# Overview

# Experimental Design

The plan of this experiment is to first sort participants into groups. Each group will then be randomly assigned to either a tournament or a race. Contestants would face same prizes, same problem to solve, same duration, etc. The only difference would be that, for those assigned to a race, the first to improve a given benchmark by a fixed percentage (e.g., say 10%) wins, and for those in a tournament, the “best” submissions at the end of a given period wins.

# Motivations

Tournaments and races may provide participants with very different incentives and potentially produce different outcomes.

*Competition & Incentives*

As regards incentives to produce high-quality outputs, it is difficult to say which among these two kinds of competitions is superior. In a tournament, the quality of the output might crucially depend on the ``intensity of competition’’. Assuming participants would exert effort to maximize their probability of winning the prize, they might end up caring about their position in the final ranking but would not be necessarily worried about low-quality performances as soon as the others are doing poorly. In a race, by contrast, participants would be required to hit a certain quality level to win, that is regardless of the what the other performances look like; although they may not want to improve their performance too much beyond the required level.

Competition & Timing

While a race would obviously impose participants to exert high effort from the first day, this is not necessarily the case in tournaments.

Competition & Feedbacks

If partial scores are disclosed (e.g., through a leaderboard), a tournament might induce contestants to hide their progress until the last day of the competition. . Here the economic rationale would be that an early disclosure can spur higher effort from rivals. In a race, by contrast, any attempt of reasonable quality should be submitted (and so disclosed) as soon as possible.

## Measures of Interest

Individual Performance (quality and diversity of solutions)

Individual Participation (number and timing of submissions)

## Detailed Description of the Race

Pre-contest phase

We hire an expert coder to assess reasonable quality levels that can be achieved by the crowd in a 2 weeks period and under the promise of a reward of XXXX

In practice, we need 2 or 3 quality levels or percentage improvements with respect to the benchmark:

* Lower level should be achievable by a motivated ``yellow’’ coder within one or two weeks
* Top level should be something that might be challenging for a ``red’’ coder but still feasible within a one or two weeks period

Contest Phase

A max period of two weeks

The competition ends sooner as soon as all levels have been hit.

A coder can win at most one prize.

In case a participant turns out to be the first in reaching more than one quality thresholds, the participant will be awarded the max prize. Other participants capable to hit the threshold will be still able to be awarded the unassigned prizes.

The leaderboard shows the provisional score.

We provide no feedback on the final score and the winners are not shown to contestants unless all quality levels have been hit.

If no one hit the threshold

??

Adopt measures against cheating

Full monitoring ex-post (to get diversity measures as well)

Make groups of different size

Method to deal with the rest

# Comments from call #SCRIPPS March 20

I think I can fully understand the value of engaging the crowd in solving the problem on a meaningful dataset rather than on the one produced by m-turkers, which is of uncertain quality.

Yet, inasmuch our objective is to compare the crowd outcomes with the industry gold standard (i.e., banner), this might not be judged a key issue after all.

I mean, the fact that the actual data came from m-turkers (instead of experts) is not going to change the underlying NLP problem. Consequently, the crowd might not be affected at all in the exploration of the space of algorithm solutions.

One can also think that once a general technique has been developed or improved, a second step would be to apply it to more meaningful datasets.

So overall I think that we should:

* Use (possibly ``low-quality’’) data from m-turkers for the training/validating/testing dataset (pretending they were doing their job as experts)
* Train BANNER on m-turkers data and use it as the benchmark for the competition
  + This might also suggest a clear link between m-turkers and experts
* Then set up a contest where the crowd would be competing on an equal footing with BANNER to improve predictions
* Once the contest is over, use the winning solutions on the dataset made by experts