A Generalizable Method for Validating the Utility of Process Analytics with Usability Assessments

Ryan Mullins[(✉)], Chad Weiss, Brent D. Fegley, and Ben Ford

Aptima, Inc., Woburn, MA 01801, USA
`rmullins@aptima.com`

Abstract. Crowdsourcing systems rely on assessments of individual performance over time to assign tasking that improves aggregate performance. We call these combinations of performance assessment and task allocation *process analytics*. As crowdsourcing advances to include greater levels of task complexity, validating *process analytics*, which requires replicable behaviors across crowds, becomes more challenging and urgent. Here, we present a work-in-progress design for validating *process analytics* using integrated usability assessments, which we view as a sufficient proxy for crowdsourced problem-solving. Using the process of developing a crowdsourcing system itself as a use case, we begin by distributing usability assessments to two independent, equally-sized, and otherwise comparable subgroups of a crowd. The first subgroup (control) uses a conventional method of usability assessment; the second (treatment), a distributed method. Differences in subgroup performance determine the degree to which the *process analytics* for the distributed method vary about the conventional method.

Keywords: Crowdsourcing · Usability assessment · Process analytics
Performance assessment

1 Introduction

In recent years, crowdsourcing, specifically the act of leveraging collective intelligence via computer-supported systems, has exploded in popularity. Research has shown the value of crowdsourcing approaches in a variety of domains, from word processing [1] to dataset development [10] to geopolitical event forecasting [9]. The value of these systems comes from their ability to assess individual performance over time and tailor tasking assignments to improve aggregate performance (see [5,11,13,14], among others). We call these combinations of performance assessment and task allocation capabilities *process analytics*; and our goal is to validate the utility of *process analytics* relative to some baseline.

Recent research has tried to apply crowdsourcing approaches to increasingly complex problems, for example argumentation [6] and composable teaming [13].

As such work continues, we expect to encounter sufficiently complex, defeasible, incendiary, and latent problems that require more adaptive and abstracted *process analytics*, which is to say problems where the process is more important than the individual in the workflow. Some examples include organizational change management, strategic corporate decision-making, and cultural change management (see [2,7,8], among others). We refer to these collectively as organizational problem-solving challenges.

We hypothesize that *process analytics* may be replicable validated via a proxy process that (a) presents participants with a repeatable task of meaningful complexity and (b) is not dependent on the behavioral characteristics of the crowds used for assessing system performance. Below, we present a work-in-progress design for validating the utility of *process analytics* designed to enable crowdsourced organizational problem-solving using usability assessments as the proxy process.

2   Validation of Utility Through Usability Assessment

Our primary obstacles in validating *process analytics* are (a) the limited ability to replicate organizational state and participant behaviors to support rigorous performance comparison (see [3], among others), and (b) the latency between the decision to implement a solution and the manifestation of its repercussions (see [4], among others). The following subsections discuss the suitability of usability assessment as a proxy process, the method by which usability assessment will be implemented, and initial performance measures used for comparison.

2.1   Assessing the Suitability of Usability

Fidelity and timeliness are our primary suitability measures. Regarding fidelity, sufficient proxy processes must capture the complexities and nuances of debating organizational problems and their solutions. Ideal proxies will also capture the sequenced and dependent nature of solutions to complex organizational problems. Modern, agile product management—the utilization of end-user feedback to drive future development—is an equally complex, nuanced, and interdependent process. Solutions and their prioritization must address and/or align with three critical perspectives: functionality required by end-users, technical feasibility of implementation, and the vision of various stakeholder groups. These perspectives are also proxies for perspectives found in organizational restructuring problems.

Regarding timeliness, the validation method we use must produce results at a much more rapid pace than organizational change. Modern product management and development practices are trending towards week- and month-long iteration cycles, if not faster, which is an order of magnitude increase, at minimum, compared to the latency of organizational problem-solving. Similarly, we can directly assess the impact of a change (i.e., the utility and usability of a feature) across

product iterations, a process that would require orders-of-magnitude differences in level of effort to model and validate in organizational change problems.

Given measures of fidelity and timeliness, usability assessment supporting product management objectives has sufficient character as a proxy for organizational problem-solving, particularly for validating *process analytics*.

2.2   A Method of Implementation

Our method is an extension of common practices in the product development, Agile software development, and user experience engineering communities. We assume that the crowdsourcing tool being evaluated uses some form of issue management tool (e.g., Jira[1] or GitHub[2]) to independently track the status of features, bug fixes, etc. being considered for future releases. Our method requires that knowledge elicitation mechanisms germane to usability assessment and product enhancement have been integrated into the crowdsourcing system. The goal is to have participants generate a ranked list of items (i.e., features and bugs) that should be addressed in the next release. Our generalized process for achieving this goal has three phases, derived from guerilla UX methods [12]:

1. An *elicitation phase*, where pain-points, bugs, and new feature ideas are solicited and refined;
2. An *assessment phase*, where technical cost and end-user value are calculated; and
3. A *debate phase*, where ideas are selected for inclusion in the next release of the system based on the aforementioned assessments.

Assessments are distributed to two independent and comparable subgroups of the crowd. The first group (control) uses the "conventional" method, where product owners and stakeholders engage face-to-face with participants only during the *elicitation phase* of the process. The second group (treatment) uses the "distributed" method, where participants engage in all phases (excepting the technical cost component of the assessment phase, which we assume to require significant expertise). Decisions regarding when and how to engage participants in the treatment group are made using the tool's *process analytics*. Data collected from these interactions are manually or automatically recorded, respectively, in the issue management system for life-cycle tracking and other uses discussed below. The membership of each group can and should be varied between versions of the tool in order to counteract biases.

2.3   Performance Measurement and Comparison

Process analytic validation occurs by comparing the outputs of the control and treatment groups. We expect that, over time, performance of the treatment group will exceed the performance of the control group along the following measures:

---

[1] https://www.atlassian.com/software/jira/features.
[2] https://github.com/features/project-management/.

– Time to complete a task, where "task" may be defined as brainstorming in service of feature idea elicitation, debate about competing ideas, or the use of various voting mechanisms to develop the final list of features, among other examples.
– Volume of ideas generated. While we anticipate that the total volume will reduce over time, we expect that the amount of time required to produce the same volume of ideas will be consistently lower for the treatment group.
– Reduced problem recurrence, measurement of which is enabled through the analysis of the items that have been stored in the issue management system of choice.
– Scoped scale, where ideas (i.e., problems, solutions, feedback) become more atomic and well-defined over time.
– Frequency of interaction (i.e., how often the group uses the tool).
– Degree of participation (i.e., how many tasks are being engaged).

3   Future Work

We have presented a novel method for validating the utility of *process analytics* used in crowdsourcing tools using an adapted version of usability assessments. This method provides a replicable and timely alternative to other analytic validation methods used in crowdsourcing research, while preserving the fidelity of complex problem-solving challenges. We are currently pilot-testing this method with our tool for organizational crowdsourced problem-solving and plan to publish our findings regarding the ecological validity of this method in the future. If successful, we expect to see improvements in task completion time and problem scoping, increased idea generation, and decreased problem recurrence in the treatment group when compared to the control group. This method should generalize to a broad spectrum of complex crowdsourcing tasks.

References

1. Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D., Panovich, K.: Soylent: a word processor with a crowd inside. Commun. ACM **58**(8), 85–94 (2015). https://doi.org/10.1145/2791285
2. Burns, T.E., Stalker, G.M.: The Management of Innovation. Oxford University Press, Oxford (1961)
3. Camerer, C.F., Dreber, A., Forsell, E., Ho, T.H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al.: Evaluating replicability of laboratory experiments in economics. Science **351**(6280), 1433–1436 (2016)

4. Dahl, M.S.: Organizational change and employee stress. Manag. Sci. **57**(2), 240–256 (2011)
5. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the EM algorithm. Appl. Stat. **28**, 20–28 (1979)
6. Drapeau, R., Chilton, L.B., Bragg, J., Weld, D.S.: Microtalk: using argumentation to improve crowdsourcing accuracy. In: Fourth AAAI Conference on Human Computation and Crowdsourcing (2016)
7. Jimmieson, N.L., Peach, M., White, K.M.: Utilizing the theory of planned behavior to inform change management: an investigation of employee intentions to support organizational change. J. Appl. Behav. Sci. **44**(2), 237–262 (2008)
8. Kerber, K., Buono, A.F.: Rethinking organizational change: reframing the challenge of change management. Organ. Dev. J. **23**(3), 23 (2005)
9. Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., Tetlock, P.: Identifying and cultivating superforecasters as a method of improving probabilistic predictions. Perspect. Psychol. Sci. **10**(3), 267–281 (2015). https://doi.org/10.1177/1745691615577794
10. Post, M., Callison-Burch, C., Osborne, M.: Constructing parallel corpora for six Indian languages via crowdsourcing. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, pp. 401–409. Association for Computational Linguistics, Montréal, Canada, June 2012. http://www.aclweb.org/anthology/W12-3152
11. Tetlock, P.E., Gardner, D.: Superforecasting: The Art and Science of Prediction. Random House, New York City (2016)
12. Unger, R., Warfel, T.Z.: Guerilla UX Research Methods: Thrifty, Fast, and Effective User Experience Research Techniques. Morgan Kaufmann, Burlington (2012)
13. Valentine, M.A., Retelny, D., To, A., Rahmati, N., Doshi, T., Bernstein, M.S.: Flash organizations: crowdsourcing complex work by structuring crowds as organizations. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 3523–3537. ACM (2017)
14. Yin, M., Chen, Y.: Predicting crowd work quality under monetary interventions. In: Fourth AAAI Conference on Human Computation and Crowdsourcing. The Association for the Advancement of Artificial Intelligence (2016)