Enlarging Usability for Ubiquitous Displays

Jacob Somervell, C. M. Chewar, D. Scott McCrickard, Ali Ndiwalana
Center for HCI & Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0106
{jsomerve, cchewar, mccricks, andiwala}@cs.vt.edu

ABSTRACT
As we consider ubiquitous systems that display information on large screen interfaces, we must find reasonable methods for obtaining usability assessments. Standardized, generic methods provide appeal since they allow ready application, benchmarking, and comparison of results. However, critical usability concerns for these interfaces may demand more focused evaluation methods with interface-specific evaluation tools. This work probes at the tradeoffs for usability evaluation of ubiquitous systems—particularly between using specific and generic survey tools to support a claims analysis process. Our study involves formative user interface testing of two ubiquitous large screen display notification systems, each with a generic and specific survey tool. We analyze survey tool performance in supporting immediate and long-term design needs, demonstrating the relative utility of each tool. The evidence we present clarifies the tradeoff between using specific and generic usability evaluation tools—favoring the generic tools—an important finding as tool development efforts proceed for usability evaluation of ubiquitous systems.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces–*benchmarking, evaluation/methodology.*

General Terms
Human Factors, Measurement, Design

Keywords
Claims analysis, notification, large screen, information exhibit

1. INTRODUCTION
As information presentation shifts from the desktop to ubiquitous displays, usability evaluation methods need to be tailored or newly developed to address pivotal user concerns and ensure quality software development. Ubiquitous systems bring new challenges to usability [1], mostly due to the nature of their multi-tasking use, in which attention is shared between ongoing tasks. However, there are many different types of usability evaluation methods, and it is unclear which ones would serve as the best models. One important variation in methods is whether to use an interface-specific tool or a generic tool that applies to a broad class of systems. The goal of our work is to investigate tradeoffs to these two approaches for evaluating large, ubiquitous displays.

ACM-SE Conference '03, March 7-8, 2003, Savannah, GA.

Specific evaluation tools are developed for a single application, and apply solely to the system being tested (we refer to this as a *per-study* basis). Many researchers use this approach, creating evaluation metrics, heuristics, or questionnaires tailored to the system in question [2][8]. These tools seem advantageous because they provide fine grained insight into the target system, yielding detailed redesign solutions. However, filling immediate needs is costly—for each system to be tested a new evaluation method needs to be designed (by designers or evaluators), implemented, and used in the evaluation phase of software development.

In contrast, *generic* evaluation tools are not tailored to a specific system and tend to focus on higher level, critical problem areas that might occur in systems within a common class. A generic method is created once (by usability experts) and used many times in separate evaluations. They are desirable for allowing ready application, promoting comparison between different systems, benchmarking system performance measures, and recognizing long-term, multi-project development progress. However, using a generic tool often means evaluators sacrifice focus on important interface details, since not all of the system aspects may be addressed by a generic tool. The appeal of generic methods is apparent over a long-term period—low cost and high benefit.

This apparent tradeoff for selecting usability evaluation tools for ubiquitous systems must be clarified. To this end, we conducted an experiment to determine the benefits of each approach in supporting a *claims analysis*, a key process within the scenario-based design approach [11]. In a claims analysis, an evaluator makes claims about how important interface features will impact users. *Claims* can be expressed as tradeoffs, conveying upsides or downsides of interface aspects like supported or unsupported activities, use of metaphors, information design choices (use of color, audio, icons, etc.), or interaction design techniques (affordances, feedback, configuration options, etc.). After discussing other usability evaluation method comparisons, the actual systems we evaluated, and our hypotheses and analytical process, we present our results and implications of our findings.

2. MOTIVATION & BACKGROUND
In recent years, determining effective usability evaluation methods (UEMs) for assessing usability of interfaces has been an important topic of research with human-computer interaction. Reports comparing UEMs have sparked an interesting debate on valid comparison methods [5]. Others have contributed improved

comparison techniques, such as Hartson's method that uses metrics like thoroughness, reliability, and downstream utility to compare UEMs based on real and found usability problem sets [6]. The challenge in using this method is producing problem sets in a consistent manner for each UEM and interface under investigation—a challenge that can be overcome with a structured claims analysis approach. The analysis approach we designed for this study demonstrates this and can be extended with Hartson's techniques for additional evaluation tool testing.

Other UEM research efforts have developed high level, generic evaluation procedures, a notable example being Nielsen's heuristics [10]. Heuristic evaluation has been embraced by practitioners because of its discount approach to assessing usability. With this approach (which involves identification of usability problems that fall into nine general and "most common problem areas"), 3-5 expert evaluators can uncover 70% of an interface's usability problems. However, the drawbacks to this approach (and most generic approaches) are evident in the need to develop more specific versions of heuristics for particular classes of systems. For example, Mankoff et al. [8] created a modified set of heuristics for ambient displays. These displays differ from regular interfaces in that they often reside off the desktop, incorporating parts of the physical space in their design and necessitating a more specific approach to evaluation. Similar work dealt with creating modified heuristics for groupware systems [2]. In this work, Baker modified Nielsen's original set to more closely match the user goals and needs associated with groupware systems. Again, the more application class-specific set of heuristics produced better results compared to the general set.

These successes in creating generic evaluation tools that are specific to application class represent new hope for human-computer interaction research—perhaps we can have the long-term comparison and benchmarking advantages with valuable, immediate feedback about interface usability problems. Therefore, as the field pursues UEM adaptation for ubiquitous systems, it is necessary to clarify the tradeoffs between generic and specific tools more systematically.

While we predict that each type of evaluation tool will exhibit different strengths, we hope the magnitude of strengths or weaknesses will suggest the better approach. Therefore, we begin our study with two hypotheses:

1. *Specific evaluation tools produce better interface usability evaluation and redesign conclusions than generic tools.*

2. *Generic evaluation tools provide long-term benefits of guideline and benchmark development and system comparison.*

To compare evaluation tools, we selected two ubiquitous interfaces within the large screen information exhibits application class. *Large screen information exhibits* are software interfaces created for use on large display surfaces, providing interesting or useful *everyday* information to groups or individuals in multi-use areas, such as meeting rooms, break rooms, and labs. These "off the desktop" interfaces provide context-aware, ubiquitous access to deeper information about ongoing activities in a format that allows users to decide when they want to look at the display. Specifications for information exhibits fall easily within design features for ubiquitous *everyday computing*, as discussed in [1].

The *GAWK* (Group Awareness, Work Knowledge) display was designed as part of the Virtual School [3] software suite to show student groupwork progress as icons within a timeline metaphor. As project groups complete work on documents and charts, icons appear in group rows. The systems cycles through newer icons, highlighting each and displaying a summary in the banner. This representation provides a history and current summary of the work done in each group, allowing teachers (and students) to better understand how they should help. Figure 1 shows a screenshot.

The *Photo News Board* shows photos of recent news stories arranged by news type, allowing people who use common areas such as break rooms, labs, and meeting rooms with large screen displays to gain awareness of the day's news events [7]. Highlighted stories (photos) correspond with the text descriptions at the bottom. The system polls and retrieves photos and news clips from Internet sources, introducing newer stories in the center and constantly shifting older stories toward the edge. Highlighting patterns reflect the news category the occupants of the room are most interested in. Figure 2 shows a screenshot.

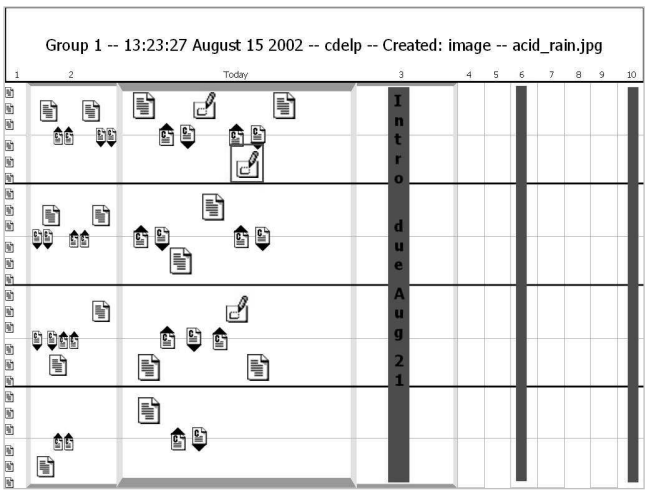

Figure 1. GAWK. Groups are horizontal rows; time proceeds horizontally; deadlines are red; banner at the top cycles details of the work artifacts indicated by the green highlight [4].

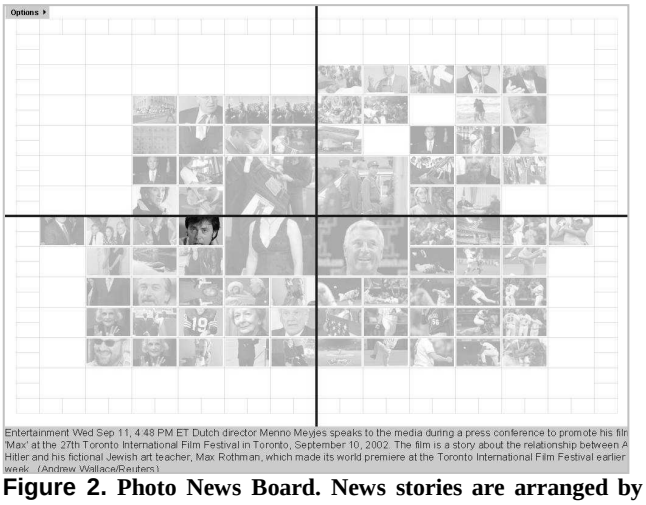

Figure 2. Photo News Board. News stories are arranged by type with newer stories in the center. Highlighted story details appear at the bottom of the screen and reflect the interests of the room occupants.

3. METHODOLOGY

We conducted an analysis of usability evaluation results on both systems to evaluate how well generic or specific surveys could support claims associated with these systems, lead to redesign conclusions, and impact long-term design processes. The overall methodology of this analysis consists of three phases: conducting the usability evaluations, assessing the claims analysis according to each result set from the usability evaluations, and recognizing potential long-term benefits.

3.1 Usability evaluations

We built several assumptions into our analytical approach that we believe to be typical of a usability study in the formative stages of system development. For instance, since participant time is quite costly, our evaluation sessions were designed to be completed within one-half hour. This made a controlled, lab-based test appealing, since we also wanted the feedback to be based on actual experience with the system rather than impressions from screenshots or storyboards. Therefore, we used scripted, rapid prototypes displayed on a 52" screen to illustrate how each system would support a real situation.

To conduct our testing, we used a 2 (system) x 2 (survey type) between-subjects experimental design. Twenty computer science undergraduate students participated in this experiment voluntarily. Participants were tested individually and asked to take on the role of a typical user for the system they were evaluating. To do this, they performed other tasks (such as reading a newspaper or recording quiz grades) that would be part of the usage context (a classroom for the GAWK system and a break room for the Photo News Board). While the participant was engaged in these tasks, the interfaces presented scripted scenarios to familiarize the participants with the information presentation as it would actually be used in the intended situation. After experiencing each of several scenarios, the participant was asked simple, free-response questions about the information displayed by the interface, reinforcing their awareness of system features. However, the only recorded feedback was answers to a nine-question survey provided to the participant once all scenarios were completed.

The between-subjects design allowed both displays to be evaluated using two separate evaluation tools—a specific survey derived for each system that focused on important system features and a generic survey based on the typical users goals for applications within the large screen information exhibit system class. Generic survey questions were based on a framework for understanding user goals of notification systems [9] (a broader class of systems that support information delivery in ubiquitous, multitasking situations). The same generic survey was used for both systems. To maintain consistency and usability study brevity, all three survey versions were developed within our research group and had nine questions. The surveys used Likert-style rating scales for various aspects of the systems. Participants read a statement and indicated their level of agreement with the statement, ranging from strongly disagree to strongly agree.

After aggregating responses for each survey, questions with ratings that clearly showed agreement or disagreement (average responses within one-standard error of the "neutral" response) were then applied to the claims analysis to determine the impact of participant responses on our claims.

Claims

Category	GAWK	PhotoNewsBoard
Supported Activities	(+) showing deadlines helps teachers focus students on tasks ***G9***, ***B4***	(+) seeing photos triggers curiosity about the story ***G9***, *A9*
Font/Icon Usage	(–) size constrains message length to ~80 characters, lack of detail causes distraction *G3, B8*	(–) smallest pictures on outer edges may not be recognizable *G6*, ***A3***

Survey Questions

G9: Appropriate reactions were obvious and intuitive.
A3: I could easily tell which news stories were recent and which were older.
B4: If I were busy with something, changes in the display would NOT distract me.

Figure 3. Example claims and survey questions[1], with upside (+) and downside (-) tradeoffs that correspond to sample questions from the generic (G9) and specific (A3, B4) surveys.

3.2 Claims analysis assessment

To determine the impact of survey responses to understanding usability problems, we had to perform a claims analysis [11] on each interface. Within the scenarios of use developed for each system, claims were made about the various design choices. These claims indicate how the design choices were thought to positively or negatively impact users. Claims analyses produced 58 design tradeoffs for GAWK and 56 for Photo News Board—each addressing system-specific claims based on activity design (e.g. supported or unsupported activities), information design (e.g. font/icon usage), and interaction design considerations. Examples of two categories of claims for each system are shown in Figure 3. Numbers of upside and downside tradeoffs by category can be seen in Table 1's left-most column for each system.

Next, survey questions from both the generic and specific surveys were mapped to each system's claims, although some claims were not addressed by questions on a given survey. This mapping was then used to determine whether or not claims were supported or refuted according to participant opinion. After capturing these numbers for the two types of evaluation tools we compared how thoroughly the surveys addressed the claims analysis, gauging the impact of generic or specific survey tools on targeting immediate, per-study usability concerns and suggesting redesign conclusions. This approach allowed conclusions about hypothesis 1.

3.3 Recognizing long-term benefits

To assess hypothesis 2, we compared generic survey responses for both systems. We started by identifying questions that exhibit low response variance, since these could be candidate questions for benchmark establishment. Then, we looked for cases where the two systems demonstrated similar results (average response value and amount of response variance) on questions that map to similar design tradeoffs, allowing recognition of potential general guidelines that would be useful in designing new systems. We also looked for questions that had wide response variation, since the associated claims might allow detection of design artifacts that are responsible for the usability concern. Finally, we thought about how the two systems compared to each other. This allowed appraisal of the generic survey's impact on long-term design processes—by suggesting guidelines, benchmarking response values, and allowing overall system comparison.

[1] All questions at: http://research.cs.vt.edu/ns/questions.html

4. RESULTS

For the first phase of the study, the average user responses from the four usability surveys can be seen in Figure 4. Participants ranked both displays highly with specific and generic surveys. While individual responses included negative ratings, response averages reflected no "disagree" ratings. Considering both the generic and specific survey response averages, seven out of nine questions (generic = 1, 4, 5, 6, 7, 8, 9; specific = 1, 2, 3, 4, 7, 8, 9) were rated above neutral within one-standard error (indicating at least "Somewhat agree") for the GAWK display on each survey, with five of nine questions showing agreement on each Photo News Board survey (generic = 2, 4, 5, 7, 8; specific = 1, 6, 7, 8, 9). Comparing the amount of variance within all questions by survey (apparent by the average length of the error bars), we see that the GAWK survey responses were quite a bit more consistent than Photo News Board's (a variance difference of .6 on the specific survey and 1 on the generic).

4.1 Hypothesis 1—Per-study impact

In the second phase of the study, during which we matched the claims to the questions, we found that for the GAWK system, the specific survey addressed 56 of 58 (97%) claims and the generic survey addressed 52 of 58 (90%) claims. For the Photo News Board, the specific survey addressed 43 of 56 (76%) claims and the generic survey addressed 37 of 56 (66%). Table 1 provides this data by claim category in the "of" column for each survey and system. Using the question-to-claims mapping (example shown in Figure 3), we recorded the number of claims supported (Table 1 "S" columns) or refuted (Table 1 "R" columns) by the questions receiving consistently conclusive user responses (those listed above). That is, if a question's average response indicated agreement, all corresponding upside (+) claims were then counted as supported. Corresponding downside (–) claims were further analyzed to determine if agreeing with the question indeed meant refuting a negative claim, or if agreeing simply meant that the question no longer applied to the downside claim. For GAWK, the specific survey supported or refuted 37 of 58 (64%) claims, while the generic survey supported or refuted 35 of 58 (60%) claims. For Photo News Board, the specific survey supported or refuted 15 of 56 (27%) claims, while the generic survey supported or refuted 17 of 56 (30%) claims.

In examining what each survey suggested for each interface's redesign efforts, we find valuable courses of action inspired by all four survey response results. For instance, GAWK's specific survey reveals usability concerns related to users understanding how each group's progress evolved over time (#6). This could be an issue with an associated claim about wasted space for future days, prompting a redesign approach such as a fisheye view of the timeline. The generic survey brought out the difficulty in parsing the single-line banner and the inconvenience of not knowing what point in the highlighting/banner association cycle the display is situated (associated with #3)—both of which can be remedied with banner redesigns. For Photo News Board, the specific survey pointed out that users are unable to always recognize a new story by the movement of stories (#4), which can be addressed by making the new picture subtly pulse for a few seconds after entering the collage. The generic survey indicated a problem in using the interface during natural breaks in ongoing tasks (#1), most likely caused by the random highlighting pattern of stories, fixable with a top-bottom or left-right approach.

4.2 Hypothesis 2—Long-term impact

In the final phase of the study, we analyzed the responses to the generic survey (since it alone was used on both systems) to see if guidelines and benchmarks were starting to emerge and to compare the two systems with each other. A visual inspection of the generic responses in Figure 4 shows that three questions had similar ratings on both systems: #4—"the interface provides an overall sense of the information," #7—"the interface support easy understanding of links between different types of information," and #8—"the interface supports rapid reaction to the information." These questions are possible candidates for identifying information exhibit benchmark performance, allowing other systems to be gauged according to how well they score on each question. Other questions had averages that were too far apart or variances that were too high.

We were also able to identify a few questions that suggest design guidelines for information exhibits, both with responses that agreed between systems and responses that differed. For instance, the consistent agreements with question #4 indicate that both interfaces provided an overall sense of important information within each usage scenario. After reconsidering the specific

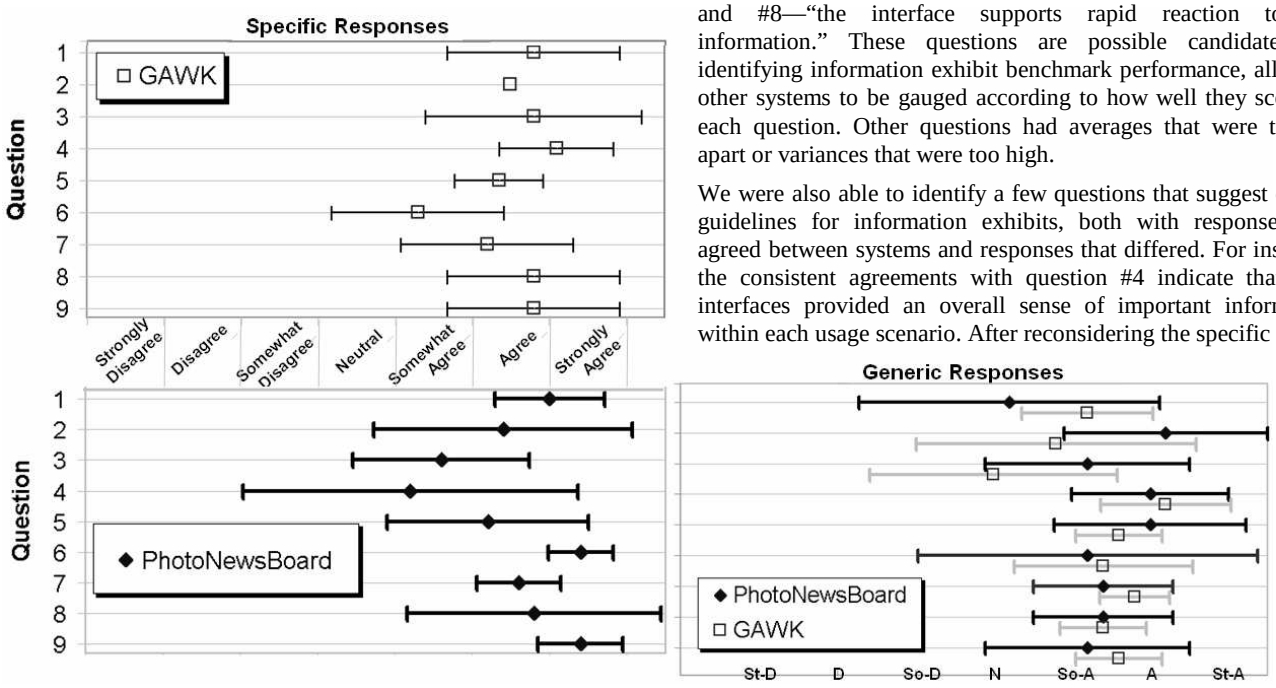

Figure 4. Participant response averages (with one-standard error) for specific and generic surveys. Questions on specific surveys varied according to the interface, while the same generic survey was used for both interfaces.

	GAWK Claims		Supported or Refuted						PNB Claims		Supported or Refuted					
			Specific			Generic					Specific			Generic		
	+	-	of	S	R	of	S	R	+	-	of	S	R	of	S	R
Supported activities	5	1	6	5	0	5	5	0	6	2	8	3	0	7	3	0
Metaphors	3	3	6	1	0	4	2	1	3	2	5	2	0	4	2	1
Layout	5	4	8	2	3	8	3	2	6	3	8	2	1	7	1	1
Colors	6	2	7	5	0	8	3	2	4	0	3	3	0	0	0	0
Fonts/Icons	3	2	5	3	2	5	2	0	2	2	4	0	0	4	1	1
Audio	1	1	2	1	0	2	0	1	1	1	1	0	0	2	1	0
Animation	4	3	7	3	2	7	3	2	3	3	6	2	0	6	2	1
Affordance	1	2	3	1	1	2	1	0	1	1	1	0	0	0	0	0
Transition of states	4	2	6	2	2	6	3	1	5	3	5	0	0	5	1	0
Feedback	2	0	2	1	0	2	2	0	2	2	2	0	2	2	0	2
Config.	2	2	4	2	1	3	1	1	2	2	0	0	0	0	0	0
Subtotals	36	22	56	26	11	52	25	10	35	21	43	12	3	37	11	6
Totals	*58*		*37 (64%)*			*35 (60%)*			*56*		*15 (27%)*			*17 (30%)*		

Table 1. Survey result impact on claims analysis: numbers of claims are shown for claim analysis categories. Specific surveys addressed slightly more claims (*a*), but the generic survey supported/refuted similar percentages of claims (*b*).

claims that are associated with information design, we recognize design features that contribute to this success, such as the use of time scales and the reductionary photo organization. We can encapsulate this as a guideline: *presentation of many information items through a strong organizational metaphor or theme can result in an overall sense of the information's meaning*. Likewise, for #7, design strategies such as highlighting techniques for both interfaces helped users realize links between different types of information (icons or photos with banner information and previous versions of work artifacts with a new submission). This suggests that: *information exhibits can help users understand the relationship between different parts of the display with coordinated, cyclical highlighting of icons that are summarized in a banner rather than tooltips*. Similarly, agreement on question #8 implies that the animation used to introduce new items was effective in supporting reaction. This can be summarized in another potential guideline: *subtle, distinctive animation patterns allow users to rapidly detect and react to newly presented information.*

Guidelines can also be inspired by large differences in responses to a given question, since these questions prompt examination of reasons why a given system's score contrasted with another (or a benchmark). The first three questions on the generic survey all dealt with attention interruption and produced the greatest differences between systems (as seen in Figure 4). Question #1 specifically asked about support for self-defined interruption: "I could find natural break points in my task to look at the display so I wouldn't miss important information." Although the GAWK system scored much higher on this question, it scored much lower on questions #2 and #3 ("the interface did not distract my attention from my current task," and "I was able to notice when new information appeared on the display without stopping my current work"). These results suggest strengths of the constant, rhythmic motion of the photos within Photo News Board (preferable for low interruption and glanceable recognition of interface changes) and the timeline metaphor of GAWK (use of position to organize information that can be spotted during natural breaks)—potential for other guidelines that may be supported in future studies of other information exhibits.

The final consideration for the third phase of the study was the overall system comparison. While no responses on the generic survey were statistically different between systems, comparing the average responses suggests that GAWK supported typical information exhibit user goals better, although Photo News Board may be less interruptive.

5. DISCUSSION

This experiment investigated the tradeoffs associated with using specific and generic evaluation tools for ubiquitous systems—in terms of immediate, per-study contributions to the usability engineering process and impact to long-term design processes.

Hypothesis 1. Surprisingly, the difference between specific and generic claims coverage, regardless of system, was roughly the same, with the specific survey supporting or refuting two more claims than the generic survey for GAWK, while the opposite case was true for Photo News Board. This shows that, although the specific survey applied to more claims than the generic survey (56 to 52 and 43 to 37, for GAWK and Photo News Board, respectively), the generic survey was comparable to the specific surveys in terms of supporting or refuting specific claims—revealing unexpected usability concerns. The comparison of redesign conclusions made available through each survey did not show any advantage for either generic or specific evaluation tools, largely because the strong mapping between questions and claims provides a rich basis for analyzing design artifact usability performance. These findings provide no clear support for hypothesis 1, suggesting no difference between the two tools for per-study usability evaluations. This means that the apparent advantages of the specific method—addressing finer details of a design, as a result of tighter coupling with a claims analysis, to reveal better redesign options—did not manifest in this study.

Hypothesis 2. The results related to the second hypothesis exhibited potential for the generic survey in impacting long-term design processes of benchmarking, guideline creation, and system comparison. Even though we only had a small number of questions and responses for two systems, we were still able to detect commonalities and disparities between the two systems. Because the survey questions were associated with claims (hence, design features), guidelines were easy to create. However, they must be verified by inspection of other systems and analysis of additional user testing results before being widely generalized. Results with the generic survey also allowed identification of three candidate questions as potentially useful benchmark values for information exhibits. However, finalizing these benchmarks will require many more studies, due to the fact that this initial evaluation was based on only two systems. Although we did not see potential for benchmarks in most of the questions, the response differences could be due to specific system design characteristics rather than an indication of a question's poor potential as a benchmark. Further testing other systems could indeed show that other questions on our generic survey may be valid benchmarks. Certainly, evaluating systems within a common application class using a common tool usually allows comparative conclusions to be drawn, and this study was no exception. Based on these observations, we find hypothesis 2 to be supported.

Other observations. Although we initially expected a more vexing tradeoff between the two approaches, our study suggests that generic surveys lose no advantage for per-study usability

evaluations, yet hold valuable potential for long-term design efforts. While the specific method addressed more of the claims for each system, the generic survey performed comparably well at supporting or refuting claims. Given the added bonus of benchmarking, guideline creation, and system comparison, the generic method seems to provide more advantages.

We can also note that the claim analysis process showed to be an extremely useful approach for supporting depth and breadth in usability study problem identification, despite the relatively small amount of data, few users, rapid prototype systems, and brief session durations. This approach to usability evaluation provides direct feedback on design artifacts. By associating user responses to specific claims through the question-to-claims mappings, we were able to determine directed redesign conclusions from both surveys. It is this mapping that provides the redesign capability and insight into the usability of an interface, broadening the analytical scope afforded by each question. Using the claims analysis approach and assessing the coverage a UEM provides to a set of claims seems complement newer UEM comparison methods (e.g. [6]).

From this study, we see that a generic approach to ubiquitous usability evaluation seems like a more logical choice. Hence, the long term benefits of these methods suggest taking the initial cost to produce them, so that they may be reused in subsequent evaluations of new versions or other systems within the application class. As refinement of usability evaluation material for ubiquitous systems proceeds, there is an impetus for carefully considering generic tools that can be created by experts and leveraged by development teams for low-cost reuse and design knowledge collection.

6. CONCLUSIONS & FUTURE WORK

The findings of our study, which compared tailored, application-specific usability surveys to generic surveys addressing critical problems of an application class, can be summarized as follows:

- There is insufficient evidence in our four usability evaluations that specific evaluation tools have an advantage over generic tools in facilitating better identification of usability concerns or redesign strategies.
- We observed the potential long-term benefits of guideline and benchmark development, as well as system comparison inherent in generic evaluation tools.
- Claims analysis proved to be an extremely useful approach for producing problem sets in a consistent manner, which is necessary for validly evaluating UEMs.
- Generic evaluation tools for ubiquitous interfaces should be researched and developed by experts to provide development teams the benefits of low-cost reuse and design knowledge collection.

We recognize many directions for future work, improving upon the actual evaluation tools, extending our UEM comparison process with complementary, metric-centered techniques, investigating other evaluation methods, and drawing out the long-term benefits that are embedded in generic approaches. Certainly, our evaluation tools can be improved upon. Surprisingly, no questions on either surveys showed an overall negative response; we suspect that inversely worded questions could evoke more thoughtful participant response (i.e. agreement indicates a usability concern). Our initial work, especially toward investigating hypothesis 2, can be extended with Hartson's equations [6], comparatively assessing UEM thoroughness, reliability, and downstream utility. In addition, this analytical process can be applied to other generic evaluation methods, such as heuristics, cognitive walkthroughs, and critical incident reports. As other systems are evaluated with generic tools, it will be especially important to collect results in a cohesive manner that empowers formulation of benchmarks, guidelines, and other reusable design knowledge.

REFERENCES

[1] Gregory D. Abowd and Elizabeth D. Mynatt. Charting past, present, and future research in ubiquitous computing, *ACM Trans. on Computer-Human Interaction 7*(1), March 2000.

[2] Kevin Baker, Saul Greenberg, and Carl Gutwin. Empirical development of a heuristic evaluation methodology for shared workspace groupware. In *ACM Conf. on Computer Supported Cooperative Work (CSCW'02).* New Orleans, LA, November 2002.

[3] John M. Carroll, George Chin, Mary Beth Rosson, and Dennis C. Neale. The development of cooperation: Five years of participatory design in the virtual school. In *Proc. on Designing Interactive Systems (DIS 2000)*, 239–251.

[4] John M. Carroll, Dennis Neale, Philip Isenhour, Mary B. Rosson, and D. Scott McCrickard. Notification and awareness: Synchronizing task-oriented collaborative activity. *International Journal of Human-Computer Studies, Special Issue on Design and Evaluation of Notification User Interfaces*, 2003 (to appear).

[5] Wayne Gray and Marilyn Salzman. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-ComputerInteraction 13*(4):203–261,1998.

[6] H. Rex Hartson, Terence S. Andre, and Robert C. Williges. Criteria for evaluating usability evaluation methods. *Intl. Jnl of Human-Computer Interaction 13*(4):373-410, 2001.

[7] William Luebke, Michael Richmond, and D. Scott McCrickard. Collaborative environments supported by large screen displays. In *ACM Conf. on Computer Supported Cooperative Work (CSCW 2002)*, New Orleans, LA.

[8] Jennifer Mankoff, Anind Dey, Gary Hsieh, Julie Kientz, Scott Lederer, and Morgan Ames. Heuristic evaluation of ambient displays. In *Proc. of ACM Conf. on Human Factors and Computing Systems (CHI 2003)*, Ft. Lauderdale, FL (to appear).

[9] D. Scott McCrickard and C.M.Chewar. Attuning notification design to user goals and attention costs. *Comm. of the ACM 46*(3), March 2003 (to appear).

[10] Jakob Nielsen and R. L. Mack. *Usability Inspection Methods*. John Wiley and Sons, New York, NY, 1994.

[11] Mary Beth Rosson and John M. Carroll. *Usability Engineering: Scenario-Based Development of Human-Computer Interaction*. Morgan Kaufman, NY, 2002