Artificial Swarm Intelligence Technology Enables Better Subjective Rating Judgment in Pilots Compared to Traditional Data Collection Methods

Kendra Befort
Boeing Co.
Mesa, AZ. USA
Kendra.l.befort@boeing.com

Camila Proffitt
Boeing Co.
Huntington Beach, CA. USA
camila.j.proffitt@boeing.com

David Baltaxe
Unanimous A.I.
San Francisco, CA. USA
david@unanimousai.com

David Durbin
Army Research Laboratory
Ft. Rucker, AL. USA
david.b.durbin2.civ@mail.mil

Ratings provided by Pilots on workload scales and usability surveys can be biased by subjective differences in perception, experience, skill, emotional state, motivation, and estimation of risk/cost that may be associated with performing a task. Personality dynamics can further compound polarization of issues during pilot debriefings. What if these unwanted effects could be filtered out of pilot data collection and we could cost-effectively access a higher-order, collective 'pilot brain' made up of a combined pilot intellect, intuition, and experience to provide more accurate insight into workload and usability? Swarm AI technology was used in a high fidelity pilot simulation event and compared against a traditional methodology for collecting workload and usability survey data. Pilot and Subject Matter Expert workload and usability survey ratings were collected during the event and compared to a post-event pilot swarm. The results of the study showed pilots engaging in collective intelligence were found to be more effective at rating workload, and also more aligned with Subject Matter Expert workload ratings. This initial workload testing suggests that Swarm AI technology and techniques have great potential for usability research by activating the collective intelligence of groups, which can exceed that of the individual performing alone. The usability survey sample was limited, therefore further study is recommended to validate the generalizability of this technology to Likert Scale data.

INTRODUCTION

Artificial Swarm Intelligence (ASI) is an emerging area of human intellect study derived from nature's phenomenon of large groups that form real-time closed-loop systems with continuous feedback to converge on solutions together, such as bees swarming or fish schooling (Couzin, 2008; Marchall et al., 2009; Seeley, & Visscher, 2003). Unlike Swarm Intelligence (SI) that focuses on the development of autonomous drones or simulated agents, ASI seeks to amplify human intellect by networking groups of humans in a closed-loop system that can answer questions, make predictions, reach decisions, or take actions with greater accuracy and optimized satisfaction among participants (Rosenberg, 2015).

Over the last few years, a series of research studies have further explored ASI using Swarm AI™ technology, developed by Unanimous AI. Unanimous' Swarm AI technology enables distributed populations of users to convene online in real-time as swarms to explore decision-spaces and arrive at solutions. Among its published successes, Swarm AI technology was used to predict the 2015 Academy Awards with a 73% success rate using a swarm comprised of seven individuals randomly selected from a group of 48 movie fans. In comparison, the average participant in the larger group had a 40% success rate, and a standard poll (frequency of responses) of the individuals produced only a 47% success rate (Rosenberg, 2015).

A series of follow-up studies published in 2016 and 2017 explore the impacts of novice vs expert swarms,

small vs large sample sizes, and the results of individuals vs swarm on polling accuracy using Swarm AI technology. All studies showed positive results regarding Swarm AI technology's ability to enable groups to generate higher accuracy output. A swarm of random sports fans out-performed experts by collectively predicting seven of ten 2016 College Bowl games (70% accuracy), whereas experts working independently only predicted five of the ten games (50% accuracy) (Rosenberg, 2016). A study that compared the predictions of 469 NFL fans against a subset swarm of 29 participants on the outcome of 19 Prop Bets in the 2016 Super Bowl found that despite being 16 times larger, the polled crowd returned significantly less accurate results (47% correct) than the swarm (68% correct). Furthermore, the swarm outperformed 98% of individual predictions (Rosenberg, Baltaxe, & Pescetelli, 2016). A separate study compared the benefits of swarm and surveys for prioritizing political objectives. In that study, 68% of the participants rated the swarm-based result as a more accurate reflection of the group's priorities than the individual vote-based result (Rosenberg & Baltaxe, 2016). A 2017 study focused on the consistency of swarm performance over an extended period of time. In that research, which examined forecasts of outcomes of sporting events over time, found that individuals acting alone averaged 55% accuracy, but increased their accuracy to 72% when using the Swarm AI platform. The cumulative accuracy over 50 games (five weeks) resulted in a 131% amplification above individual predictions (Rosenberg & Pescetelli, 2017).

Literature reviews on workload have validated that pilot ratings provided on workload scales and surveys can be biased by subjective differences in perception, experience, skill, emotional state, motivation, and estimation of risk/cost that may be associated with performing a task (Cain, 2007; Gawron, 2008; Kruger, 2008). Some experts recommend multiple measures of workload (subjective and objective), to accurately characterize the demands of a task, because there is no unanimous agreement on workload or the set of scales that should be used to get reliable ratings (Farmer & Brownson, 2003; Gawron, 2008). In addition, survey methods can be susceptible to bias ranging from coverage error, sampling error, non-response error, specification error, measurement error, adjustment error, and processing error (Visser, Krosnick, Lavrakas, & Kim, 2013), which can be cost prohibitive to control. What if these unwanted effects could be filtered out of pilot data collection and we could cost-effectively access a higher-order, collective 'pilot brain' made up of a combined pilot intellect, intuition, and experience to provide more accurate insight into workload and usability?

METHOD

As a parallel to the ongoing research on ASI prediction accuracy, it was hypothesized that Artificial Swarm Intelligence methodologies and technologies could be generalizable to usability data collection methods involving ordinal scales, such as Bedford Workload Rating Scale (BWRS) and Likert Surveys, to achieve greater subjective rating judgment in pilots.

To validate the assumption, Swarm AI technology was used as a post-test in two full-mission high-fidelity usability tests conducted in the Boeing simulator. In the initial test, Swarm AI technology was used to prioritize usability issues viewed as most critical by the users (six Army pilots) that were found during simulated missions. In the second test, BWRS and 5-pt Likert Scale Survey data for individual vs swarm ratings were collected and analyzed for simulated missions (see Table 1).

Table 1. Traditional vs Swarm Method of Comparison

	Pilots (n=6)	Subject Matter Experts (n=2)	Swarm AI Technology (n=6)
Workload	*Provided Individual ratings for all designated segments during missions*	*Provided Observer ratings for all designated segments during missions*	*Pilots visually re-enacted a subset of selected segments, then swarmed on rating*
Survey	*Took online Likert Scale Survey at the end of each mission*	*N/A*	*Pilots swarmed on a subset of questions from the Likert Scale Survey*

Prior to their simulated missions, the six pilots were trained to use the BWRS. During the simulation, they provided workload ratings to evaluate new helicopter functionality. As an additional comparison, two highly experienced Army pilots (subject matter experts) provided observed workload ratings to evaluate the new helicopter functionality. At the end of each mission, the six pilots participated in an online usability survey where the attributes, of 'locating', 'interpreting', and 'interacting' were used to measure acceptability of the new helicopter functionality.

After the simulated missions, a Boeing SME led a review of nine mission segments where the six pilots were asked to "visualize" flying that segment in the simulator. A brief discussion was held to ensure all pilots had fully engaged in recall of that segment. When the discussion was over, the instruction for silence in the

room was given and the swarm commenced. The pilots were tasked again with assigning workload ratings, this time collectively. In the online Swarm AI environment, the pilots worked in synchrony, and continuously assessed and reassessed their workload ratings with respect to each of the possible outcomes, weighing their personal confidence and preferences. Ultimately, the swarm converged on solutions that reflected the collective will of the group, tuned by each individual's unique level of confidence, which was recorded as the Pilot Swarm Rating. The same method was used for Pilot Swarms tasked with a subset of the usability survey questions. At the close of the event, a pilot debrief was conducted where the pilots provided feedback about their perceived subjective judgment for measuring workload when using the new helicopter functionality, and the data collection methods.

RESULTS

The data collected in this study reflect small unequal sample sizes rating on an ordinal scale, thus to meet analysis assumptions a Kruskal Wallis One-Way Analysis of Variance (ANOVA) was used to analyze the workload ratings provided by pilots ($n = 6$), subject matter experts ($n = 2$), and the Pilot Swarms for nine mission segments. The analysis showed a significant difference in one or more groups ($KW = 7.30$, $p = .026$) with a post hoc paired comparison showing that the Pilot group provided significantly lower workload ratings than the SMEs (KW_{CRT} 1.834, $p = .018$), and the Pilot group also provided significantly lower workload ratings than those same pilots participating in the Swarm (KW_{CRT} 1.834, $p = .021$) with an 80% desired confidence (see Figure 1).

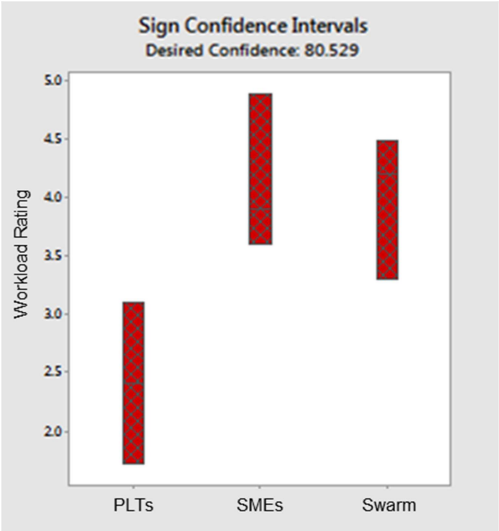

Figure 1. Kruskal Wallis Post Hoc Workload Analysis

Time constraints during the event limited the swarm to considering only five usability survey questions. The smaller sample size favored a non-parametric Mann-Whitney Test. The Likert ratings from the online usability survey were compared with the Pilot Swarms on the same survey questions, and though there was a trend toward more conservative ratings (i.e., less extremes), significance was not found ($W = 26.0$, $p = .832$).

At the conclusion of the simulation event, all pilots (100%) agreed their swarm workload and usability survey ratings were the most realistic ratings.

DISCUSSION

We believe the results have shown that using swarm technology likely resulted in more realistic workload ratings than individual pilot workload ratings. The Pilot Swarm and SME workload ratings were not significantly different, and judged more realistic during the pilot debrief and subsequent discussions among test personnel who observed the missions. It is our assumption that Swarm AI technology empowered the pilot to think as a representative of the pilot community and not as just an individual with an opinion. In addition, we observed that the Swarm AI technology enabled the pilots to 'negotiate' workload and usability survey ratings in a group setting, but still allowed each pilot to maintain his anonymity in regard to his stance on the workload or survey rating during a swarm. Therefore, results support the premise that Swarm AI technology provides a platform where groups working together in closed-loop systems, with real-time feedback control, can more effectively optimize solutions, or in the case of subjective data, provide more realistic workload and usability survey ratings.

Going forward, we plan to further evaluate the Swarm AI technology in simulation studies to assess its utility in broader contexts and to expand upon its online capabilities to explore virtual research objectives.

DEMONSTRATION MATERIALS

If wifi is available, Authors will present an interactive demonstration of the Swarm AI platform and enable attendees to participate in a real-time swarm using tablet computers. To conduct the demonstration, 30-40 tablet computers will be provided to attendees by the presenters. Participants with their own wifi-connected tablets and laptop computers can also participate. Participants will log into the Swarm AI platform anonymously using either standard Chrome or Firefox browsers or a pre-loaded app. No other materials are

required. Prior to the demonstration, we will confirm that wifi is available and that facility network security is compatible. This is not typically an issue.

DEMONSTRATION METHOD

Presenters will guide attendees through familiar investigation and experimental scenarios using Swarm AI applications for data collection and insight exploration. The Presenter will ask questions and the attendees will work together to answer the questions as a "swarm" by simultaneously interacting with the Swarm AI interface on the tablets. The demonstration will cover a variety of question types, media support, and question-asking techniques specific to swarm research methodologies.

The second part of the demonstration will focus on analysis of collected data (e.g., behavioral dynamics, cohort analysis, confidence and decision alignment metrics, etc.) and the unique insights and visualizations available from the Swarm AI platform. It will also highlight unique differences and benefits of using Swarm AI technology compared to traditional data collection and analysis methodologies.

REFERENCES

Cain, B. (2007, July 1). *A Review of the Mental Workload Literature* (Technical Report No. RTO-TR-HFM-121-Part-II). Toronto, Canada: Defense Research Development. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a474193.pdf

Couzin, I.D. (2008). Collective Cognition in Animal Groups. *Trends in Cognitive Sciences, 13*, 36.

Farmer, E., & Brownson, A. (2003). *Review of Workload Measurement, Analysis and Interpretation Methods* (Technical Report No. CARE-Integra-TRS-130-02-WP2). Brussels, Belgium: European Organization for the Safety of Air Navigation (Eurocontrol Integra). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.3382&rep=rep1&type=pdf

Gawron, V.J. (2008). *Human Performance, Workload, and Situational Awareness Measures Handbook* (2nd ed.). London, England: Taylor & Francis.

Kruger, A. (2008). *A Systems Approach to the Assessment of Mental Workload in a Safety-Critical Environment* (Dissertation). University of Pretoria, South Africa.

Marchall, J.A.R., Bogacz, R., Dornhaus A., Planque R., Kovacs T., & Franks, N.R. (2009). On optimal decision making in brains and social insect colonies. *Journal of the Royal Society Interface*, *6*, 1065.

Rosenberg, L.B. (2015, September). Human swarming, a real-time method for parallel distributed intelligence. *Swarm/Human Blended Intelligence Workshop (SHBI)*. doi: 10.1109/SHBI.2015.7321685

Rosenberg, L.B. (2016, July). Artificial Swarm Intelligence vs Human Experts. *International Joint Conference on Neural Networks (IJCNN), IEEE*. doi: 10.1109/IJCNN.2016.7727517

Rosenberg, L.B., & Baltaxe, D. (2016, December). Setting Group Priorities – Swarms vs Votes. *Swarm/Human Blended Intelligence Workshop (SHBI)*. doi: 10.1109/SHBI.2016.7780279

Rosenberg, L.B., Baltaxe D., & Pescetelli, N. (2016, October). Crowds vs Swarms, a Comparison of Intelligence. *Swarm/Human Blended Intelligence Workshop (SHBI)*. doi: 10.1109/SHBI.2016.7780278

Rosenberg, L.B., & Pescetelli, N. (2017, September). Amplifying Prediction Accuracy using Swarm A.I. *Intelligent Systems Conference (IntelliSys)*. London, UK. Retrieved from http://unanimous.ai/publications/

Seeley, T.D., & Visscher, P. K. (2003). Choosing a home: How the scouts in a honey bee swarm perceive the completion of their group decision making. *Behavioral Ecology and Sociobiology, 54(5)*, 511-520.

Visser, P.S., Krosnick, J.A., Lavrakas, P.J., & Kim, N. (2013). Survey Research. In H.T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 402-440). New York: Cambridge University Press.