Survey Measures for Evaluation of Cognitive Assistants

Aaron Steinfeld, Pablo-Alejandro Quinones, John Zimmerman,
S. Rachael Bennett, Dan Siewiorek
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA, USA
{steinfeld@, paq@andrew, johnz@cs, srbennet@andrew, dps@cs}.cmu.edu

***Abstract*—** A survey designed to measure subject perception of benefit, ease of use, usefulness, collaboration, disorientation, flow, and assistance was used to evaluate two releases of an integrated machine learning cognitive assistance system. The design and validity of this evaluation survey is discussed in the context of an information overload experiment.

***Keywords*:** *subjective performance, intelligent systems, evaluation*

I. Introduction

As part of the RADAR project, a cognitive assistant equipped with integrated machine learning capability is regularly evaluated in human subject experiments. This effort is driven by the belief that machine learning, especially when implemented in complex integrated systems, needs to be evaluated on realistic tasks with a human in the loop. Furthermore, the evaluation is designed to examine the impact of machine learning under information overload conditions.

Unfortunately, research utilizing human subjects to evaluate machine learning centric digital assistants with demanding tasks of this nature is limited. As such, few comparison cases are available. Worse, survey tools to measure user perception of such systems are even harder to find in the literature. Validated surveys are especially valuable in that cross-domain and cross-application comparisons are often more appropriate than purely objective metrics.

Evaluations of many machine learning systems are largely based on simulation (e.g., [1, 2]), comparison to traditional methods (e.g., [3]), and subject judgments on system performance (e.g., [4]). It is quite possible that this is generally the result of the kind of system that is built – something that is not meant to be an assistant but, rather, is designed to perform a task that has specific rules. An assistance system, when designed and evaluated, should be tested with humans in the loop (e.g., [5]).

There is relatively little literature on evaluation results of cognitive digital assistants and their focus tends to be specific to a narrow range of machine learning (e.g., [6, 7]). This may be because most of assistants of this nature are design exercises, lack resources for comprehensive evaluation, not evaluated with humans in the loop, and/or proprietary and unpublished.

Likewise, explorations of suitable exit surveys (e.g., [8-11]) provided promising survey questions but uncovered few measures validated for cognitive personal assistants. NASA-TLX was considered but deemed too narrow for examination of certain system assistance nuances.

This paper addresses the subsequent efforts by the RADAR testing team to develop and validate a survey for evaluating complex technologies under information overload.

A. System and Conditions

Radar, the project's implemented system, is specifically designed to assist with a suite of office tasks. In most cases, the specific technologies are designed to be domain agnostic (e.g., email categorizing, resource scrounging, etc). However, for the purposes of the evaluation, the base data present in Radar and used for learning is centric to the domain of conference planning. As such, certain components appear to be domain-specific but their underlying technologies are more extensible (e.g., conference-related email categories, room finding, etc).

In order to show the specific influence of learning on overall performance, there were two Radar conditions – one with learning (+L) and one without (-L). In the context of the evaluation test, learning was only "learning in the wild" (LITW). Such machine learning is specific to learning that occurs through the course of daily use. Brute force spoon-feeding and code-driven knowledge representation is not LITW. To count as LITW, learning must occur through regular user interaction and user interfaces present in Radar.

The other experimental condition described here is which version of Radar (1.0 or 1.1) was tested. There were significant improvements in both usability and engineering from Radar 1.0 to 1.1.

II. Method

A. Materials and Storyline

Extensive detail on the protocol, materials, and findings on other metrics, especially those specific to overall task performance, can be found in [12, 13]. As mentioned, this paper is focused on the survey design and results.

The general scenario for the evaluation was that the subject

was filling in for a conference planner, who was indisposed, to resolve a crisis in the current conference plan. This crisis was major enough to require a major shuffling of the conference schedule and room assignments that, in turn, triggered secondary tasks. These included supporting plans (e.g., shifting catering, AV equipment delivery, adjusting room configuration, etc), reporting (e.g., make changes to the website, issue a daily briefing, etc), and customer handling (e.g., "here is the campus map"). Noise stimuli were also present in the form of unrelated email, unusable rooms, unrelated web pages, and other clutter content.

The materials included an email corpus and simulated world content. The need for repeatability over time led to the requirement for a simulated world. This consisted of facts about the world (e.g., characteristics of a particular room) and conference (e.g., characteristics of each event).

The simulated world and the initial conference were designed to provide clear boundaries on the types of tasks subjects would need to complete, yet also permit large-scale information gathering, high resolution on learned fact variation, and the opportunity to induce a substantial crisis workload.

The conference itself was a 4-day, multi-track technical conference complete with social events, an exhibit hall, poster sessions, tutorials, workshops, plenary talks, and a keynote address. The conference was populated with over 130 talks/posters, each with a designated speaker and title. All characters were provided with email addresses and phone numbers. Many were also given fax numbers, website addresses, and organizations.

The physical space was a modification and extension of the local university campus. In addition to modifying the student union, two academic buildings and a hotel were created and populated. These latter three buildings were instantiated to protect against campus entry knowledge in the subject pool. This information was presented to the subject in the form of revised university web pages easily accessible from the subject's home page.

Other static web content included a conference planning manual (complete with documentation of standard task constraints), a PDF of the original schedule, and manuals for the tools used by the subjects.

Subjects were also given access to a working, realistic "university approved" vendor portal where goods and services could be ordered for the conference. These included audio-visual equipment, catering, security, floral arrangements, and general equipment rentals. Email receipts, complete with hyperlinks to modification/cancellation pages and computed prices, were delivered to the subject's mail client in real time. All vendor interactions were via web forms since automatic or Wizard of Oz handling of subject e-mails can lead to problems with stimulus consistency and realism. This had face validity since many real-life counterparts are web-based, including the subject signup website used during recruitment.

The corpus initialization for each experiment included:

- The predecessor's conference plan in the file format of the condition toolset,
- Other world state information – e.g., room reservation schedule, web pages detailing room characteristics, etc.,
- Stored e-mail from the original conference planner, including noise messages and initial vendor orders,
- The vendor portal, loaded with the initial orders, and
- Injected e-mail, including details of the crisis, new tasks, and noise.

B. Survey Metrics

The survey questions, and their respective categories, are shown in Table 1. All ratings were a 7-point scale with anchors at 1, 4, and 7 (Strongly agree, Neutral, Strongly disagree). Categories – e.g., metrics – were not revealed to the subjects.

Questions in the Ease of Use, Usefulness, Disorientation, and Flow categories were drawn from surveys validated in other fields [10, 11]. Questions 10, 11, and 13 in the Collaboration section were adapted from surveys validated in computer supported cooperative work research [8, 9]. Given the dramatic differences from the fields in which these survey questions were validated, there was some concern that adaptation for complex intelligent systems would not result in valid measures.

For the purposes of analysis, responses to each question within each category were flipped to have the same positive/negative direction and averaged as a group. This category level rating is referred to as an index (e.g., Ease of Use index). The exception is the General category – these are not designed to measure a common metric, so they are left independent.

Questions 16 and 17 were specifically designed to examine how the specific mixture of user interaction, machine learning, and automation affected perceived relationships within collaboration. Ideally, a good mixture will lead to a low score for Question 16 and a higher score for Question 17. This would mean the system was perceived as behaving as an assistant, rather than a taskmaster. The fear with machine learning, and in fact all assistance software, is that the needs of the software (e.g., confirmation, corrections, reminders, etc) will lead to user perception that the locus of control is with the software, rather than the user. It is possible to envision cases where a system has good usability and excellent machine learning, but the nature of the interaction leads the user to feel that they are serving the software.

D. Procedure

Each subject was run through approximately 3 hours of testing (1 for subject training and 2 for time on task). The survey was given at the end of the session. Each cohort of subjects for a particular session was run on a single condition (COTS[1], Radar -L, or Radar +L). When possible, cohorts

[1] Conventional Off The Shelf, see [13] for more details.

were balanced over the week and time of day to prevent session start time bias. Follow-up analyses on this issue revealed no apparent bias. The nominal cohort size was 15 but was often lower due to dropouts, no-shows, and other subject losses (e.g., catastrophic software crash). Cohorts were run as needed to achieve approximately 30 subjects per condition.

Motivation was handled through supplemental payments for milestone completion (e.g., the conference plan at the end of the session satisfies the constraints provided). Subjects were given general milestone descriptions but not explicit targets.

All subjects were recruited from local universities and the general public using a local human subject recruitment website. Subjects were required to meet the following criteria:

- Between the ages of 18 and 65,
- Do not require computer modifications,
- Fluent in English, and
- Not affiliated with or working on the RADAR project.

III. Results

There were several test windows during the period reported here. The survey results data in this document correspond to Radar 1.0 and 1.1 tested on the stimulus package referred to as Crisis 1. The survey reliability data is for the Radar 1.1 test only. Details on Radar 1.1 and Crisis 1 can be found elsewhere [12, 13].

The Radar 1.0 subject pool used for results analysis, after exclusions and dropouts, was 31 and 47 (-L, and +L). Radar 1.1 pool size was 34 and 32. As such, these two tests accumulated 158 cumulative hours worth of time on task by subjects with a multi-task machine learning system.

A two-way ANOVA model on Version (1.0, 1.1) and Learning (-L, +L) was run. Differences between the latter on the survey measures were largely not significant. The exception to this was Usefulness which was viewed as better for Radar +L (F-Ratio, 5.05; p-value 0.026). However, almost every survey measure reported that Radar 1.1 was an improvement over Radar 1.0 (Table 2). Only Question 1 (Confident did task well) was marginally significant.

Figure 2 shows the corresponding means for Version and

Table 2. Improvement for new system version

General Survey Questions	**F-Ratio**	**p-value**
1. Confident did task well	3.89	0.051
2. Task difficult to complete	5.31	0.023
3. As good without software	17.3	<0.0001
Survey index	**F-Ratio**	**p-value**
Ease of Use	10.9	0.0012
Usefulness	4.88	0.029
Collaboration	6.03	0.015
Disorientation	4.13	0.044
Flow	4.31	0.040
Relationship Metric	**F-Ratio**	**p-value**
Assistant vs. Taskmaster (Q17 – Q16, higher is better)	10.2	0.0018

Table 1. Survey Questions

General	
1. I am confident I completed the task well. *(r)*	
2. The task was difficult to complete. *(r)*	
3. I could have done as good of a job without the software tools. *(r)*	
Ease of Use	*Cronbach's alpha:* 0.87
4. Learning to use the software was easy. *(r)*	
5. Becoming skillful at using the software was easy. *(r)*	
6. The software was easy to navigate. *(r)*	
Usefulness	0.94
7. Using similar software would improve my performance in my work. *(r)*	
8. Using similar software in my work would increase my productivity. *(r)*	
9. I would find similar software useful in my work. *(r)*	
Collaboration	0.69
10. I disagreed with the way tasks were divided between me and the computer.	
11. Tasks were clearly assigned. I knew what I was supposed to do. *(r)*	
12. The software did exactly what I wanted it to do. *(r)*	
13. I found myself duplicating work done by the software.	
14. I could trust the software. *(r)*	
15. The software kept track of details for me. *(r)*	
16. The software was assisting me. *(r)*	
17. I was assisting the software.	
Disorientation	0.81
18. I felt like I was going around in circles.	
19. It was difficult to find material that I had previously viewed.	
20. Navigating between items was a problem.	
21. I felt disoriented.	
22. After working for a while I had no idea where to go next.	
Flow	0.57
23. I thought about other things.	
24. I was aware of other problems.	
25. Time seemed to pass more quickly. *(r)*	
26. I knew the right things to do. *(r)*	
27. I felt like I received a lot of direct feedback. *(r)*	
28. I felt in control of myself. *(r)*	
All responses on 7-point scales:	
1 Strongly agree 2 3 4 Neutral 5 6 7 Strongly disagree	
(r) = scale reversed for index averages and analysis	

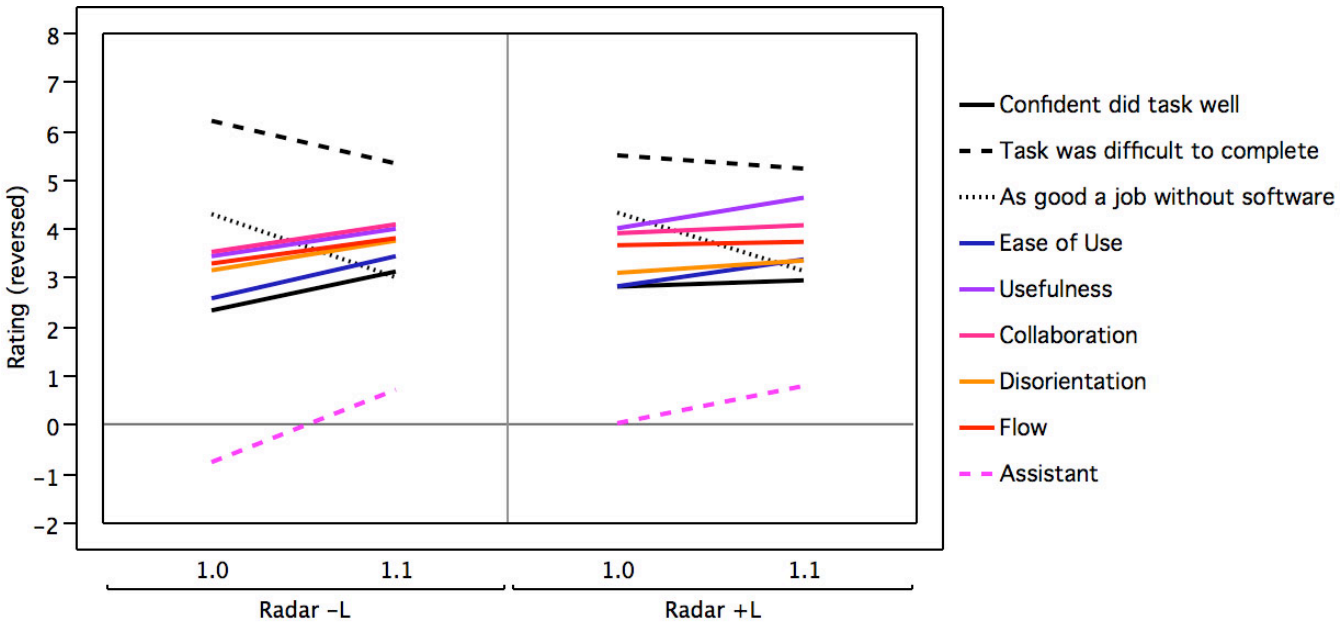

Figure 1. Mean survey responses (Assistant & indices, higher is better; statements, 1=strongly disagree, 7 strongly agree)

Learning. While the interaction comparisons were not significant, it is worth noting is that there is an apparent overall pattern where improvements across versions are less pronounced when machine learning is present in Radar. This matches ground truth in that the majority of the user detectable improvements between versions were in the usability area.

Also, Radar 1.0 -L has a negative Assistant value; subjects felt this instance of Radar was more of a taskmaster than an assistant. The latter finding is not surprising in that the Radar 1.0 user interaction was extremely onerous and only marginal assistance was provided by the software due to the lack of machine learning. This suggests that the machine learning in Radar 1.0 was enough to offset these known deficiencies.

In general, the index collections performed reasonably well when tested for measurement reliability using the Radar 1.1 data (Table 1). Only the Flow index was markedly below the 0.7 reliability acceptance threshold used in the literature. Collaboration was right on the edge.

An initial estimate of the validity of the Assistant vs. Taskmaster relationship metric is to examine how well it correlates to Question 3 (As good without software). Theoretically, ratings on this metric should decrease as Question 3 increases – i.e., software that is considered a taskmaster will not be regarded as valuable by the end user. This was indeed the result for this data set and these measures were correlated (-0.42; p-value <0.0001; Figure 2). As such, early indications are good with respect to metric validity. However, additional research is needed with more precise measures of assistant/taskmaster ground truth.

IV. Discussion

At the time of the Radar 1.1 test there were still unaddressed issues in usability and engineering. The limited perceived differences in the Learning effect beyond Usefulness, contrary to findings from performance metrics [12, 13], may be due to these remaining issues. Possible explanations include: (a) the poor user experience depressed positive machine learning influences and (b) the improvements in machine learning were not perceptible in a between subjects study design.

At the time of this writing, the next round of annual Radar experiments is underway and additional data on issues like the impact of machine learning and index reliability will become available. Early indications are especially promising on the ability of these metrics to capture the precieved value of machine learning. A larger Learning effect is expected since both the user experience and machine learning aspects of Radar have improved substantially. Unfortunately, final data and analyses are not available yet.

There was a clear feeling within the team that the user interfaces for Radar 1.0 and 1.1 were masking the value provided by the machine learning. To some degree, the results presented here confirm this suspicion and reinforce the importance of good user interaction design.

Having said this, the improvement in survey scores from Radar 1.0 to 1.1 mirrors the ground truth improvements made to the system itself. This, combined with the good reliability results, suggests that these survey measures have merit for other experiments on human use of intelligent assistance systems.

Acknowledgments

At Carnegie Mellon, Robert Kraut provided valuable help with previously validated surveys. Kyle Cunningham, Django Wexler, and Matt Lahut assisted with the development of the study materials.

Jordan Hayes (Bitway, Inc) and Othar Hansson integrated Radar and assisted with experiment execution. Julie Fitzgerald

(JSF Consulting), Mike Pool, and Paul Cohen (University of Southern California) served as external evaluators and generated the crisis stimulus. They, with Mark Drummond (SRI International), provided significant input on the protocol.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DARPA or the Department of Interior-National Business Center (DOI-NBC).

REFERENCES

[1] Clymer, J. R. Simulation of a vehicle traffic control network using a fuzzy classifier system. In Proc. of the IEEE Simulation Symposium. 2002.

[2] Clymer, J. R. and Harrsion, V. Simulation of air traffic control at a VFR airport using OpEMCSS. In Proc. IEEE Digital Avionics Systems Conference. 2002.

[3] Zhang, L., Samaras, D., Tomasi, D., Volkow, N., and Goldstein, R. Machine learning for clinical diagnosis from functional magnetic resonance imaging. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2005.

[4] Hu, Y., Li, H., Cao, Y., Meyerzon, D., and Zheng, Q. Automatic extraction of titles from general documents using machine learning. In Proc. of ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL). 2005.

[5] Schrag, R., Pool, M., Chaudhri, V., Kahlert, R., Powers, J., Cohen, P., Fitzgerald, J., and Mishra, S. Experimental evaluation of subject matter expert-oriented knowledge base authoring tools. In Proc. NIST Performance Metrics for Intelligent Systems Workshop. 2002. http://www.iet.com/Projects/RKF/PerMIS02.doc

[6] Shen, J., Li, L., Dietterich, T. G., and Herlocker, J. L. A hybrid learning system for recognizing user tasks from desktop activities and email messages. In Proc. International Conference on Intelligent User Interfaces (IUI). 2006.

[7] Yoo, J., Gervasio, M., and Langley, P. An adaptive stock tracker for personalized trading advice. In Proc. International Conference on Intelligent User Interfaces (IUI). 2003.

[8] Fussell, S. R., Kraut, R. E., Lerch, F. J., Sherlis, W. L., McNally, M., and Cadiz, J. J. Coordination, overload and team performance: effects of team communication strategies. In Proc. of the ACM Conference on Computer Supported Cooperative Work (CSCW). 1998.

[9] Kraut, R. E., Fussell, S. R., Lerch, F. J., and Espinosa, A., Coordination in teams: Evidence from a simulated management game. Journal of Applied Psychology, under review.

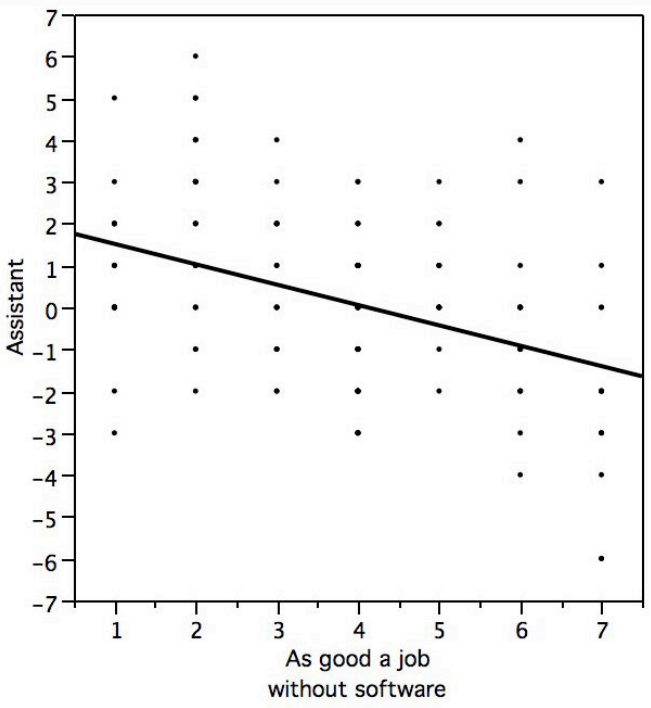

Figure 2. Assistant vs. Taskmaster metric as compared to "I could have done as good of a job without the software tools" (negative slope is better)

[10] van Schaik, P. and Ling, J., Using on-line surveys to measure three key constructs of the quality of human–computer interaction in web sites: psychometric properties and implications. Int. Journal of Human-Computer Studies, 2003. 59: p. 545-567.

[11] van Schaik, P. and Ling, J., Five psychometric scales for online measurement of the quality of human-computer interaction in web sites. Int. Journal of Human–Computer Interaction, 2005. 18(3): p. 309-322.

[12] Steinfeld, A., Bennett, S. R., Cunningham, K., Lahut, M., Quinones, P.-A., Wexler, D., Siewiorek, D., Hayes, J., Cohen, P., Fitzgerald, J., Hansson, O., Pool, M., and Drummond, M. Evaluation of an Integrated Multi-Task Machine Learning System with Humans in the Loop. In Proc. NIST Performance Metrics for Intelligent Systems Workshop (PerMIS). 2007.

[13] Steinfeld, A., Bennett, R., Cunningham, K., Lahut, M., Quinones, P.-A., Wexler, D., Siewiorek, D., Cohen, P., Fitzgerald, J., Hansson, O., Hayes, J., Pool, M., and Drummond, M., The RADAR Test Methodology: Evaluating a Multi-Task Machine Learning System with Humans in the Loop. 2006, Carnegie Mellon University, School of Computer Science: Pittsburgh, PA. http://reports-archive.adm.cs.cmu.edu/anon/2006/abstracts/06-125.html