Making a Gamble: Recruiting SE Participants on a Budget

Madeline Endres
University of Michigan
Ann Arbor, Michigan, USA
endremad@umich.edu

Westley Weimer
University of Michigan
Ann Arbor, Michigan, USA
weimerw@umich.edu

Amir Kamil
University of Michigan
Ann Arbor, Michigan, USA
akamil@umich.edu

ABSTRACT

Human studies in software engineering, especially those on a budget, often struggle to recruit participants. We investigate allocating a $100 budget for a remote program comprehension study to maximize the number of high-quality responses. We compare seven incentive structures, including various raffle-based and first-come-first-serve methods. We focus on computer science undergraduates, a common population in software engineering studies. Incentive structure does have significant effects on the number of participants and on data quality. We conclude with concrete guidelines for incentive allocation for online software engineering human studies.

KEYWORDS

participant recruitment, computer science, incentives

ACM Reference Format:
Madeline Endres, Westley Weimer, and Amir Kamil. 2018. Making a Gamble: Recruiting SE Participants on a Budget. In *1st Workshop on Recruiting Participants for Empirical Software Engineering, 2022, Pittsburgh, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recruitment is one of the most challenging aspects of human subjects research in software engineering [1, 5, 7]. For instance, Buse *et al.* found that recruitment was a barrier for conducting human evaluations for almost 60% of software engineering researchers. One method commonly used to increase recruitment rates is monetary compensation [10]. However, many software engineering researchers may lack the budget for large numbers of participants.

In this paper, we report the results of a controlled experiment on the effect of seven different financial incentive structures on recruitment given a fixed budget of 100 US dollars. The incentive structures vary by both incentive method (drawing vs. first-come-first-serve) and number of possible monetary awards. Studies comparing incentive structure efficacy have been conducted in other fields [6]. However, to the best of our knowledge, this is the first such experiment of computer science participants.

We find, given a fixed participant recruitment budget, that the incentive structure has a significant effect on amount of data collected: we obtain up to 3× more valid participants from the best incentive structure vs. the worst one. Additionally, we find that data quality is higher for raffle-based approaches: of those participants who start the survey, drawings result in a higher proportion of usable and complete data than do first-come-first-serve structures.

RoPES, 2022, Pittsburgh, USA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 EXPERIMENTAL DESIGN

We controlled the recruitment conditions while carrying out a previously-published program comprehension survey (see Endres *et al.* [3]). The original study was on the impact of iterative or recursive framing on programming performance. The study took around 30 minutes to complete, and asked participants to write the output of 6–7 functions. It also asked participants to complete a short spatial reasoning test and a demographics questionnaire. The study was aimed at undergraduates at the University of Michigan who had taken a computer science course in the last few semesters. Recruitment took place exclusively over email, and all emailed students members were eligible to participate in the study. Participants were unaware that there were multiple different incentive structures. In total 5,639 emails were sent to potential participants.

To understand the impact of incentive structures on recruitment with a limited budget, we randomly divided the participant recruitment pool into seven groups. For each group, we allocated US$100 for incentives. For all groups, the recruitment procedure was the same other than the incentive: incentives varied in the maximum possible compensation amount per participant and were either structured as a drawing or as first-come-first-serve for a finite number of gift cards. Table 1 describes all seven conditions.

As all groups had the same fixed budget, the number of possible awards is inversely related to the potential award amount (e.g., two $50 awards vs. one $100 award). However, both groups 3 and 4 had 10 $10 potential awards, allowing us to directly compare the two incentive structures with the same monetary properties.

3 EXPERIMENTAL RESULTS

We organize our analysis around the following questions:

(1) Does the rate of usable data differ significantly by incentive type?
(2) Do drawings or first-come-first-serve yield higher quality data?
(3) How does the maximum monetary value listed in a recruitment communication relate to the amount of usable data?

RQ1—Which incentive was best? Incentive structure *does* significantly effect the number of valid data points obtained (Table 1). We conducted a 2×7 χ^2 test comparing the proportion of the participant pool that provided valid data for each incentive. The χ^2 test result, $p = 0.0001$, is well below our significance threshold.

To investigate which incentive conditions drive this result, we compute adjusted standardized residuals following best practices [9]. We then converted these z-score values to p-values. After correcting for multiple comparisons using the Holm adjustment [2], we find **participants provided significantly more valid data when**

ID	Group	Description	Size	Started	Finished	Valid	% Valid	χ^2 residual	p-value
1	F_$1	First 100 valid participants get $1	805	30	11	11	37%	-2.71	**<0.01**
2	F_$5	First 20 valid participants get $5	805	56	30	27	48%	0.95	0.34
3	F_$10	First 10 valid participants get $10	805	**64**	32	25	39%	0.50	0.62
4	D_$10	In drawing for 1 out of 10 $10 cards	805	29	14	14	48%	-2.03	*0.04*
5	D_$25	In drawing for 1 out of 4 $25 cards	806	38	25	23	61%	0.03	0.98
6	D_$50	In drawing for 1 out of 2 $50 cards	806	36	25	24	67%	0.26	0.80
7	D_$100	In drawing for 1 $100 card	806	53	**38**	**36**	**68%**	3.00	**<0.01**

Table 1: Results for first-come-first-serve ("*F*") and drawing ("*D*") recruitment. Statistics reports the number of survey starts, finishes, valid responses, and valid rate. The largest value in each column is bolded and highlighted green; the smallest is highlighted pink. Residual reports adjusted standardized z-scores: positives indicate higher-than-expected valid responses. p-value is the significance of the z-score, highlighted at $p < 0.05$ and bolded if significant after correction for multiple comparisons.

offered a drawing for a single $100 **incentive** (D_$100). By contrast, we note that for F_$1, we had fewer responses than awards available! Our results align with findings in other fields (e.g., lotteries) that humans prefer the chance of a large award [8]. Using our most productive incentive structure results in over three times the number of valid responses than using our least productive incentive.

RQ2—Drawings vs. Response Order: Data Quality: We compare drawing incentives vs. first-come-first-serve incentives using a two-tailed t-test for population proportions. We do not find evidence that drawing vs. first-come-first-serve produces more valid data points overall ($p = 0.37$). However, we do find significant differences in data quality between the two groups. While significantly more first-come first-serve participants start the study ($p = 0.02$), **drawing-incentive participants data is of significantly higher quality**. Of those who start the study, drawing incentive participants are more likely to both finish the study ($p = 0.003$) and also to pass data quality thresholds ($p = 0.0004$): see "% Valid" in Table 1, comparing 68% valid (best drawing) to 48% valid (best FCFS).

RQ3–Response Rate vs. Maximum Award: We use Pearson's r to correlate the incentive's maximum award amount with the number of valid responses collected. These two values have a strong positive correlation with $r = 0.74$: **the bigger the potential award, the more valid responses** (even though bigger awards are less likely to be obtained by any one participant). This result has significant implications on how software engineering researchers design incentive structures on a fixed budget.

4 GUIDELINES AND DISCUSSION

First, our results support using a *random drawing* (raffle) for the potential to win a larger sum, rather than paying the first n valid participants a fixed rate. Second, our results support using a single larger drawing sum: this increases data quality (the number and percentage of valid responses, as assessed via a manual quality threshold). Although $10 \times \$10$ and $1 \times \$100$ are mathematically equal, following results in social science [8], we find they are not equal when recruiting software engineering participants. The single larger advertised value resulted in more valid responses (36 vs. 27) and more valid data among those responding (68% vs. 48%).

Limitations: We acknowledge the low overall response rate (more indicative of email/remote surveys than in-person studies), but, for reasons of space, focus on one primary threat to generality: our population of students. Industrial or professional developers may not be motivated by monetary amounts in the ranges considered (compared to their salaries or wages); instead, recent reports suggest that such participants may be motivated more by social good aspects of the proposed work and how societal benefits are communicated (cf. [4]). Our results may not generalize beyond students.

5 CONCLUSION

Human studies are critical to much of software engineering, and recruiting can be a critical barrier to human studies [1]. Informally, the primary contribution of this work is that we conducted the same experiment seven times, advertising to 800 different students each time, changing only the incentive structure: a controlled investigation of the effect of structuring a fixed budget. We find that random drawings (i.e., raffles, lotteries) produce more data ($p = 0.0001$), result in participants more likely to finish the study ($p = 0.003$), and result in more valid data ($p = 0.004$). While 10×10 and 100×1 are equal numbers, the bigger the potential reward, the more valid responses are generated ($r = 0.74$). Software engineers should use *drawings for one large prize* to recruit students for human studies.

REFERENCES

[1] Raymond PL Buse, Caitlin Sadowski, and Westley Weimer. 2011. Benefits and barriers of user evaluation in software engineering research. In *Object oriented programming systems languages and applications*. 643–656.

[2] Shi-Yi Chen, Zhe Feng, and Xiaolian Yi. 2017. A general introduction to adjustment for multiple comparisons. *Journal of thoracic disease* 9, 6 (2017), 1725.

[3] Madeline Endres, Westley Weimer, and Amir Kamil. 2021. An Analysis of Iterative and Recursive Problem Performance. In *Computer Science Education*. 321–327.

[4] Yu Huang, Denae Ford, and Thomas Zimmermann. 2021. Leaving My Fingerprints: Motivations and Challenges of Contributing to OSS for Social Good. In *International Conference on Software Engineering*. 1020–1032.

[5] Andrew J Ko, Thomas D LaToza, and Margaret M Burnett. 2015. A practical guide to controlled experiments of software engineering tools with human participants. *Empirical Software Engineering* 20, 1 (2015), 110–141.

[6] Kypros Kypri and Stephen J Gallagher. 2003. Incentives to increase participation in an Internet survey of alcohol use: a controlled experiment. *Alcohol and Alcoholism* 38, 5 (2003), 437–441.

[7] Norsaremah Salleh, Rashina Hoda, Moon Ting Su, Tanjila Kanij, and John Grundy. 2018. Recruitment, engagement and feedback in empirical software engineering studies in industrial contexts. *Info. and software tech.* 98 (2018), 161–172.

[8] Zur Shapira and Itzhak Venezia. 1992. Size and frequency of prizes as determinants of the demand for lotteries. *Organizational Behavior and Human Decision Processes* 52, 2 (1992), 307–318.

[9] Donald Sharpe. 2015. Chi-square test is statistically significant: Now what? *Practical Assessment, Research, and Evaluation* 20, 1 (2015), 8.

[10] Carl L Tishler and Suzanne Bartholomae. 2002. The recruitment of normal healthy volunteers: a review of the literature on the use of financial incentives. *The Journal of Clinical Pharmacology* 42, 4 (2002), 365–375.