From expert-derived user needs to user-perceived ease of use and usefulness: A two-phase mixed-methods evaluation framework

Mary Regina Boland a, Alexander Rusanov b, Yat So a, Carlos Lopez-Jimenez c, Linda Busacca d, Richard C. Steinman c,f, Suzanne Bakken a,e,f, J. Thomas Bigger c,f, Chunhua Weng a,f,*

a Department of Biomedical Informatics, Columbia University, New York, NY, USA
b Department of Anesthesiology, Columbia University, New York, NY, USA
c Department of Medicine, Division of Cardiology, Columbia University, New York, NY, USA
d The Clinical Trials Office, Columbia University, New York, NY, USA
e School of Nursing, Columbia University, New York, NY, USA
f The Irving Institute for Clinical and Translational Research, Columbia University, New York, NY, USA

Article history:
Received 31 July 2013
Accepted 6 December 2013
Available online 12 December 2013

Keywords:
Evaluation studies
Clinical trials
Workflow
Needs assessment
Medical informatics

Abstract

Underspecified user needs and frequent lack of a gold standard reference are typical barriers to technology evaluation. To address this problem, this paper presents a two-phase evaluation framework involving usability experts (phase 1) and end-users (phase 2). In phase 1, a cross-system functionality alignment between expert-derived user needs and system functions was performed to inform the choice of “the best available” comparison system to enable a cognitive walkthrough in phase 1 and a comparative effectiveness evaluation in phase 2. During phase 2, five quantitative and qualitative evaluation methods are mixed to assess usability: time-motion analysis, software log, questionnaires – System Usability Scale and the Unified Theory of Acceptance of Use of Technology, think-aloud protocols, and unstructured interviews. Each method contributes data for a unique measure (e.g., time motion analysis contributes task-completion-time; software log contributes action transition frequency). The measures are triangulated to yield complementary insights regarding user-perceived ease-of-use, functionality integration, anxiety during use, and workflow impact. To illustrate its use, we applied this framework in a formative evaluation of a software called Integrated Model for Patient Care and Clinical Trials (IMPACT). We conclude that this mixed-methods evaluation framework enables an integrated assessment of user needs satisfaction and user-perceived usefulness and usability of a novel design. This evaluation framework effectively bridges the gap between co-evolving user needs and technology designs during iterative prototyping and is particularly useful when it is difficult for users to articulate their needs for technology support due to the lack of a baseline.

1. Introduction

Evaluation is important to all innovations [1], including Health Information Technology (HIT) interventions. However, there are significant barriers for evaluating HIT, such as the lack of a reference HIT gold standard [2], the paucity of knowledge regarding user needs prior to the development of an HIT intervention [3], and the complexity of socio-technical systems and multi-stakeholder teams, which can affect the intended outcomes of the HIT intervention [4].

Five levels of usability evaluation have been described in the literature: task-based, user-task, system-task, user-task-system, and user-task-system-environment [5]. The first three levels occur early in prototype development, focusing on task identification, how users perform their tasks, and if a system supports the task it was designed for [5]. The fourth level addresses how users perform a set of tasks using the system and how users perceive the usefulness of the system [5]. Building on these, the fifth level evaluates how the task, user, and system interact within the workplace environment [5]. The fifth level usually occurs after system deployment [6], while the fourth level occurs during the prototype deployment stage.

Abbreviations: IMPACT, Integrated Model for Patient Care and Clinical Trials; CRC, Clinical Research Coordinator; CTMS, Clinical Trial Management System; SUS, System Usability Scale; UTAUT, Unified Theory of Acceptance of Use of Technology; PRN, Pro Re Nata (as needed); HIT, Health Information Technologies; STARE-HI, STAtement on Reporting of Evaluation studies in Health Informatics.

* Corresponding author. Address: Department of Biomedical Informatics, Columbia University, 622 W 168th Street, VC-5, New York, NY 10032, USA. Fax: +1 212 305 3302.
E-mail address: cw2384@columbia.edu (C. Weng).
Often, HIT prototypes are not fully comparable with existing systems because of their inherent novelty and uniqueness. Identifying an appropriate baseline or comparison system (when a baseline is lacking) for evaluation purposes is difficult for emerging HIT interventions [7]. However, it is important to overcome this problem and select the “best available” system as a reference standard for comparing the usability and effectiveness of various systems against.

Various evaluation methods and strategies have been developed [8]. Evaluations that mix methodologies are considered robust [9,10] and particularly useful in the medical setting [11,12]. There are many ways to combine methods, such as mixing qualitative and quantitative methods [13], involving users of varying perspectives for data collections [14], or using various data collection methods to achieve greater data validity. The mixed-methods approach is superior to either qualitative or quantitative research methods alone [15] because it ensures comprehensive data collection and avoids unnecessary a priori assumptions often made by researchers [13]. In a mixed-methods evaluation, qualitative data can be used to identify unmet needs [16–18], while quantitative data can measure workflow impact [18–22]. Data triangulation further allows verification of derived user needs [23].

In addition, evaluation designs can involve different types of evaluators, including usability experts and intended users. Several evaluation methods utilize usability experts. Cognitive task analysis (CTA) is an evaluation method performed by usability experts for assessing usability and has been successfully applied in healthcare settings [24]. Cognitive walk-through also involves usability experts but is less intensive than CTA. In a cognitive walk-through, an expert who is already familiar with the system performs a set of predefined tasks and notes the number of steps required by certain tasks and any usability and design problems with the interface [25]. Other evaluation methods make use of the intended end-users themselves. Time-motion analysis is a quantitative method that measures the amount of time users spend performing a task [26]. Results of time-motion analyses provide insight into the likeliness of system adoption and can be used to identify areas in users’ workflow amenable to an informatics intervention [27]. The advantages of surveys, emails, and think-aloud protocols in evaluating informatics interventions is well-established [28]. Software log analysis is another useful evaluation method that can capture behind-the-scenes interactions with the system and is not intrusive to evaluators [29]. Questionnaires can assess users’ perceptions of a system’s usability [30] and the likelihood of acceptance of the technology [31]. Qualitative information in the form of unstructured interviews and think-aloud protocols are especially useful during system evaluation because they allow users to provide additional information not specified a priori via a structured questionnaire [32]. Mixing qualitative and quantitative evaluation methods to further enhance the evaluation result is a well-established approach [15].

To address the evaluation challenges with emerging HIT, where user needs are vague and clinical workflow is complex, we describe a two-phase mixed-methods evaluation framework to bridge the gap between co-evolving user needs and technology designs during iterative prototyping. This novel evaluation framework enables an integrated assessment of both expert-derived user needs satisfaction and the user-perceived usefulness and ease of use of emerging HIT interventions [33]. It supports formative evaluation of HIT before the release of a fully-fledged system. We applied our methodology to evaluate the prototypes of a novel clinical research decision support system called Integrated Model for Patient Care and Clinical Trials (IMPACT), which is designed to provide decision support for scheduling research visits [34]. We followed the STAtement on Reporting of Evaluation studies in Health Informatics (STARE-HI) guideline for reporting evaluation studies where applicable [35] since our framework was ideally suited for formative evaluations of software prototypes. We then describe this evaluation framework and its use in evaluating IMPACT prototypes.

2. Materials and methods

Our evaluation framework consists of two phases. In phase 1, a usability expert collects user needs and compares the intervention with related systems by aligning system functions with derived user needs for each system. This enables the selection of a suitable comparison system followed by a cognitive walk-through involving a task analysis and a comparison of interface design differences between the innovation and the comparison system. Phase 2 involves the system’s end-users, Clinical Research Coordinators (CRCs) to collect quantitative and qualitative data. Fig. 1 illustrates our mixed-methods evaluation framework.

Table 1 shows the types of data collected at each phase. Two measures are assessed during phase 1: the number of steps required by each task and interface features used while performing each task (e.g., screen transitions and pop-ups). Analysis during phase 1 allows developers to assess how well the system performs in a laboratory setting. If phase 1 identifies many critical system functions that require improvement, the system can be refined prior to testing with end-users. This approach prevents end-users from being adversely affected by a system requiring critical improvements. Since phase 1 of the IMPACT evaluation revealed no such deficiencies, we were able to proceed directly to phase 2 of the evaluation.

2.1. The IMPACT system and its environment

Columbia University Medical Center (CUMC) is an academic medical center where many patients are also research participants. The IMPACT system, developed at CUMC, was designed to integrate information from both patient care and clinical research to facilitate the scheduling of research visits and coordination of patient care and research workflows. It incorporates temporal constraints from the research protocol’s visit schedule and availability of research resources (e.g., rooms, equipment, and personnel) into a calendar interface. Designed for use by CRCs and schedulers, IMPACT automatically calculates resource availability and recommends suitable dates and times for the next research visit. IMPACT’s complete functionality has been published elsewhere.

2.2. Phase 1: Usability expert component

2.2.1. Cross-system feature vs. derived user needs alignment

We recruited a usability expert to derive comprehensive user needs for scheduling decision support. This usability expert was independent from the design team but was present in the participatory design meetings to understand user needs. To guide user needs identification, the expert surveyed existing scheduling systems and anticipated problems that the user is likely to encounter using knowledge of CRCs’ workflow. Each system’s features (including those of IMPACT) were compared to this set of usability-expert derived user needs. Four relevant systems currently being used for scheduling at CUMC were included to quantify how well user needs were satisfied by each system: Microsoft Outlook Calendar, AllScripts Study Manager [36], Velos eResearch [37], and WebCAMP [38].

2.2.2. Comparison system selection

The cross-system feature alignment was used to identify a competent system to compare IMPACT with. This was done by
comparing IMPACT’s key features – user support during scheduling and calendar visualization – with those of other systems to identify the “best available” system for comparison with the intervention during phase 2.

2.2.3. Cognitive walkthroughs

Based on the previously reported CRC workflow [34], the usability expert developed scenarios for eight tasks: logging in, locating a participant, scheduling a screening visit, scheduling a randomization visit, viewing visit details, moving a visit on the calendar, updating a visit’s status, and rescheduling a visit. These scenarios ensure that the iterative evaluations stay focused on these important functionalities and remain relevant throughout the lengthy software design cycle. Appendix 1 provides sample scenarios. The usability expert counted the number of steps required by each task for both systems and noted their design differences. During phase 2, these same scenarios were used by end-users to evaluate each system.

2.3. Phase 2: End-user component

We recruited CRCs and schedulers from various clinical research settings to participate in scenario-based evaluations, 30 min or so each. This evaluation was conducted after IMPACT had undergone a 2-year participatory design process and after all key features had been implemented. We provided evaluators with a task-based scenario (the same scenario used by the usability expert in phase 1) and asked them to complete each task using IMPACT and the comparison system. We asked all evaluators to use both systems. Columbia University Medical Center’s Institutional Review Board approved this study (IRB-AAAK6000).

Our evaluation framework integrated three quantitative and two qualitative methods to assess four research constructs (Table 1). We studied usability and user acceptance at three conceptual levels: (1) human–computer interaction design, (2) team work and workflow, and (3) the socio-technical systems around Health Information Technology. We identified four research constructs and mapped them to these three levels as follows: user anxiety and user-perceived ease of use were mapped to the first level, human–computer interaction design; workflow was mapped to the second level, team work and workflow; and function integration was mapped to the third level, the socio-technical system issues. We selected each of the five methods because they contribute data for a measure that was not assessable by other methods. For instance, time-motion analysis contributes information regarding task-completion-time that could not be assessed by any of the remaining four methods. Similarly, only unstructured interviews allowed users’ “wish-list” to be elicited from evaluators. Two measures were assessed by multiple methods, namely usability and user acceptance.

Specifically, we used a think-aloud protocol to identify and record usability and interface problems [21]. Using unstructured interviews, we elicited user feedback and needs [39]. In addition, we used time-motion analysis to assess the impact of HIT on workflow [40,41] by comparing the time needed to perform a set of tasks with and without using IMPACT, since task-completion-time is related to user satisfaction [42]. We also used questionnaires to assess perceived usability [30] and user acceptance [31] and a software log to record users’ interactions with IMPACT [29]. We then triangulated data [43] from these diverse data sources.

2.3.1. Qualitative data collection

Evaluators were asked to “think aloud” while completing the scenario. Evaluators vocalized their difficulties with the system and in some cases recommended changes to IMPACT’s interface design and functionalities. At the end of the think aloud session, an unstructured interview was conducted during which users were asked about their overall impression of the application and their suggestions for improving IMPACT or the comparison system. Both
the think aloud session and unstructured interview were recorded and transcribed by a professional transcription service.\(^1\)

### 2.3.2. Quantitative data collection

During each evaluation, we collected the time spent per task using an iPad v.2.0 (Apple Inc., Cupertino, CA) application called: ATTracker [44]. User tasks were defined using prior knowledge of CRCs workflow [34]. The complete list of tasks included logging in, locating participant's visit schedule in system, scheduling a regular visit, scheduling a PRN (Pro Re Nata: Latin meaning “as needed”) visit, viewing visit details, tabbing through visit pages, updating visit status, rescheduling a visit from the visit details page, rescheduling a visit using drag-and-drop from the main calendar page, viewing reminders, changing account settings, searching for note paper, talking on phone, using REDCap [45,46], and miscellaneous activities. Data were collected for all evaluators interacting with either system.

After each scenario and subsequent interview, each evaluator completed two questionnaires: the System Usability Scale (SUS) [30] and the Unified Theory of Acceptance of Use of Technology (UTAUT) [31]. The UTAUT enhances the well-known Technology Acceptance Model [47]. Users ranked their responses on a scale of one (strongly disagree) to five (strongly agree). SUS scores were normalized so that four indicated the optimal response while zero indicated the lowest possible response.

Additionally, we analyzed IMPACT’s software log of user activities recorded during each evaluation session. Ten actions were logged: logging in, changing password, viewing calendar, viewing visit, scheduling a visit, interacting with resource optimizer, rescheduling a visit using drag-and-drop, scheduling a personal event, viewing user reminder(s), and logging out. We analyzed the action transition frequencies [41] to assess the integration of functions within IMPACT and as another assessment of IMPACT’s effect on workflow. Software logs were unavailable for the comparison system.

### 2.3.3. Mixed-methods data triangulation

We triangulated data [43] across evaluation methods to compare results obtained across methods [13] and assess their complementarity and look for convergence. We also compared results obtained across evaluation phases (usability expert vs. end-users).

### 3. Results

#### 3.1. Phase 1: Usability expert evaluation

Usability experts derived a set of user needs based on the knowledge of CRC workflow obtained through the participatory design process. The identified needs represent anticipated workflow problems encountered by CRCs during research visit scheduling. We provided three user needs as examples below.

The first user need is sharing calendars among CRCs. When a CRC is sick, research participants must be seen by another CRC. CRCs noted that it would be helpful if the covering CRC could access the entire visit schedule. Enabling CRCs to select and view any other CRC’s work calendar within the system would help address this need.

The second user need is defining CRC qualification. Some tasks performed during a research visit, such as phlebotomy, recording ECG, etc., require a research staff member certified to perform that task. An ideal system would synchronize multiple CRCs’ calendars enabling a specialized CRC to be scheduled for part of the research visit.

The third user need is scheduling a PRN visit. This can occur, for example, when a research participant arrives for a visit that requires a fasting glucose test but cannot be tested because the participant did not fast. The CRC may complete all other required tasks, e.g., blood pressure, weight, and then schedule a PRN visit for another day to perform the fasting glucose test. A system that enables scheduling of a PRN visit would be ideal.

We used these derived needs, perceived by experts, to compare several scheduling software systems in use at CUMC and IMPACT. Table 2 shows the alignment result between system functionality and scheduling needs. Only IMPACT addressed all 18 user needs, while other scheduling systems addressed at most 10.

Besides IMPACT, no existing clinical research visit scheduling system was designed specifically to suit CRCs scheduling needs. Ve-los eResearch [37] and Allscripts’ Study Manager [36] are Clinical Trial Management Systems (CTMSs) developed for billing or data management purposes and have limited scheduling functionality. This made them unsuitable baseline systems for IMPACT. We chose WebCAMP [38], developed at Weill-Cornell, as the comparison system because it shared several key functions including scheduling and calendar visualization, two main components of IMPACT. We focused on WebCAMP’s outpatient scheduling features because they were more comparable to IMPACT.

---

\(^1\) www.synergytranscriptionservices.com.
We compared the process of searching for a research participant between IMPACT and WebCAMP. In both systems, the search begins by entering a participant's name into a search box. IMPACT (Fig. 2a) displays a list of matching participants directly below the search bar. After selecting a participant, the user is shown that participant's next upcoming unscheduled visit. In contrast, WebCAMP (Fig. 2b) displays a pop-up window with a list of related search results. After the user selects the participant, the participant's history of visits is shown in another pop-up window. The user must close each pop-up window to return to the main calendar, where the user can then schedule a visit. The user cannot see the next visit required by the protocol. WebCAMP uses three pop-up windows while IMPACT uses only one screen transition to achieve equivalent functionality.

3.2. Phase 2: End-user evaluation

We recruited 12 evaluators from CUMC who were unfamiliar with IMPACT before and who represented a balanced distribution of males (5 users) and females (7 users), because gender and user response are sometimes related on usability questionnaires [48]. Most of our evaluators (75%) were moderately experienced (2–8 years), and the remaining 25% were highly experienced (15–20 years). Because of our affiliation with the division of Cardiology, 50% of evaluators were from Cardiology and 25% from Behavioral Cardiology. To improve the heterogeneity of end-users in the sample, we recruited two end-users specialized in cancer studies and one in diabetes research. All 12 evaluated IMPACT. However, because of CRCs time constraints, only three IMPACT evaluators were also able to evaluate WebCAMP. All were males from Cardiology with moderate experience.

We summarized evaluators' comments recorded during subsequent interviews or upon scenario completion (i.e., “think aloud” protocol) into a “wish list” that users would like to be addressed in a future scheduling system and a list of suggested system improvements shown in Table 3. The “wish-list” and “suggestions for improvement” provided additional information for the development team. Four key usability improvements were recommended by users: (1) highlighting important information by changing information display, (2) achieving system compliance with HIPAA, (3) automatically populating default values for coordinator name and current date and time when scheduling a new research visit, and (4) making buttons look more like a button rather than like text. All four were incorporated into IMPACT as part of its development update.

Eight of 12 evaluators provided feedback pertaining to particular system features. We organized the feedback distinguishing between specialized (unsupported by other systems) and general (e.g., adding or deleting a task) system features shown in Table 4. Some evaluators liked features that were unsupported by other systems, such as scheduling multiple coordinators simultaneously, stating, “This is very helpful. And on top of that, you could have access on the Internet. You could go anywhere and have access.” Others enjoyed IMPACT’s ability to combine protocol-specified visit-specific information into the visit schedule stating that IMPACT is “useful for tracking down the checklist for the patient.” One evaluator perceived IMPACT’s usefulness for multiple study coordination by stating that, “you have to be on multiple studies and you have lot of scheduled things...that have different (tasks) ...then you can have the list and the other list and (check) if things are incomplete.”

Evaluators using IMPACT provided statements such as, “I think it can work. I think it’s great...it’s pretty simple.” However, evaluators using WebCAMP reported much frustration and anxiety. One such evaluator said, “It’s too small. I can’t click in there...It won’t let you go back” Another stated, “Those squares are microscopic...You can’t see the name; it’s too many steps just to input one appointment...I can’t use this system.”

3.3. Triangulating results

Triangulating results from across methods allowed us to assess how well IMPACT performed [13,43]. Table 5 shows a subset of

---

**Table 2**

Cross-system alignment between system functionality and scheduling needs. X indicates that the corresponding system provides functionalities to satisfy the selected user needs.

<table>
<thead>
<tr>
<th>System</th>
<th>Visits</th>
<th>Participants</th>
<th>Import existing participants</th>
<th>Import clinical appointment for patients</th>
<th>Specify patient scheduling preferences</th>
<th>Multiple coordinators</th>
<th>Other scheduling needs</th>
<th>Reminders and notifications</th>
<th>Resource allocation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scheduling needs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microsoft Outlook calendar</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Allscripts Study manager</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Velos eResearch</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Cornell WebCAMP</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>CUMC IMPACT</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

---

- * means that this need was not fully incorporated into the evaluation prototype. Bold text indicates a feature unique to IMPACT. PRN: Pro Re Nata (as needed).
- a Function supports workflow flexibility.
- b Function facilitates collaboration and group awareness.
- c The recent release of eResearch [37] supports visualization of research appointments.
evaluation results obtained from triangulating IMPACT’s evaluation data using our framework. Each method contributes a piece of unique information that affects a research construct. IMPACT’s ease-of-use (or lack thereof) was assessed by four of five methods. Our framework revealed quantitative differences in task-completion time for scheduling, high usability scores from the SUS questionnaire, and so on. The think-aloud protocol (qualitative method) also revealed indicators of ease-of-use for scheduling. By triangulating results from these methods, we gained greater confidence in assessing IMPACT’s ease-of-use. We only include a subset of evaluation results in Table 5 (one main result per method) as the focus of this paper is on our evaluation framework.

Likewise, the system’s integration of functions and users’ anxiety during use can be assessed (Table 5). The most complex research construct, effect on workflow, was affected by the results of the other three constructs, i.e., ease-of-use, integration.
4.2. Generalizability of the framework

4.3. Confirming results between phase 1 and 2

Our framework enabled findings to be validated across phases. Phase 1, with the usability expert, allowed us to assess the differences among systems with regard to the number of steps that are required to schedule a visit and in interface designs through indicators of ease-of-use (Table 1). In our application of this framework on IMPACT (using only a subset of results in Table 5), phase 1 revealed a reduction in the number of steps to schedule a visit, and a more streamlined interface with one screen transition for finding a participant as opposed to three pop-ups in WebCAMP. Phase 2, with end-users, allowed quantification of ease-of-use using time-motion analysis, SUS questionnaire, and the think-aloud protocol. Phase 2 revealed that IMPACT uses 60 s less than WebCAMP to reschedule a visit (time-motion analysis), scores for frequency of use increased with IMPACT, 3.33 vs. 1.33 (SUS questionnaire), and CRCs enjoyed IMPACT’s ability to schedule a visit to multiple CRCs (think-aloud protocol) (Table 5).

4.4. Discussion

4.4.1. Generalizability of the framework

Formative evaluation of HIT prototypes is especially necessary during development as clinical needs are both implicit and complex [49–51]. Prototype evaluation allows developers to probe users to elicit and refine their needs. Our frameworks enabled the validation of expert-derived user needs, elicitation of functions, and anxiety. Assessment of performance and effort expectancy from the UTAUT questionnaire were also added to further assess the effect on workflow.

3.4. Confirming results between phase 1 and 2

Our framework enabled findings to be validated across phases. Phase 1, with the usability expert, allowed us to assess the differences among systems with regard to the number of steps that are required to schedule a visit and in interface designs through indicators of ease-of-use (Table 1). In our application of this framework on IMPACT (using only a subset of results in Table 5), phase 1 revealed a reduction in the number of steps to schedule a visit, and a more streamlined interface with one screen transition for finding a participant as opposed to three pop-ups in WebCAMP. Phase 2, with end-users, allowed quantification of ease-of-use using time-motion analysis, SUS questionnaire, and the think-aloud protocol. Phase 2 revealed that IMPACT uses 60 s less than WebCAMP to reschedule a visit (time-motion analysis), scores for frequency of use increased with IMPACT, 3.33 vs. 1.33 (SUS questionnaire), and CRCs enjoyed IMPACT’s ability to schedule a visit to multiple CRCs (think-aloud protocol) (Table 5).

4. Discussion

4.1. Generalizability of the framework

Formative evaluation of HIT prototypes is especially necessary during development as clinical needs are both implicit and complex [49–51]. Prototype evaluation allows developers to probe users to elicit and refine their needs. Our frameworks enabled the validation of expert-derived user needs, elicitation of functions, and anxiety. Assessment of performance and effort expectancy from the UTAUT questionnaire were also added to further assess the effect on workflow.

3.4. Confirming results between phase 1 and 2

Our framework enabled findings to be validated across phases. Phase 1, with the usability expert, allowed us to assess the differences among systems with regard to the number of steps that are required to schedule a visit and in interface designs through indicators of ease-of-use (Table 1). In our application of this framework on IMPACT (using only a subset of results in Table 5), phase 1 revealed a reduction in the number of steps to schedule a visit, and a more streamlined interface with one screen transition for finding a participant as opposed to three pop-ups in WebCAMP. Phase 2, with end-users, allowed quantification of ease-of-use using time-motion analysis, SUS questionnaire, and the think-aloud protocol. Phase 2 revealed that IMPACT uses 60 s less than WebCAMP to reschedule a visit (time-motion analysis), scores for frequency of use increased with IMPACT, 3.33 vs. 1.33 (SUS questionnaire), and CRCs enjoyed IMPACT’s ability to schedule a visit to multiple CRCs (think-aloud protocol) (Table 5).

4. Discussion

4.1. Generalizability of the framework

Formative evaluation of HIT prototypes is especially necessary during development as clinical needs are both implicit and complex [49–51]. Prototype evaluation allows developers to probe users to elicit and refine their needs. Our frameworks enabled the validation of expert-derived user needs, elicitation of functions, and anxiety. Assessment of performance and effort expectancy from the UTAUT questionnaire were also added to further assess the effect on workflow.

3.4. Confirming results between phase 1 and 2

Our framework enabled findings to be validated across phases. Phase 1, with the usability expert, allowed us to assess the differences among systems with regard to the number of steps that are required to schedule a visit and in interface designs through indicators of ease-of-use (Table 1). In our application of this framework on IMPACT (using only a subset of results in Table 5), phase 1 revealed a reduction in the number of steps to schedule a visit, and a more streamlined interface with one screen transition for finding a participant as opposed to three pop-ups in WebCAMP. Phase 2, with end-users, allowed quantification of ease-of-use using time-motion analysis, SUS questionnaire, and the think-aloud protocol. Phase 2 revealed that IMPACT uses 60 s less than WebCAMP to reschedule a visit (time-motion analysis), scores for frequency of use increased with IMPACT, 3.33 vs. 1.33 (SUS questionnaire), and CRCs enjoyed IMPACT’s ability to schedule a visit to multiple CRCs (think-aloud protocol) (Table 5).

4. Discussion

4.1. Generalizability of the framework

Formative evaluation of HIT prototypes is especially necessary during development as clinical needs are both implicit and complex [49–51]. Prototype evaluation allows developers to probe users to elicit and refine their needs. Our frameworks enabled the validation of expert-derived user needs, elicitation of functions, and anxiety. Assessment of performance and effort expectancy from the UTAUT questionnaire were also added to further assess the effect on workflow.

3.4. Confirming results between phase 1 and 2

Our framework enabled findings to be validated across phases. Phase 1, with the usability expert, allowed us to assess the differences among systems with regard to the number of steps that are required to schedule a visit and in interface designs through indicators of ease-of-use (Table 1). In our application of this framework on IMPACT (using only a subset of results in Table 5), phase 1 revealed a reduction in the number of steps to schedule a visit, and a more streamlined interface with one screen transition for finding a participant as opposed to three pop-ups in WebCAMP. Phase 2, with end-users, allowed quantification of ease-of-use using time-motion analysis, SUS questionnaire, and the think-aloud protocol. Phase 2 revealed that IMPACT uses 60 s less than WebCAMP to reschedule a visit (time-motion analysis), scores for frequency of use increased with IMPACT, 3.33 vs. 1.33 (SUS questionnaire), and CRCs enjoyed IMPACT’s ability to schedule a visit to multiple CRCs (think-aloud protocol) (Table 5).

4. Discussion

4.1. Generalizability of the framework

Formative evaluation of HIT prototypes is especially necessary during development as clinical needs are both implicit and complex [49–51]. Prototype evaluation allows developers to probe users to elicit and refine their needs. Our frameworks enabled the validation of expert-derived user needs, elicitation of functions, and anxiety. Assessment of performance and effort expectancy from the UTAUT questionnaire were also added to further assess the effect on workflow.

3.4. Confirming results between phase 1 and 2

Our framework enabled findings to be validated across phases. Phase 1, with the usability expert, allowed us to assess the differences among systems with regard to the number of steps that are required to schedule a visit and in interface designs through indicators of ease-of-use (Table 1). In our application of this framework on IMPACT (using only a subset of results in Table 5), phase 1 revealed a reduction in the number of steps to schedule a visit, and a more streamlined interface with one screen transition for finding a participant as opposed to three pop-ups in WebCAMP. Phase 2, with end-users, allowed quantification of ease-of-use using time-motion analysis, SUS questionnaire, and the think-aloud protocol. Phase 2 revealed that IMPACT uses 60 s less than WebCAMP to reschedule a visit (time-motion analysis), scores for frequency of use increased with IMPACT, 3.33 vs. 1.33 (SUS questionnaire), and CRCs enjoyed IMPACT’s ability to schedule a visit to multiple CRCs (think-aloud protocol) (Table 5).

4. Discussion

4.1. Generalizability of the framework

Formative evaluation of HIT prototypes is especially necessary during development as clinical needs are both implicit and complex [49–51]. Prototype evaluation allows developers to probe users to elicit and refine their needs. Our frameworks enabled the validation of expert-derived user needs, elicitation of functions, and anxiety. Assessment of performance and effort expectancy from the UTAUT questionnaire were also added to further assess the effect on workflow.
unanticipated users' needs (e.g., cell phone app), collection of users' perceptions of the system. Our framework also enabled the characterization of IMPACT's potential for reducing workarounds by addressing more user needs, perceived by usability experts, when compared to other existing systems and by assessing IMPACT's effect on CRC workflow. Reducing workarounds and improving workflow are two key components for successful implementation of a decision support system [52, 53] for the clinical research environment.

Many CRCs expressed the need for a system to aid them in scheduling research participants but had difficulty articulating their needs as implementable system functionalities for the development team. When users interacted with the functional prototype during the evaluation, they were able to identify aspects of the system's interface and functionality they liked and disliked. They were also better able to express what they would like to see ("wish-list"). This experience is not unique to our assessment [54]. We posit two primary explanations for this phenomenon in the clinical research environment. First, our users (mainly CRCs) are familiar with performing many tasks using tedious manual processes. This has become the status quo and many users are unaware of the possible benefits HIT can bring to their workplace making it difficult to elicit their technological needs. Second, some CRCs experienced more anxiety using the comparison system.

Results are shown in parentheses [ ]. The IMPACT result is shown first followed by the WebCAMP result according to the following form: (IMPACT result vs. WebCAMP). All results follow this form. For example, 'Schedule a visit (5 vs. 17 steps)' indicates that scheduling a visit required 5 steps in IMPACT and 17 steps in WebCAMP. Another example, 'Freq. of use (SUS) (3.33 vs. 1.33)' indicates that the SUS measure for frequency of use was 3.33 when using IMPACT vs. 1.33 when using WebCAMP.

Table 5

Example of data triangulation across methods by research construct.

<table>
<thead>
<tr>
<th>Research construct</th>
<th>Phase 1 (usability expert)</th>
<th>Phase 2 (end-user)</th>
<th>Qualitative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Quantitative</td>
<td>Quantitative</td>
<td>Qualitative</td>
</tr>
<tr>
<td>Task analysis</td>
<td>Interface Design</td>
<td>Time-motion analysis</td>
<td>Think-aloud protocol</td>
</tr>
<tr>
<td>Ease-of-use</td>
<td>Schedule a visit (5 vs. 17 steps)</td>
<td>Find a participant (1 screen transition vs. 3 pop-ups)</td>
<td>Rescheduling a visit from the main calendar (60 s. less in IMPACT)</td>
</tr>
<tr>
<td>Function integration</td>
<td>Find a participant link difficult to locate in WebCAMP</td>
<td>IMPACT's central action is viewing the calendar</td>
<td>Integration of functions (SUS) (3.33 vs. 0.67)</td>
</tr>
<tr>
<td>Anxiety</td>
<td></td>
<td>Anxiety (UTAUT) (1.58 vs. 2.67)</td>
<td>Performance (3.78 vs. 2.22) and effort expectancy (UTAUT) (4.33 vs. 2.83)</td>
</tr>
<tr>
<td>Workflow</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results are shown in parentheses [ ]. The IMPACT result is shown first followed by the WebCAMP result according to the following form: (IMPACT result vs. WebCAMP). All results follow this form. For example, 'Schedule a visit (5 vs. 17 steps)' indicates that scheduling a visit required 5 steps in IMPACT and 17 steps in WebCAMP. Another example, 'Freq. of use (SUS) (3.33 vs. 1.33)' indicates that the SUS measure for frequency of use was 3.33 when using IMPACT vs. 1.33 when using WebCAMP.

4.2. Knowledge gained when using the framework

Like many HIT interventions, IMPACT contains features that are not available in existing related systems, which made it difficult to compare it with existing systems. This precludes the use of traditional pre-/post-design evaluation methods since no baseline system existed for such comparisons [2, 4, 1, 5, 6, 5, 7]. Identifying a suitable baseline or comparison system for evaluation purposes is a challenging problem for emerging HIT interventions [7]. Using our cross-system alignment between system function and user need enabled us to select a suitable comparison system for the IMPACT prototype. This alignment was necessary for selecting the system, which was then used in phase 2 to compare with IMPACT. We chose WebCAMP as the "best available" comparison system because it possessed both scheduling functionality and visualization, two key features of IMPACT.

Two methods are used in phase 1 and five in phase 2 of the evaluation framework. Overall, five methods contributed data for a measure that was unique (Table 1). We triangulated measures to assess four research constructs: ease-of-use, integration of functions, anxiety during use and effect on workflow. Each of the four constructs is multi-faceted requiring multiple methods to capture various aspects of it. For example, ease-of-use can be assessed using the SUS questionnaire. However, system usability is affected by many factors and is complex [58]. Also, questionnaires like SUS generally measure users’ perceived system usability and may not be measuring the ‘actual’ [33]. Therefore, measuring the time to complete key tasks and analyzing the interface design bolster our understanding of the system’s actual vs. perceived ease-of-use. By triangulating methods, especially by combining qualitative
and quantitative results, we strengthen our results of the system’s performance along the research constructs.

4.3. Limitations

One limitation of this study is that our cross-system function vs. needs alignment was performed using systems currently in use at CUMC. Other CTMSs and scheduling systems that are not used by CUMC were excluded from this study. However, this limitation should not affect the overall evaluation framework. A second limitation is that we used scenario-based evaluations. Scenarios guide users to perform certain tasks within each system. This type of step-by-step guidance for performing tasks is not present in the “real world”. Ideally, users should evaluate systems during their regular workflow. While we used scenario-based evaluations to test our framework on IMPACT, we believe that our two-phase mixed-methods evaluation framework would also be useful for “real world” evaluations in situ.

5. Conclusions

We present a two-phase evaluation framework that combines evaluations by usability experts (phase 1) and evaluations by end-users (phase 2). Our framework is particularly relevant for early prototype evaluations for emerging HIT interventions, when users are unclear about their needs and when a baseline is lacking, both frequently encountered problems. Our framework enables an integrated assessment of user needs identification by usability experts (phase 1) and user needs refinements by end-users (phase 2). By triangulating results from mixed-methods, our framework measures the ease-of-use, integration of functions, user anxiety, and workflow impact.

Funding

The research described was supported by grants R01LM010815 and R01LM009886 from the National Library of Medicine, R01HS019853 from the Agency for Healthcare Research and Quality, and UL1 TR000040 from the National Center for Advancing Translational Sciences.

Competing interests

The authors declare that they have no competing interests.

Contributions

CW conceptualized the evaluation framework, supervised the evaluation study, and wrote the manuscript. MRB performed data collection and analysis and wrote an initial draft of the manuscript. AR participated in the evaluation research and edited the manuscript. YS developed the software log used in this evaluation. RCS, CL, LB contributed ideas to the design and provided feedback on the manuscript. JTB and SB provided substantive methodological contributions to the manuscript.

Acknowledgments

We would like to thank the Clinical Research Coordinators from CUMC and NYP who evaluated the IMPACT prototype. We would also like to thank Matthew Paulson for allowing us to observe his interactions with Velos eResearch.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2013.12.004.

References

[16] Cimino JJ, Li J, Bakken S, Patel VL. Theoretical, empirical and practical approaches to resolving the unmet information needs of clinical information system users. AMIA Annu Symp Proc 2002;172(9).