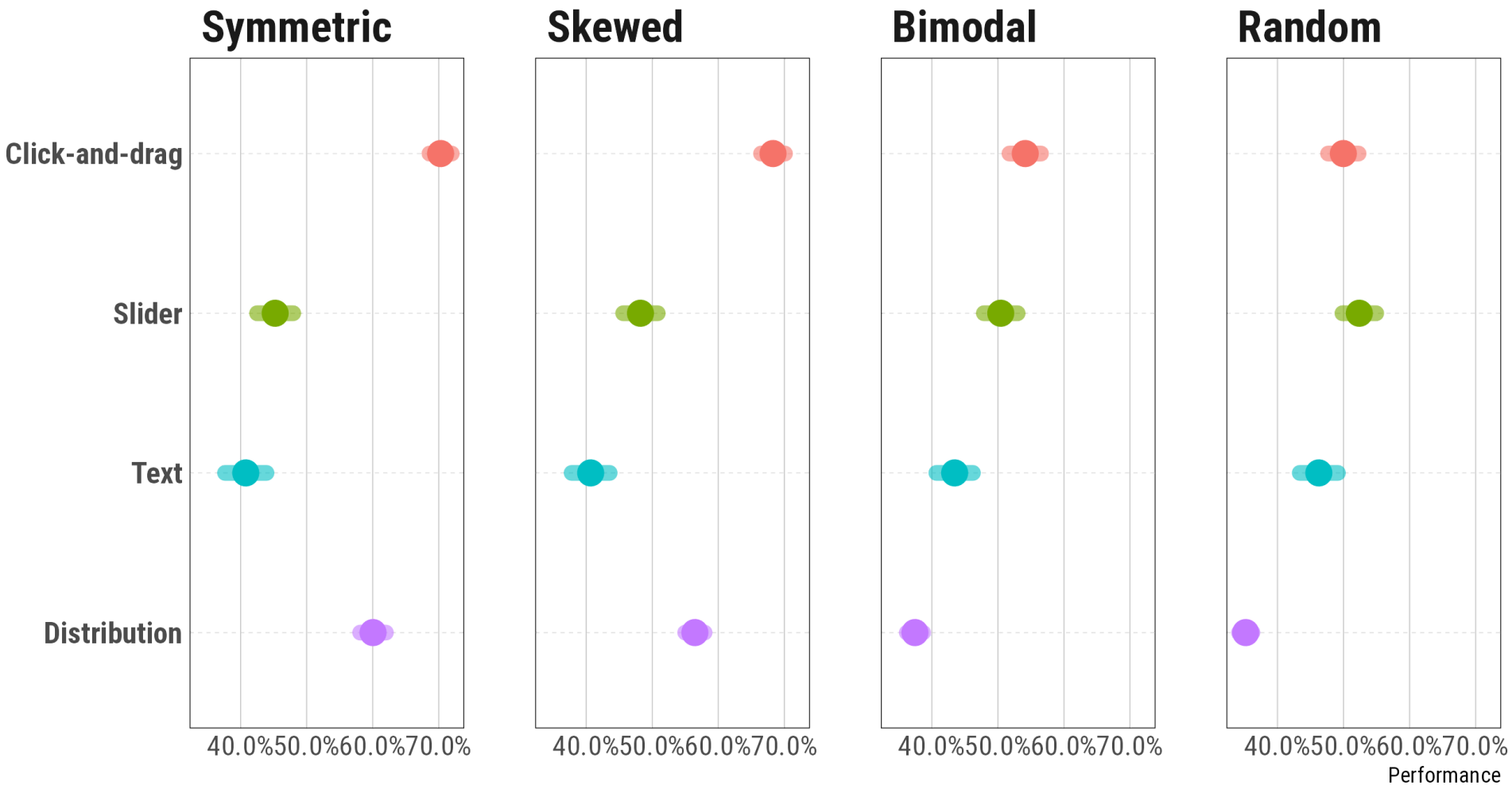
Accuracy – by time constraint



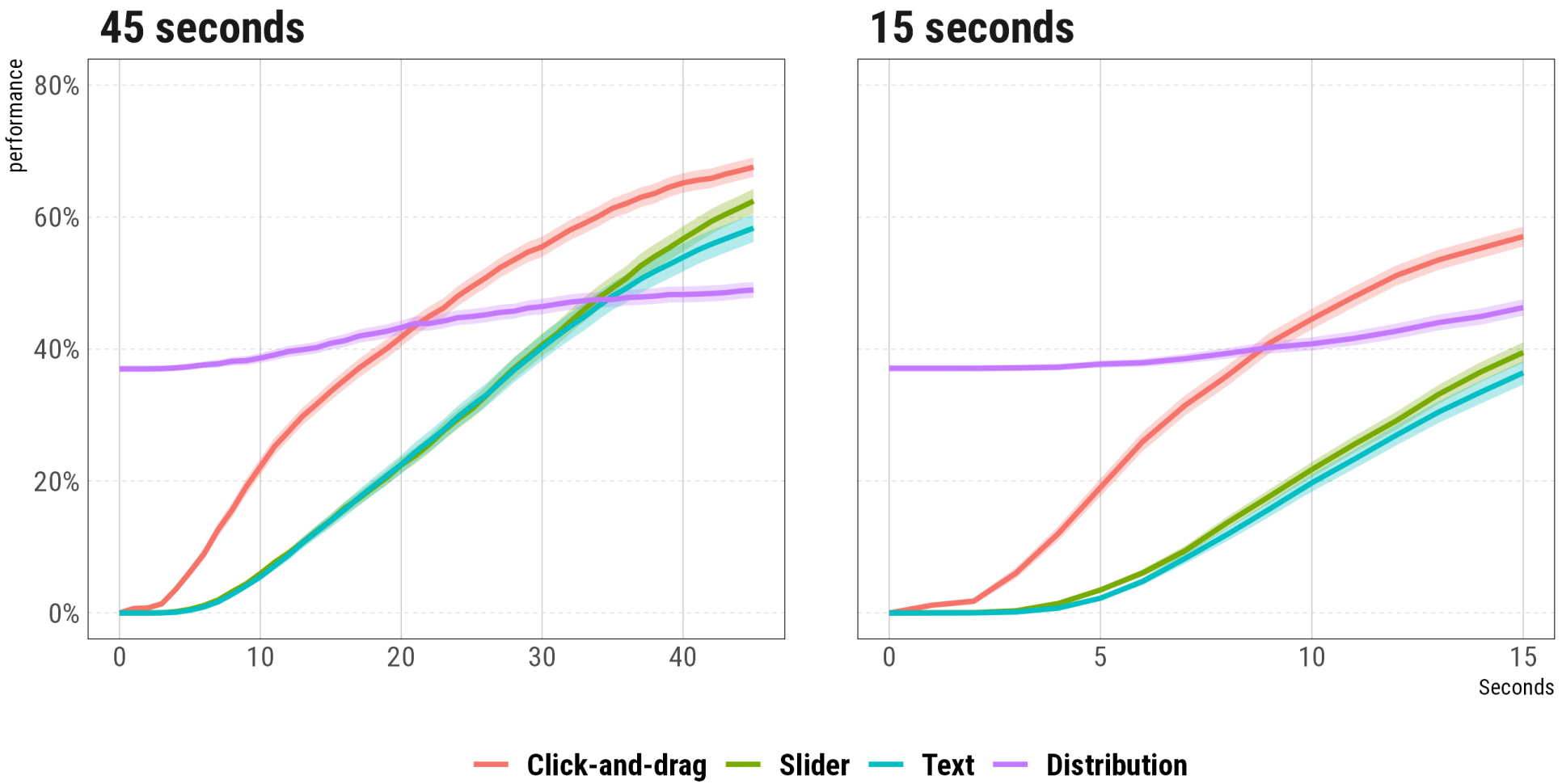
Accuracy – by number of bins



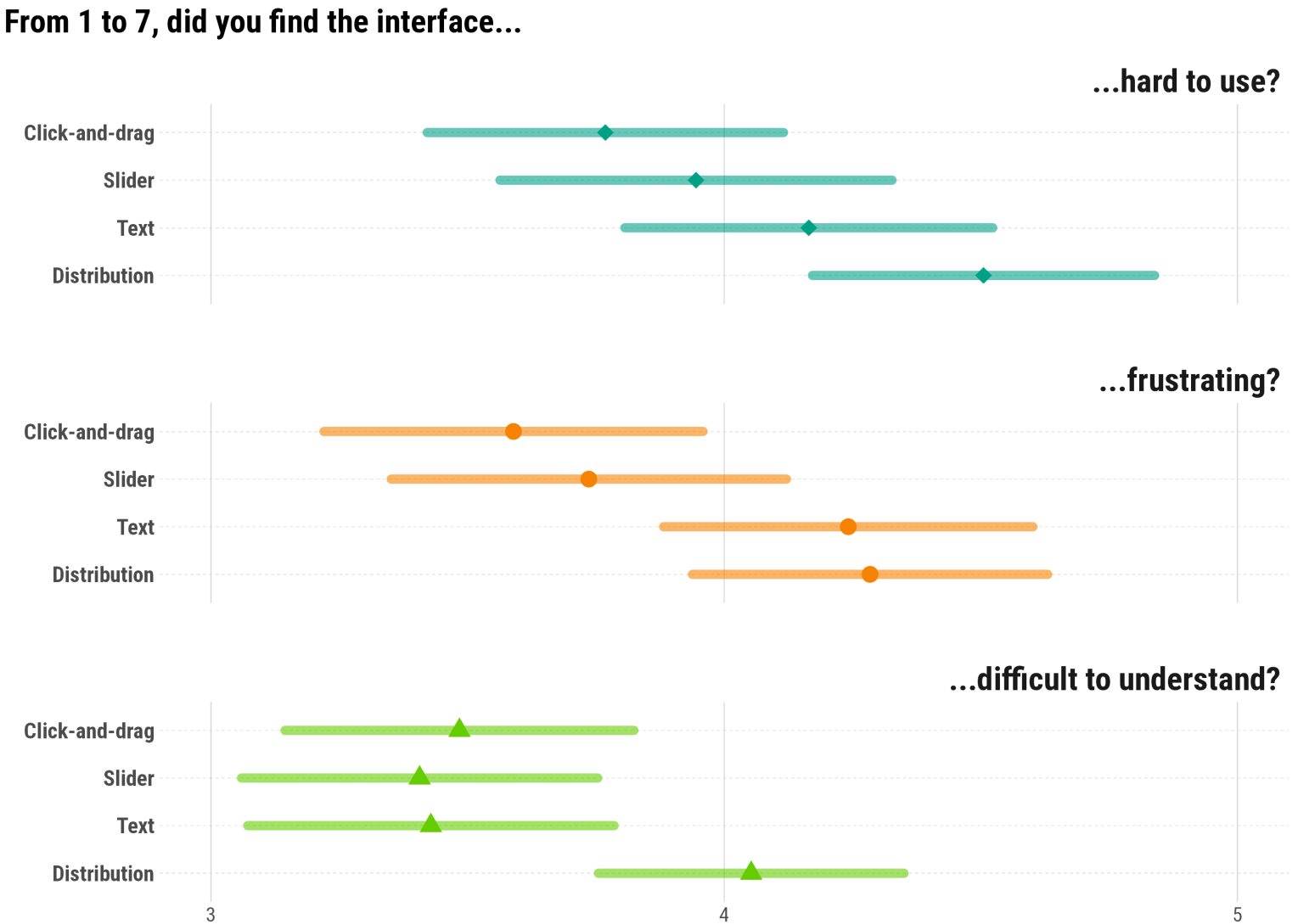
Accuracy – by distribution shape



Path of adjustment – all screens



Self-reported assessment



The interfaces in action: temperature in NYC

Eliciting beliefs: climate change

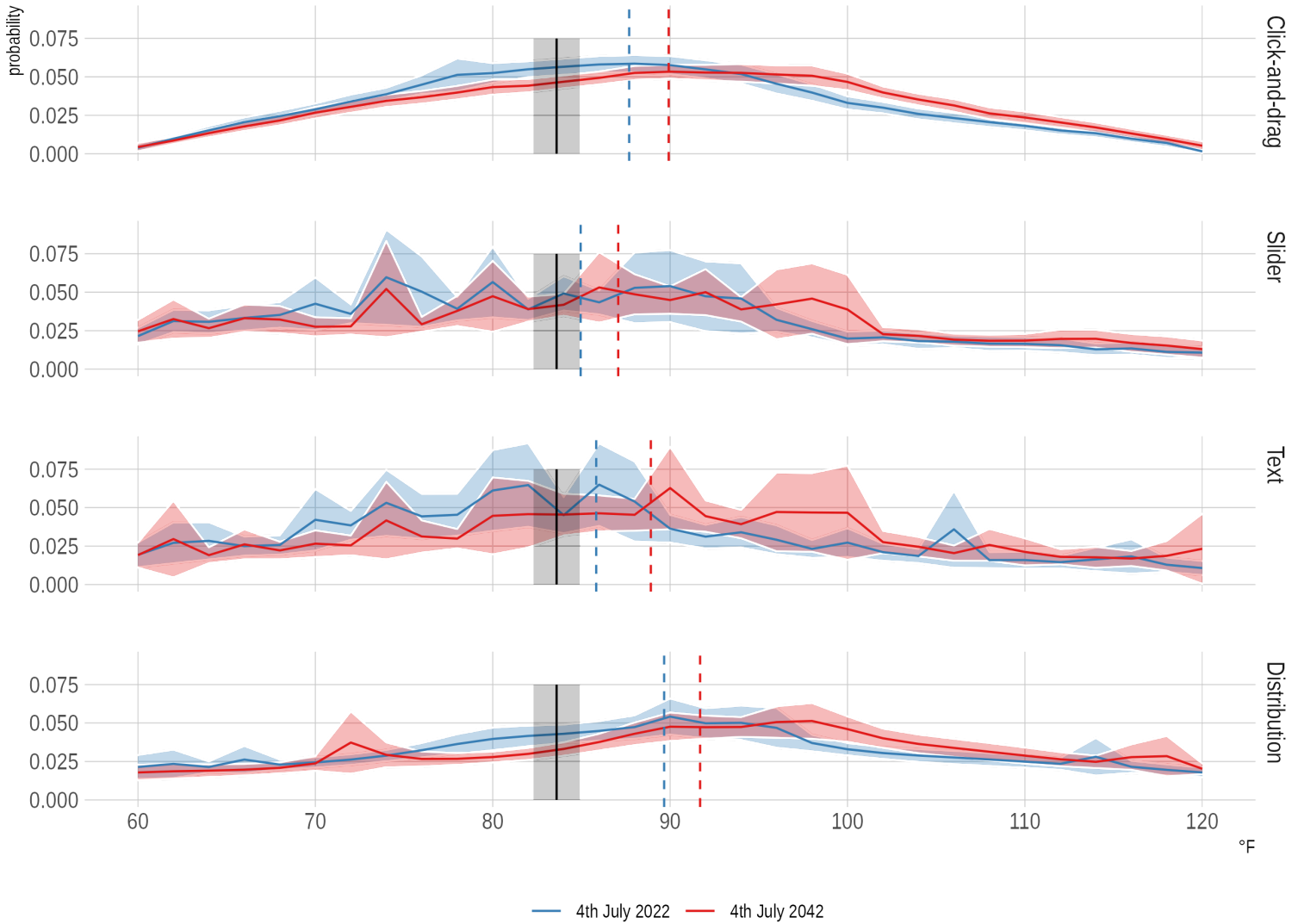
We use the 4 interfaces to elicit the *same* belief:

Maximum temperature in NYC on July 4th, 2022 and 2042

Why doing so?

- do the interfaces *impact* the elicited *beliefs*?
- can we see any *bias* introduced?
- do people hold *correct beliefs*?
- *first test* in a real, homegrown belief setting

2.5˚F warming expected in 20 years



Summing up

Summing up

Click-and-drag kicks ass. Think about using it!

Summing up

- we believe in **experimentally testing** our very tools
- not usually done in ExEc: **time to grow up**!
- Click-and-drag beats all other interfaces
  - faster
  - more accurate
  - less frustrating
  - more appreciated
- In practice, tool does not seem to matter
  - so just use the most intuitive tool

In practice

- Open data analysis tools at github

- Drop-in, minimal-configurations plugins available for:
  - oTree
  - Qualtrics
  - Limesurvey

Uses so far

> The interface is out there in the wild since June 2022. What happened?

- Agreement with **social norms*** (in progress, Bologna)
- **Political beliefs** under polarization (in progress, Bergen)

- **Hazard rate** elicitation (planned, Bergen)
- Beliefs of French farmers about **pesticide** cost & benefit (planned, Bordeaux)
- Central Bank **inflation** forecasts (planned, Bank of Norway)

Within the bigger picture

Scoring rules for forecasts horse race (**see you at SEET 2024**!)


thank
you