Non-Cooperative Behaviour and Acceptability Judgement Tasks

Non-cooperative behaviour reduces the quality of experimental data (cf. e.g., Kazai et al., 2011). This issue affects any behavioral task and (online) judgement studies in particular. For instance, Downs et al. (2010) report that up to 39% of their participants were non-cooperative. However, very few (syntactic) judgement studies employ strategies to discourage or detect non-cooperative behaviour: A Google Scholar query (<syntax “acceptability judgment task”>) returned 49 papers that were published in 2014, were available through our university networks, and included a syntactic judgement task. Of those papers, not a single one reports the use of some kind of reaction time strategy to exclude subjects, only 5 use “booby trapping”, and only 5 use other detection strategies.

We argue that implementing two detection strategies, booby trapping and a median-based reaction time criterion, is critical to detect non-cooperative behaviour. We also present an on-line warning mechanism that discourages non-cooperative behaviour. The following is based on data from three experiments using Amazon’s Mechanical Turk (12 sessions, including 422 subjects, all of which had completed >5000 tasks and an approval rate of >98%).

The first detection strategy is *booby trapping*, i.e. the inclusion of non-critical items whose status is well-established (e.g. clearly acceptable, clearly unacceptable). Participants who fail to distinguish “bad” and “good” items should be excluded. In our experiments, this led to the exclusion of 4% of our subjects.

Extreme reaction times point to non-cooperative behaviour and Figure 1 and Figure 2 illustrate why the *median reaction time* is more effective for such non-cooperative behaviour detection. The median is less vulnerable to “clever spammers” who adjust their mean reaction time by pauses. “Simple spammers” are detectable by both mean and median reaction times, whereas clever spammers can only be detected by his/her median RT. We suggest to exclude participants whose median reaction time is 1.5 standard deviations below the mean of all participants’ median reaction times. In the example, this would lead to the exclusion of all four spammers. In our experimental sessions, a median-based criterion more than doubled the detection rate of non-cooperative participants (a mean-based criterion detected 11 spammers, a median-based criterion detected a total of 26 spammers).

Implementing these detection criteria has a real effect on the results. The mean ratings by non-cooperative participants significantly differ from those by cooperative participants for 9 of our 12 sessions (Wilcoxon Signed Rank Tests: $p < 0.05$) and their variance across ratings was significantly lower, too (F-Tests: $p < 0.05$).

However, the most effective way of dealing with non-cooperative behaviour is to discourage it. We do so by letting participants know that we are able to detect non-cooperative behaviour. To this end, two of our experiments included an *on-line warning mechanism* that produced an alerting pop-up window when a participant’s reaction times repeatedly fell below 400 ms (cf. Figure 3). This measure reduced the rate of (detected) non-cooperative behaviour from 14.1% to 7.1%, slashing detected non-cooperative behaviour in half.

We conclude that booby trapping and a median based reaction time criterion are effective strategies to detect and exclude non-cooperative behaviour. Further, including some kind of on-line warning mechanism is an effective way to discourage non-cooperative behaviour in the first place. While implementing these strategies requires that the researcher adjusts his/her experimental design, it also improves the quality of one's data significantly.

References

Downs, J. S., Holbrook, M. B., Sheng, S., Cranor, L. F.. Are Your Participants Gaming the System?: Screening Mechanical Turk Workers. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 10): 2399-2402.

Kazai, G., Kamps, J., Milic-Frayling, N., 2011. Worker Types and Personality Traits in Crowdsourcing Relevance Labels. Proceedings of Twentieth International Conference on Information and Knowledge Management (ACM CIKM): 1941-1944.

Figures

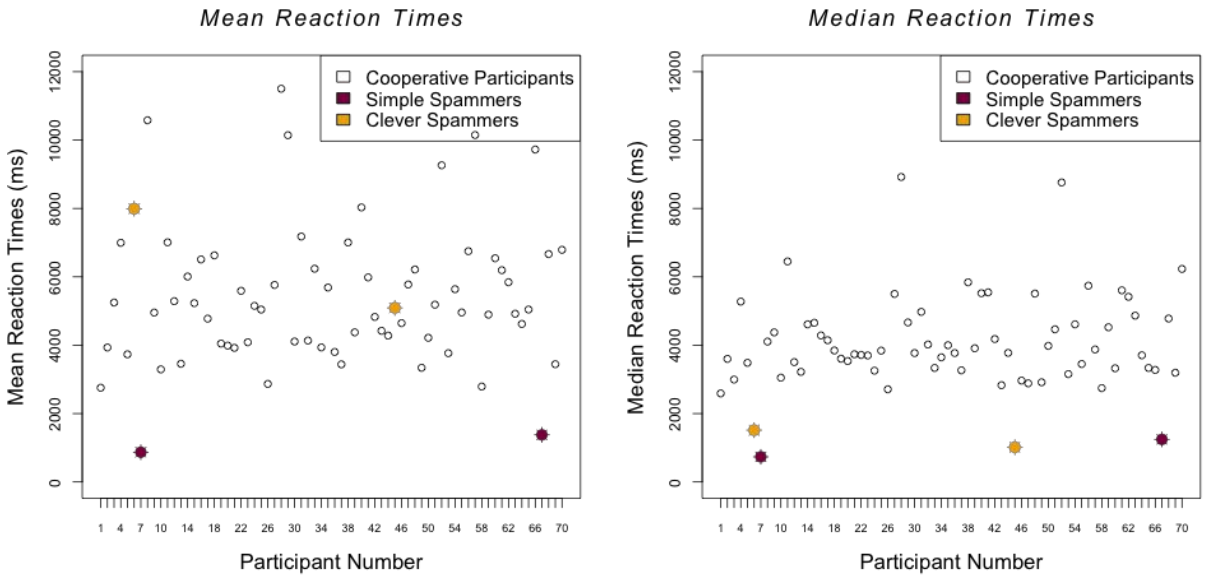

Figure 1: Distribution of mean response times (left) and median response times (right) (both from Experiment 1). While a mean-based exclusion criterion is not able to detect clever spammers, a median-based criterion is able to do so.

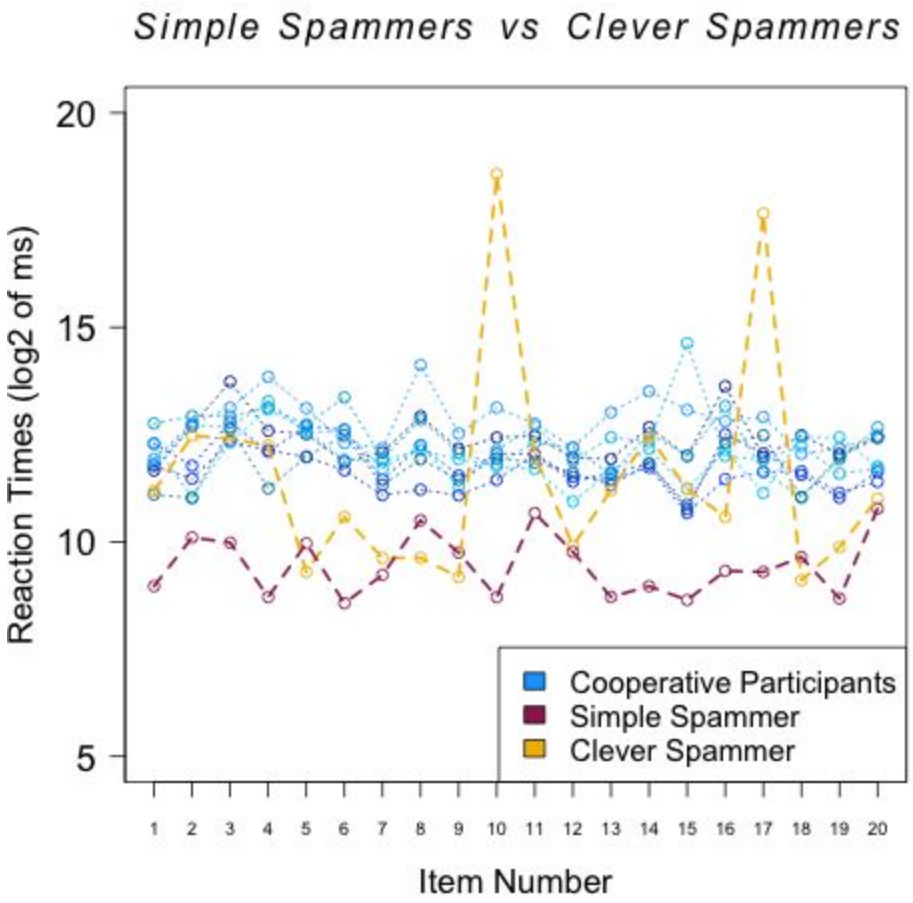

Figure 2: Response times of eight cooperative and two non-cooperative participants (from Experiment 1).

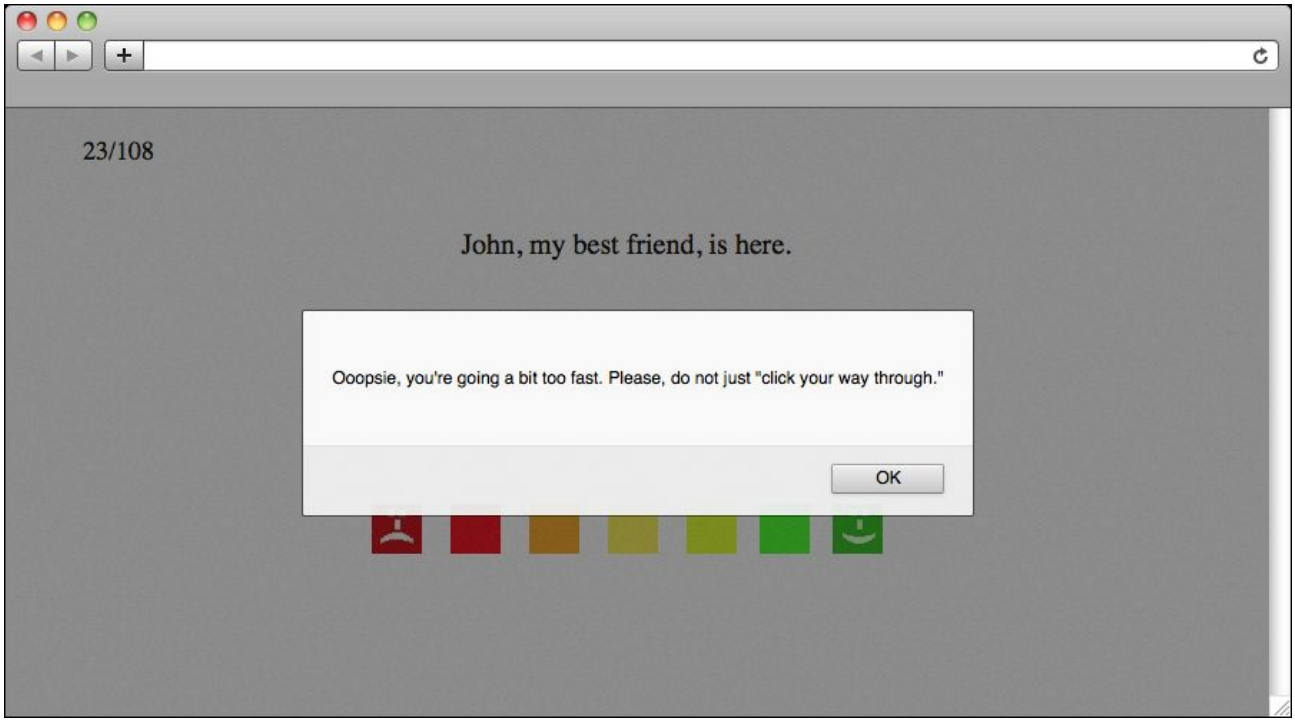

Figure 3: An illustration of the on-line warning mechanism.