Races vs. Tournaments. Field Experimental Evidence from Online Competitions.

Andrea Blasco, Kevin Boudreau, Karim Lakhani, Michael Menietti,

# Experimental Design

The plan of the experiment is to sort participants into groups homogenous by distribution of skills, experimental task, and by prize pool. The only difference would lie in the criterion used to rank order efforts at the end of the competition. This in turn would affect the dispensation of prizes and therefore the strategic interactions among participants.

In half of the groups, submissions below a fixed quality level *Q>0* are ranked below the others and do not qualify for prizes. Submissions of quality above this level are rank ordered by *time* *of submission*: earlier submissions are ranked higher.

In the other half, submissions are rank ordered just by *quality.* That is, regardless of the time of submission, solutions of higher quality are also ranked higher.

We name the first setting a *race* and the second a *tournament*.

# Motivations

Of course, there are a number of possible mechanisms through which a race could differ from a tournament (self-selection, effort decisions, timing of entry, dynamic competition, the role of feedback, etc.). This experiment would allow us to remove differences due to self-selection into groups (an other interesting question, do participants prefer a race or a tournament?), to focus on how strategic interactions can affect outcomes. To study this issue we need a theoretical framework to better interpret the observed outcomes.

Under the simplifying assumption that competitors would only care about being awarded the final prizes, Moldovanu and Sela (2001) have developed a game theoretic model of tournament competition in which *n* agents simultaneously choose a costly bid *q* to top the rivals and win a prize of value V>0. In their model, agents are *strategic* and choose the bid on the basis of their expectations about the abilities of their opponents. This can be seen as both a weakness and a strength of a tournament setting. In fact, when competition is relatively high. Expected quality is also high. By contrast, when competition is relatively low, expected quality is low. In other words, outcomes tend to follow closely some sort of average *intensity of competition*, which is parameterized in the model by the distribution of abilities.

However, this cannot happen in a race where participants do not compete on quality but on the time to deliver. To study this setting in detail, we extend the MS model to encompass a race setting. Now participants have to make two decisions: the quality of the costly bid *q* and some costly effort *t*to reduce the time to deliver. Again the intensity of competition is an important driver of decisions. But it is not crucial ffoOf course, if someone enters the competition, then he has no incentive to deliver a quality level below or above the threshold. So the max quality is simply the thresholds, conditional on entry.

**Overall, it is not clear whether a race would produce higher max quality than a tournament. From the model, it seems that outcomes in a race can be higher when the (expected) intensity of competition is somewhat low (e.g., few participants, highly skewed distribution of abilities). Otherwise the tournament seems a better**

## Setting the right threshold.

In a race the sponsor is required to establish (and commit to) a quality threshold before the contest starts. This activity involves taking some risks. However it does not seem very different from deciding on the extent of the prize pool. In fact, one criterion could be to look at a relative measure, as if asking the following question: “how much I want to pay to improve my gold-standard solution of 1% (if there’s entry in a 2-week race)?”. In a tournament one should ask the same question but in expectation terms: “how much I want to pay to improve my gold-standard solution by x% on average?”.

## Competition & Feedbacks

One limitation is that we don’t have dynamic model. Participants don’t take simultaneous decisions, because they could benefit form feedback generated during the competition by others.

For example, 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

The unit of analysis.

Individual Performance (quality and diversity of solutions)

Individual Participation (number and timing of submissions)

## Details of the Experiment on TopCoder

**Pre-contest phase**

A crucial aspect of the race is setting the right quality level. If the level is too low, then all tournaments would dominate the race. If the level is too high, then we might have too few participants submitting solutions and again it would be difficult to make a comparison with the tournaments. So my initial proposal was to have two thresholds (easy and hard). The room winners would have different prizes depending on hitting the easy or hard threshold. So to reduce the likelihood of ending up in one of these two limiting cases. The alternative would be to have one single threshold for the two winners.

Regardless of whether we choose to have one or two thresholds, we still need to pin down one (or more) quality level for the races. To do that we will use Rinaat’s Brain plus we send invitations to some expert coder to “bootstrap” the problem and to come up with a reasonable number.

**Contest Phase**

|  |  |  |
| --- | --- | --- |
|  | Tournament | Race |
| Advertising |  |  |
| Submission Period | Two weeks. | Two weeks.  The submission period stops if 2 members hit the quality threshold. |
| Prizes | Top 2 ranked solutions by quality.  #1 prize > #2 prize | First 2 solutions to hit the quality threshold.  #1 prize > #2 prize |
| Provisional Score | The leaderboard shows the provisional score. | The leaderboard shows the provisional score.  In the background we compute the final score.  If someone hits the threshold we send notification to all participants in that group. |
| On each submission | Pop-up asking how many hours worked. | Pop-up asking how many hours worked. |
| Group size | Groups of different size: 7 and 30 | Groups of different size: 7 and 30 |
|  |  |  |
| Prevent cheating | Ex-post we control winning submissions | Ex-post we control winning submissions |

# 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

In this paper we experimentally two alternative ways to reward effort: rank order tournaments and races.

Since recently, have become quite popular for providing incentives in employment relationships.

A *tournament* is a contest in which prizes are awarded to the top N contestants as ranked by the quality of their solution.

A *race* is a contest in which prizes are awarded to the first N contestants to hit a certain quality level.

Assuming no feedback, contestants will be exerting effort if the expected time to develop a solution is sufficiently low or equivalently the cost of speeding up production is sufficiently small.

As soon as top coders face lower costs of production, they should enter the competition.

On the one hand, the fastest contestant may not be the one also capable of providing the best quality. For example, top coders may have more constraints on the time to spend on a competition.

We plan to gather data on

1. Number & quality of the solutions submitted.
2. Self-reported measures of hours worked.

## Possible outcomes

|  |  |  |
| --- | --- | --- |
| **Race\Tournament** | Someone’s solution is above the required level | No solution is above the required level |
| Someone’s solution is above the required level | (A) Race < Tournament? | (B) Race > Tournament? |
| No solution is above the required level | (C) Race < Tournament? | (D) |

If (A) occurs, this implies that competition in the tournament was not very intense. The race

# Experimental Design

* Registration phase.
* Split into groups:
  + If sample size is large enough, we create groups of different size to study variation in group size.
  + Randomize by rating? Or experience?
* **Race**
  + **The first 2 to hit a quality threshold are awarded a prize.**
  + **Evaluation of solutions** 
    - Participants receive feedback based on the testing dataset.
    - In the background we also compute the final score.
    - As soon as someone hit the threshold, we send a notification.
* **Tournament**
  + **The top ranked 2 are awarded a prize.**
    - Minimum quality level for their solutions?

# Literature Review

## Theory

* Moldovanu & Sela (2001)
  + We extend to incorporate a race.
* Harris & Vickers (ReStud, 1985) “Racing with uncertainty”
  + Interplay of *uncertainty* in the outcomes of effort and *strategic interaction* between competitors.
  + Leaders make greater efforts as the gap with the followers widens.
* Fudenberg, Gilbert, Stiglitz, Tirole (EER, 1983)
  + When races are neck-to-neck and when degenerate into monopoly?
* Zizzo (IJIO, 2002)
  + lab. experiment on H&V predictions.
  + Not as predicted!
* Baye & Hoppe (GEB, 2003)
  + Tullock contest function races are equivalent to tournaments.

## Surveys

* Konrad (book)
* Decheneaux, Kovenock, Sheremeta (2012)
  + Dynamic contest: one paragraph on races. Mainly Zizzo’s results.

## Something to read

* Hoppe & Lehman-Grube (JET, 2005), “Innovation timing games: a general framework with applications“