Taking the Long View: Structured Expert Evaluation for Extended Interaction

Alan Dix
Computational Foundry
Swansea University, Swansea, Wales, UK
alan@hcibook.com

ABSTRACT

This paper proposes first steps in the development of practical techniques for the expert evaluation of long-term interactions driven by the need to perform expert evaluation of such systems in a consultancy framework. Some interactions are time-limited and goal-driven, for example withdrawing money at an ATM. However, these are typically embedded within longer-term interactions, such as with the banking system as a whole. We have numerous evaluation and design tools for the former, but long-term interaction is less well served. To fill this gap new evaluation prompts are presented, drawing on the style of cognitive walkthroughs to support extended interaction.

CCS CONCEPTS

• **Human-centered computing**: HCI design and evaluation – Walkthrough evaluations • **Human-centered computing**: Systems and tools for interaction design

KEYWORDS

Long-term interaction, expert evaluation, cognitive walkthrough, interaction design

ACM Reference format:

Alan Dix. 2020. Taking the Long View: Structured Expert Evaluation for Extended Interaction. In *Proceedings of the International Conference on Advanced Visual Interfaces (AVI '20). ACM, New York, NY, USA, 9 pages.* https://doi.org/10.1145/3399715.3399831

1 INTRODUCTION

Some interactions with digital systems, for example getting cash from an ATM, are short and constrained: a number of steps are executed, screens or pages viewed, and the user's goal is achieved. However, our interactions with many systems extend over days, weeks or years. Even an ATM is revisited multiple times and is part of a broader interactive experience of a banking system, which includes both online and face-to-face sessions.

We have many tools to conceptualise and design micro-interactions from Norman's seven stage model [39] to heuristic evaluation [37] and cognitive walkthroughs [45], and this scale is also well served by practical development techniques such as user stories in agile processes. We also have theoretical frameworks for more extensive interactions such as activity theory [25], as well as classic user-requirements methods combining interviews and workplace observations. However, practical tools are less well developed for the design and evaluation of long-term interactions, where challenging issues such as re-engagement and multiple interfaces are critical.

This paper attempts to fill this gap; building on long-term theoretical research on long-term interactions, a framework and set of evaluation prompts are presented, which draw on the style of cognitive walkthroughs. These prompts aim to aid designers in performing early evaluation of systems that involve multiple periods of intense interaction, spaced over long periods. The prompts have been validated in a commercial context.

In the next section we explore in more detail single phases of direct interaction and the longer-term multi-phase interactions of which they form episodes of use. Section 3 reviews key theories and models of long-term interaction and the following section will recap the standard cognitive walkthrough, which together form the theoretical and practical background for the novel multi-phase prompts described in the section 5. The new prompts have been used in a project with a major industrial client and the paper ends looking at some of the insights from this and the implications for further development of the tools for long-term interaction.

2 FROM SINGLE-PHASE TO MULTI-PHASE INTERACTION

One can view an interaction at multiple scales. Motor-level models [33] such as Fitts' Law [16] and Card, Moran and Newell's keystroke-level model [5], have proved very powerful for the finest level of interaction with Fitts' Law being generalised for many types of interaction techniques beyond simple mouse movement [20].

At a slightly wider level Norman's, classic, seven-stage model of interaction takes into account the action-by-action intentions of the user and how these are constantly monitored and adapted in a

 ISBN 978-1-4503-7535-1/20/09…$15.00
https://doi.org/10.1145/3399715.3399831

tight loop of interactions [39]. Although not necessarily referring to Norman's model, the concept of tightly focused interaction cycles is deeply embedded in interaction design.

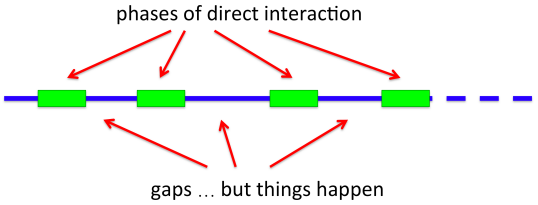

Figure 1: Multi-phase interaction (CC-BY, Alan Dix)

As noted these tight periods of focused use of a digital system are usually part of a far larger pattern of activity. These phases of intense interaction are often separated by gaps (Figure 1) where there is either no activity related to the system, or off-line activity, or both. Understanding these gaps is important, because:

1. There may be significant activity related to the system happening – for example, in an university learning environment, the student may have attended lectures, read books or written an essay
2. The user needs to re-engage with the system, this raises many issues, for example re-establishing context, and indeed whether the user re-engages at all.

However, despite this ubiquity of long-term interaction the core design tools and methods are largely focused on screen-to-screen interactions within each phase. In Taylor Palmer's annual Design Tools Survey of more than 3000 UX practitioners [41], the popular design tools they use are focused almost entirely at the level of screen design and the flow between screens or web pages. Even storyboarding and user stories seem to be free-hand activities. Crucially compared with Campos and Nunes 2007 survey [4], tool use seems to be far more central to a modern UX designer's practice, and so the fact that tools do not actively support broader physical and temporal context is disturbing.

User testing of long-term interaction is of course hard. By definition an in-lab study is of limited duration and may include short interruptions, but cannot easily mimic days or weeks of interaction. Sometimes it is possible to design short-term tasks that create effects normally seen in long-term uses [9], but this is rare, the best practice is often to use interviews or occasionally bring back participants after a period of a few weeks. In the wild studies [47, 6], as well as addressing the broader physical and social context, are more able to address longer-term phenomena, and long-term deployments feature in both academic research and commercial development. However, these are typically even more expensive than lab-based studies, especially if a representative number of users are required.

The exception to costly long-term evaluation is where trace data can be obtained for analysis. This has been used for many years in an investigative manner to uncover long-term trends and patterns [51], but has become far easier in web deployments and where connectivity of mobile apps allows widespread data gathering. A–B testing has reached a level where many small-decisions such as choice of colours and item placement is a matter of empirically-informed automated decisions rather than UX expertise [32]. However, paradoxically, the long-term data of A–B testing is most strongly directed towards these fine-scale and short-timescale design choices.

Various forms of expert or discount evaluation are often used, especially for early evaluation where the cost of design changes are more acceptable. Most well known amongst these are Neilsen's heuristic evaluation [37, 52] and forms of cognitive walkthrough [31, 45]. Both of these are focused at the within screen and screen-to-screen flow of interaction, in their normal forms they can be used for single phases, but not the wider picture of long-term interaction. Later in the paper we will describe cognitive walkthroughs in more detail and then how these can be adapted for more extended interaction by the use of additional prompts and steps. However, these prompts are themselves based on existing work on models and theories for long-term interaction.

3 MODELS AND THEORIES OF LONG-TERM INTERACTION

Although most evaluation and design methods are focused on shorter-term interaction, the broader methods of inquiry with HCI and interaction design are applicable to longer timescales. User-centred design is about obtaining a rich understanding of the users and the context in which they operate. Techniques for understanding users such as persona development [7, 38] apply equally well for nearly any scale of activity although, while in principle scenarios can extend beyond a single focused interaction, in practice they are often more commonly applied to fairly short time-scale activities [22, 1].

Just as with system evaluation, it is hard to study long-term interactions by direct observation, but this is possible, certainly for periods of days or weeks. Some of the most successful ethnographic studies, for example, Heath and Luff's analysis of the London Underground control room [21], have included multiple whole day observations. In a user research setting shadowing someone for an extended period can be very costly, but may also have substantial payoffs, as in Alison Kidds' seminal "The marks are on the knowledge worker" [27] and diary methods effectively recruit the subjects of research as active participants [29, 24]. Physical and digital artefacts such as meeting minutes or documents scattered on a desk can, in combination with shorter observations and user interviews, be used to piece together a larger-scale view just as a dendrologist will understand the growth and development of trees even though their life-span may be hundreds of years [46].

There are a wide variety of task analysis methods [8] and the majority of practical use is focused on single periods of use; indeed hierarchical task analysis (HTA) [49] had some of its most successful early use in creating training for highly procedural tasks such as assembling a rifle [36]. Although most frequently used for short-term tasks, some notations, notably ConcurTaskTrees (CTT) [42, 43] have provision for off-line single user and multi-user activities interspersed amongst computer-focused actions, and have been used to analyse, activities such as flight control and holiday booking. Similar

techniques to task analysis are often applied in information systems and management science to document large-scale organisational workflows and are a key part of techniques such as business process re-engineering [34] and its more recent offshoots.

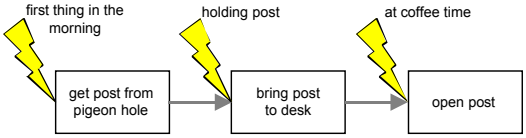

Figure 2: Triggers for actions in a process (from [10])

Trigger analysis extends standard task-analysis or workflow models by focusing on the issues that arise when a sub-task that logically follows on from another is not necessarily performed immediately [10, 12]. In some cases these gaps are planned (maybe reading post picked up first thing over morning coffee, see Fig. 2) and sometimes unplanned (interruptions). There are classes of failures in both cases that are due to the delayed action never being executed. Trigger analysis focuses on the *triggers* that make an action happen when it does, and the potential ways in which the larger process of which it is a part is, or is not, resilient if triggers fail. It includes a rich analysis of different types of triggers and potential failures for each (Fig. 3). In the context of multi-phase interactions (Fig. 1), we need to understand the trigger that operates after a gap to initiate the next phase of direct interaction.

- ***immediate***: –activity begins immediately after previous one.
- ***temporal*** – periodic or actions after a particular delay
- ***sporadic*** – when the person remembers
- ***external event*** – alarm, notification or specific event.
- ***environmental cue*** – things that remind us that something ought to be done, whether explicitly recorded or implicit.

Figure 3: Types of Trigger from [12]

Within HCI and CSCW, activity theory [25] is probably the most widely used theoretical framework that emphases the larger arc of human engagement with technology. Activity theory operates at three levels: activity, action and operation. The *action* level roughly corresponds to the overall unit of task analysis: time-bounded with a clear and typically explicit goal. The *operation* level is roughly the lowest level of task analysis, using a tool, or filling in a field on a computer form and corresponds to the unit of the Normal seven-stage model. However, the highest level, *activity*, is most interesting for this paper. Whereas a university student may attend a lecture or complete a piece of course work, these time-limited and goal-directed actions are set within three or more years of study or even a lifetime of education, the activity. These high-level activities typically have more diffuse motives (such as being a more educated or employable person) rather than easily measurable goals.

Frameworks for experience and emotion typically also have an element of longevity as our emotional experience of a system or context build slowly over time (e.g. [2, 28]). McCarthy and Wright [35] emphasise the way that a single experience is part of a larger temporal pattern that includes anticipation and memory. Studies of photologging suggested that these episodes of focused experience (corresponding to the phases in Fig. 1), link together leading to *extended episodic interaction* [26], where each episode builds on the emotional experience of previous episodes and frames itself within anticipation of future interactions.

Finally, but by no means least, service design has become increasingly important as service industries have come to dominate both online and offline commerce [23, 17]. Service design by its nature stretches over the extended period during which an organisation engages with a customer to deliver a service: whether of fixed duration, such as going on a holiday, or unlimited, such as banking. Core to service design are the ideas of *touchpoints*, the particular times when contact is made between organisation and customer, and the *customer journey*, the story of how the service is provided to a customer through multiple touchpoints. From an interaction design perspective, service design has both 'front of stage' interactions with the customer or end-user and 'back stage' interactions within the organisation.

So, every service design will involve one or more long-term, multi-phase interactions, where roughly the touchpoints correspond to the phases of direct interaction (with system or people) in Figure 1. However, there are long-term interactions that would not normally be regarded as service design, for example, the way you might use a word processor and other tools to write a paper for a conference or a book, potentially over many weeks or even years.

4 STANDARD COGNITIVE WALKTHROUGH

The cognitive walkthrough is a widely used expert evaluation technique. It was originally conceived by Lewis, Polson, Rieman, and Wharton [31, 45] based on exploratory learning theory and subsequently developed by themselves and others as a practical usability method [54, 50, 3]. The aim of the cognitive walkthrough is to create a series of well-defined steps that user interface experts can work through in order to guide them in working out whether a user will have the required information and resources to achieve their goals.

The details of the walkthrough vary from source to source and between practitioners, but consist of two main stages. Prompts for these are given in Figures 4 and 5 which are themselves based on [45, 54, 11]. The first analysis stage (Fig. 4) is about describing the overall system (what), the intended users (who), the tasks they are expected to be performing (why) and the concrete actions needed to achieve these (how). The list of actions can be thought of as a simple task analysis.

1. specification or prototype of the system
2. description of the users
3. description of the task the user is to perform
4. complete, written list of actions needed

Figure 4: Standard cognitive walkthrough – System Description (based on [45, 54, 11])

In the second walkthrough stage, the action sequences are followed through one by one. For each action the evaluator considers a number of prompt questions (Fig. 5) designed to expose whether the action meets the actual goals of the user and how difficult it is for the user to follow the right action.

goal match
Is the effect of the action the same as the user's goal at that point?
visibility
Will users see that the action is available?
identification
Once users have found the correct action, will they know it is the one they need?
feedback
After the action is taken, will users understand the feedback?

Figure 5: Standard cognitive walkthrough – Per-Action Prompts (based on [45, 54, 11])

The analyst is encouraged to think of both typical and worst case situations: for example the difference between an expert user who knows what actions are available and a new user to a system. This distinction was very obvious during the writing of this paper as the ACM digital library changed its interface. The author knew that it must be possible to get a formatted citation for a listed article, and this was indeed the case (goal match), but the actions are now simply shown as icons (visibility), so the only way to find the right action was by hovering over each in turn. Happily a tool tip confirmed the correct icon (identification, Fig. 6) and clicking it opened a box with the citation (feedback).

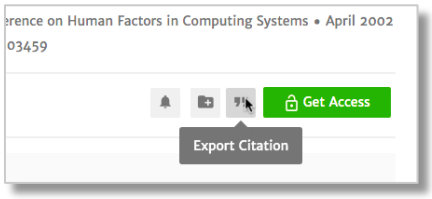

Figure 6: ACM DL – finding a formatted citation.

5 A WALKTHROUGH FOR MULTI-PHASE INTERACTION

Recalling Figure 1, a standard cognitive walkthrough is focused on the phases of direct interaction. There are typically different kinds of phases of direct interaction. In an online learning system there may be learning-focused periods of interaction phases where the student is watching videos, but also assessment interactions involving timed tests. Similarly for a banking application visiting an ATM to get cash out, viewing an online bank statement, or making a mobile-app payment. Standard cognitive walkthrough checklist can be applied to each of these individually. However, there is also the larger arc of long-term interaction and additional prompts are required to deal with this including the gaps between direct interactions (Fig. 7).

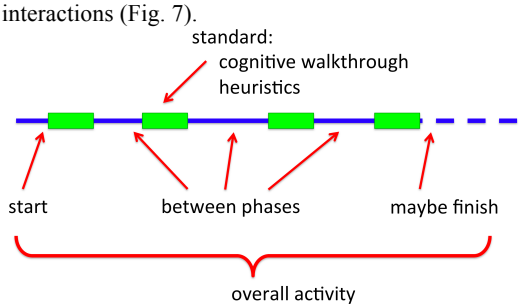

Figure 7: Additional places for walkthrough (CC-BY, Alan Dix)

The extended walkthrough is divided into five parts that can be seen in Figure 7.

- ***overall activity*** – These are a set of questions, building on Figure 4, that are about the wide arc of system use.
- ***per-phase prompts*** – This is the standard cognitive walkthrough (Fig. 5) applied to each type of phase
- ***start prompts*** – The per-phase prompts may concern special start-up actions (e.g. registration or opening a bank account). In contrast, the start prompts are about the lead up to this, why and how the user starts using the system in the first place.
- ***between phase prompts*** – Similarly, these are about the reasons and means by which a user re-engages after a period when they are doing other activities.
- ***end prompts*** – Finally, for some systems, such as email, there may be no planned end, but for others there may be a clear end point such as graduating from an online course or closing a bank account.

We will look at each in turn. The new prompt cards in Figures 8–12 are provided[1] in addition to the standard cognitive walkthrough prompts in Figures 4 and 5.

5.1 Overall activity

Prompt cards in Figures 8 and 9 relate to the overall system and an overview of the wide arc of the activity with it.

Starting with Figure 8, it builds on the standard prompts in Figure 4. Steps 1 and 2 of Figure 4, refer to the system as a whole: step 1 "*specification of prototype of the system*" is so that we have something to evaluate and step 2 "*description of the users*" is so that the expert performing the walkthrough knows who will be using the system and hence the level of knowledge,

[1] See also https://alandix.com/longterminteraction/ for prompt cards and furher resources related to the extended walktrhough.

skills, etc. to expect. These are therefore performed once for the system as a whole (Fig. 8).

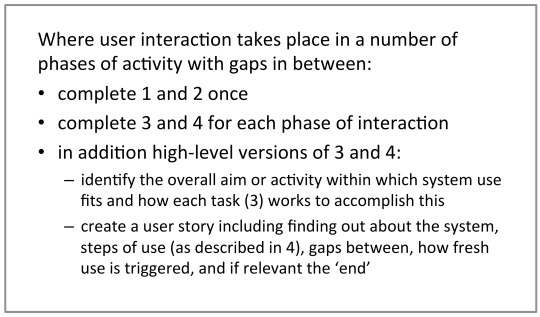

Figure 8: Multi-phase walkthrough –System Prompts (numbers refer to standard CW system description, Fig. 4)

The normal steps 3 and 4 in a cognitive walkthrough (Fig. 4) relate to individual phases of interaction, and are returned to below. However, they suggest high-level variants that are performed once for the overall activity.

The first (penultimate bullet, Fig. 8), "*identify the overall aim or activity ...*" helps the analyst contextualise the entire arc: for a banking system this might be about managing finances, for a learning system it may be about obtaining a qualification or building personal skills. The focus here is on the user's aims, but many long-term interactions involve multiple users, so the analyst may need to think about groups of users.

The second (last bullet, Fig. 8), "*create a user story ...*" asks the analyst to describe the overall use of the system over the full period of use. There may be different patterns of use, so this could involve a formal description with alternatives and multiple branches; however, for the purposes of a walkthrough this should be reduced to one or more linear narratives. This effectively corresponds to describing the *customer journey* in service design [17]. The user story will include some episodes that are about phases of direct interaction, some about activity in the gaps, and some about the transitions between the two.

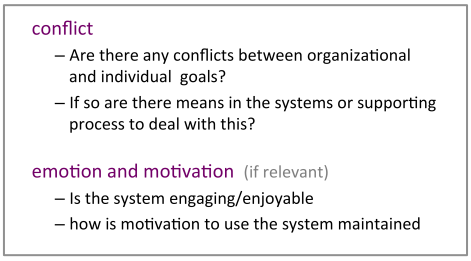

Figure 9: Multi-phase walkthrough – Overall Activity

The second new prompt card in Figure 9 adds elements that arguably should be present in a standard cognitive walkthrough.

The first heading "*conflict*" is suggested by the experience within CSCW research that there may be conflicting aims between the organisation that choses the use of a particular system, and the end users [15]. This is particularly true of a system that has in some way been imposed on employees or others working with an organisation. A good example of this is the virtual learning environment (VLE) in a university for a member of staff suggesting an item on a reading list the primary aim may be the education of the student, but the university may also trade this against monetary costs of accessing different forms of copyright material.

The analyst is encouraged to identify such conflicts and if they exist the extent to which the system has managed or mitigated the risks. The CSCW literature is replete with stories of systems that fail because such issues have not been adequately addressed [19].

The second heading "*emotion and motivation*" will vary in its relevance. All systems need some level of motivation to use, but this could be quite crude. If you are a checkout assistant in a supermarket, you use the tills provided because that is part of the job. In contrast, if the system is in some way voluntary, such as on-the-job training courses, maintaining on-going motivation is critical. Of course good emotional engagement is always essential, but, as we have seen in the theory section, this is especially important when we consider repeated use of a system.

5.2 Per-phase prompts

Each type of direct interaction phase needs to be identified. Framing these will be aided by the overall user story (above). Crucially each episode in the overall user story that concerns direct interaction will need a corresponding detailed walkthrough.

Once this has been done, steps 3 and 4 of the standard cognitive walkthrough prompts in Figure 4 need to be performed for each different type of direct interaction phase. This gives a sequence of actions to be performed during the walkthrough and an understanding of the goal the user wishes to achieve; of course these goals are informed by the overall activity aims and the details in the user story. These can then be used to drive the standard action-by-action, screen-to-screen walkthrough prompts in Figure 5.

Typically the story may include similar phases of direct interaction at different points and, with care in case the context is different, a single walkthrough can be performed for these. Also a single phase of direct interaction may include several parts that each have a standard walkthrough: for example, in a banking application the user may view their statement and then go on to make a payment. If an aspect has already had a walkthrough, it does not need to be repeated for every phase that includes it; again taking care in case this means that the user has different knowledge, or is starting at a different part of the system,.

5.3 Start prompts

There is always a first use of a system. Sometimes this involves special direct interaction (registration for an online course), sometimes not (first visit to a football team website). However, before that first phase of direct interaction there has to be something that brings the user to the system. The start prompts in Figure 10 invite the analyst to consider the circumstances that lead to this first engagement with the system,

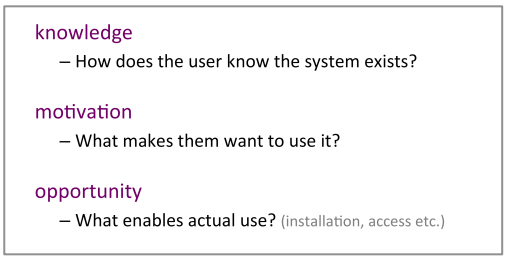

Figure 10: Start Prompts

For some systems some of these will be obvious: for example the new checkout assistant has the till in front of them. However, for others, especially purely online applications, even knowing that the system exists may be a problem. Does the user hear from friends, receive a company email telling them about the system, or seek it out?

If they know it exists, why should they choose to start to engage? Perhaps they need to be part of a group activity, or do so because of some pre-existing need. Typically the answer to this prompt will relate closely to the overall system motivation question in Figure 9.

Finally, and often far from trivial, is getting to use the system: for example I know the university has a travel booking system, but how do I navigate to it from the university intranet? For web-based systems this is often about knowing the relevant URL or the way to navigate to it. This can often be eased by direct links in an email (with security caveats of course!). For desktop systems there may be installation issues. Many years ago IBM user researchers used to show a video of engineers behind a one-way mirror laughing at users attempting to install a new PC … until they went through and tried themselves, even physically getting it out of its box was a challenge.

5.4 Between phase prompts

The between phase prompts in Figure 11 are primarily focused on re-engagement. This is informed especially by trigger analysis [10] recognising that breakdowns in long-term interactions are often due to a user failing to initiate the next stage in an extended process. They are also informed by the action analysis in [14], which identified the confluence of *drivers* (why you want to do something) and *capability* (the time and resources to do it) as essential for the successful execution of an action.

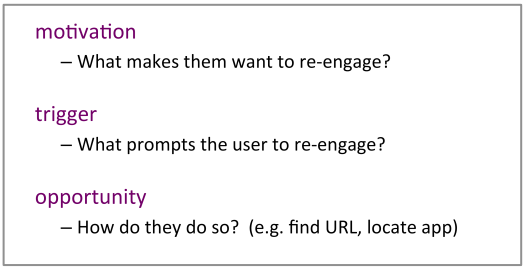

Figure 11: Between Phase Prompts

The prompts are largely self-explanatory, however, the first, *motivation* does need some expansion. Jean-Jacques Rousseau, the 19th century French philosopher, identified two kinds of will [48]: *actual will*, your momentary desire to do something, and *real will*, your longer term desires about the kind of world you wish to live in (sometimes called particular and general will). These were formulated in the context of the rights of the state to enforce general law, but the two time scales are useful when considering motivation in user interaction.

The answer to the *emotion and motivation* prompt Figure 10 is about long-term orientation of the user, for example to manage their money or improve their education, whilst the *motivation* in Figure 11 is about the precise moment when the user re-engages with the system. These are not necessarily in alignment: consider the desire to eat healthily vs. the temptation of another slice of cake! In such cases additional means may be needed to align high and low level motivation; for example the OpenBadges framework [OB19, GO15] was originally developed at P2PU (Peer to Peer University [PP19]) where students are highly self-motivated, but even so found they needed means to augment the long-term desire to learn with short term rewards, the basis of much gamification.

In some cases the motivation for re-engagement is a particular need, such as requiring more money from the bank. At other times it is more diffuse such as knowing you need to another learning session *sometime*.

In a simple system these prompts can be addressed once and apply equally to re-engagement after every gap. Even then, the answer to some of these prompts may include multiple mechanisms, for example motivation may differ between users, or there may be several potential triggers (remembering, email notification).

There also may be different kinds of phases of interaction (e.g. ATM vs. banking app.) or different kinds of gaps (working on something else vs. being on holiday) with very different contexts for re-engagement. In such cases several copies of Figure 11 would be completed for each kind of re-engagement.

5.5 End prompts

Finally we come to the prompts for when use of the system is coming to an end (Fig. 12). The first of these, *existence*, asks whether there is such an end. For some the end is merely disuse (for example Yahoo! Directory), but for others there is a potential defined end point such as completing an online courses. The aim of the system designers may sometimes be to avoid the end point occurring, but it still may need to be allowed if the user requires it (e.g. closing a bank account). Alternatively there may be an end point that is always expected (e.g. leaving a dating app when a partner has been found). In extremis the end point may be death, but often this still needs to be considered (e.g. back accounts, social media).

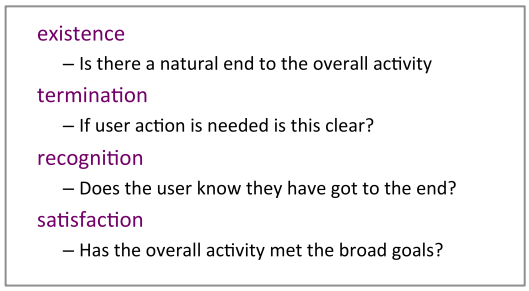

Figure 12: End Prompts

The remaining three prompts are only pertinent if there is an end point.

Termination and *recognition* prompts concern the mechanics of ending system use. In some cases this is driven by the system, for example of the user has completed all the modules in a course, but in others it is driven by a user decision, for example deregistering from an estate agent after having found a suitable house. In some cases, especially where the user is making the decisions, some explicit action is required by the user and hence the user needs to realise this is necessary and how to do this: for example, claiming a course-completion certificate, or locating the right part of the mobile-phone company site to initiate a move to a new provider. In cases where the end is intrinsic to the system, perhaps after the last credit payment for a new car, or when a student reaches the end of their course, it may be necessary to explicitly inform the user that their use of the system is at an end.

The last prompt *satisfaction* relates back to in the overall motivation for using the system in Figure 9. The analyst considers the extent to which the user actually achieves these aims or desires over their complete use of the system. Did they learn from an online course or enjoy playing on a gaming platform?

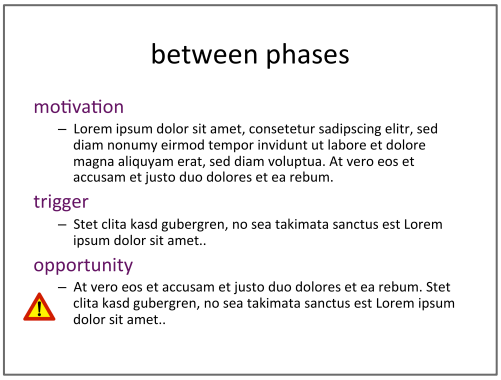

Figure 13: Anonymised completed prompt form (Lorem Ipsum text. courtesy [53])

6 LESSONS FROM USE

6.1 Applying the extended walkthrough

As noted the extended walkthrough was developed in response to the need to evaluate a system as part of a usability consultancy for a major industrial client. The specific application was an online course for company employees built on one of the major MOOC platforms, which had already undergone extensive user testing and in-use development.

A PowerPoint copy of the prompts was used for the structured evaluation duplicating prompt slides where this was required for different phases and actions. Answers were edited into the copied slides supplemented by screen shots and annotated details.

Some responses were simply comments, but were there was some form of problem or potential problem this was marked with a traffic-warning triangle (see Figure 13). This allowed rapid review of problems. Given the underlying platform already had many years of prior use and was so well established, one might have expected to only have content-specific issues, but in fact around 2/3 of the potential problems identified related to the platform, with on average one problem for each set of prompts (the slide in Figure 13 although anonymised was typical).

The fact that so many issues were identified on such a well-established platform is evidence not just that the new walkthrough prompts were useful, but of the need in general for aids for long-term interaction design support.

6.2 Cross-method insights

At the point the evaluation was performed the course had been delivered once and the structured walkthrough evaluation was performed alongside other forms of user and expert evaluations. The full details of these cannot be described for confidentiality reasons, but there were some general patterns that emerged.

User interviews revealed more content-specific comments, both positive and negative, compared to the walkthrough, which revealed more platform related issues. This is perhaps unsurprising as the aim of the walkthrough is early evaluation of the interaction and this was equally the case for the new and standard walkthrough prompts. However, there were also cross-over effects, e.g. motivation and saliency are often content related.

As well as the structured walkthrough, the expert performed an informal unstructured evaluation by following parts of the course. The expert was particularly familiar with learning systems, so it would have been ideal to compare the expert's unstructured evaluation, with another person using the structured walkthrough. However, the focus was the practical evaluation of the system, not assessment of the method, so the same person performed both and typically only mentioned issues in the informal evaluation that had been missed by the structured walkthrough; it is therefore not possible to assess how many of the issues highlighted by the walkthrough would have been spotted during informal evaluation.

The majority of additional issues identified during the informal evaluation concern detailed interactions; that is issues connected with the standard walkthrough prompts. This was largely because the evaluator was working in an exploratory manner, trying out different links, rather than following a prescribed 'standard' path through the software. This was possible because this was a deployed system, rather than simply a sketch or prototype.

This pattern was also evident in the (fewer) additional issues related to larger scale interaction. The prompts had clearly picked

up the majority of the issues, but the additional ones identified were ones that were only evident in actual use, for example, notification emails being classified by Gmail into non-important categories and receiving multiple confusing email notifications due to starting the course 'late'.

So in both short and long term interactions, we see the limits of structured evaluations, which by their nature tend to focus on a few paths out of the many options available in any relatively complex system. This will also be familiar to those involved in other forms of software testing where unit tests and test suites are essential, but typically need to be supplemented with free use.

Two more issues arose in the informal evaluation, which suggest possible additions to structured prompts.

The first concerned the point of *disengagement*, when the user finishes one of the phases of direct interaction. Some software simply continues where you left off, but some requires a 'save' action or other indication that a session has finished. There are many well-known premature closure problems where such completion actions are missed as the user gets the "I've done it" feeling before the. last system action; perhaps the most well known of these was in early ATMs which dispensed cash before the card was returned leading to many forgotten cards.

The second issue concerned a sense of *on-going progress*. This is perhaps a particular issue for learning systems, but in any system it is important that the users on-gong sense of satisfaction and emotional engagement is managed throughout use, not just at the and point as in the current prompts.

Finally, in both user interviews and informal expert evaluation there were comments or questions related to *offline activ*ity. The between phase prompts in Figure 11, are primarily about re-engagement, because this was known to be a common issue in many information systems. However, there is clearly also need to be prompts related to the gap itself.

REFLECTION AND FURTHER WORK

This paper has presented a novel set of prompts designed to extend popular cognitive walkthrough techniques to encompass issues arising from long-term interaction with multiple phases of direct interaction. Evaluation prompts are a generative artefact in the sense of [13]: a system or object that is not used for itself, but instead used to design, create, or in this case formatively evaluate, other things. The empirical evaluation of evaluation tools is therefore methodologically problematic as the success depends not just on the tool but the application and the analyst. Empirical evaluations of such tools involves a class of students applying several tools to different systems; methodologically problematic both because of the limited expertise of the student analysts and the danger of Hawthorne effects given the new tool being evaluated has typically been produced by their tutor.

However, it is possible to validate generative artefacts using a combination of justification (reasoned arguments for efficacy) and empirical use [13]. This paper has taken this approach with prompts that are derived from the theoretical literature and experience of practical use. The latter did provide evidence for the efficacy of the prompts. First because the use of the prompts uncovered a substantial number of issues in a system that had already been subject to heavy evaluation. Second because the informal expert evaluation and user interviews based on real use uncovered few issues of long-term interaction that were not exposed by the walkthrough and of the issues not captured, most related either to content or use of the deployed systems, nether of which would be expected to be uncovered by an early-evaluation method. Measured by this, the prompts can be seen as a success.

However, the practical use did highlight areas that were not well covered by the prompts presented here, suggesting future improvements.

- Currently the satisfaction of the overall motivation for use of the system is only questioned at the final end of use, which in some systems may be never. Some form of prompt relating to on-going sense of progress is also needed.
- The prompts cover re-engagement after gaps in use, the issues at the point of dis-engagement are not currently covered. A new prompt card is needed for this probing the need for explicit actions, potential for premature closure, and things that the user may be expected to remember after use.
- The prompts in Figure 11 were conceived as 'between phase', in fact they are really about re-engagement and a new card for the gaps themselves is needed probing issues such as off-line activity and potential for interfering activities.

The prompts were developed based on a real need. This is positive in establishing actual utility, but undoubtedly influenced the kind of prompts developed. For example, one of the issues highlighted in the literature is that long-term interactions often cross organisational boundaries, giving rise to potential breakdowns, for example if you are waiting for a response to a message that never comes [10]. On reflection it is clear that the reason this has not been included in the current prompts is because the context of use was within a single organisation. While this is a common situation, to be more useful in the widest possible set of circumstances some prompts concerning this need to be added to gap or re-engagement cards.

Of course adding more cards and prompts may increase the complexity of the analysis and hence reduce the likelihood of practical use. This suggests two potential areas for future work. First would be to create slight variants for different kinds of situation, perhaps selecting a particular set of prompt cards depending on an initial assessment of the system and context. Second would be to embed the prompts in an online tool managing multiple paths through the prompts, and allowing support material, such as examples or more detailed descriptions, to be available on a just-in-time bases.

Finally, while this is a complete and used practical tool, it is also a first step. Just as traditional cognitive walkthroughs have been augmented, updated, modified and evaluated by multiple authors and practitioners; feedback and further work by others is both welcome and necessary.

REFERENCES

[1] D. Benyon, 2010. Designing Interactive Systems. Addison Wesley

[2] T. Bickmore and R. Picard, 2005. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact*. 12(2):293–327. DOI: 10.1145/1067860.1067867

[3] M. Blackmon, P. Polson, K. Muneo and C. Lewis, 2002. Cognitive Walkthrough for the Web. In *Proc. CHI 2002*. 463–470. DOI: 10.1145/503376.503459

[4] P. Campos and N. Nunes, 2007. Practitioner Tools and Workstyles for User-Interface Design. *IEEE Software*, 24(1):73-80, Jan.-Feb. 2007. DOI: 10.1109/MS.2007.24

[5] S. Card, T. Moran and A. Newell, 1980. The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, 23, 396-410.

[6] A. Chamberlain and A. Crabtree (eds.), 2019. *In Into the Wild: Beyond the Design Research Lab*. Springer, pp.7–29.

[7] A. Cooper, 1999, The Inmates Are Running the Asylum. Sams, 1999

[8] D. Diaper & N. Stanton (eds.), 2004. *The Handbook of Task Analysis for Human-Computer Interaction*. Lawrence Erlbaum Associates.

[9] A. Dix and S. A. Brewster, 1994. Causing Trouble with Buttons. *Ancilliary Proceedings of HCI'94,* Glasgow, Scotland. Ed. D. England

[10] A. Dix, D. Ramduny and J. Wilkinson, 1998. Interaction in the Large. *Interacting with Computers* - Special Issue on Temporal Aspects of Usability. J. Fabre and S. Howard (eds). 11(1):9-32.

[11] A. Dix, J. Finlay, G. Abowd, and R. Beale, 2004. *Human–Computer Interaction* (3rd ed.). Pearson.

[12] A. Dix, D. Ramduny-Ellis and J. Wilkinson, 2004. Trigger Analysis - understanding broken tasks. Chapter 19 in *The Handbook of Task Analysis for Human-Computer Interaction*. D. Diaper and N. Stanton (eds.). Lawrence Erlbaum Associates, pp. 381-400

[13] A. Dix (2008). Theoretical analysis and theory creation, Chapter 9 in Research Methods for Human-Computer Interaction, P. Cairns and A. Cox (eds). Cambridge University Press, pp.175–195. ISBN-13: 9780521690317

[14] A. Dix and J.Leavesley (2015). Learning Analytics for the Academic: An Action Perspective. *Journal of Universal Computer Science* (JUCS), 21(1):48-65.

[15] S. Easterbrook (ed.). 1993. *CSCW: Cooperation or Conflict?* Springer.

[16] P. Fitts, 1954. The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology, 47, 381-391.

[17] S. Gibbons, 2017. Service Design 101. Nielsen Norman Group. https://www.nngroup.com/articles/service-design-101/

[18] D. Gibson, N. Ostashewski, K. Flintoff, S. Grant and E. Knight, 2015. Digital badges in education. *Educ Inf Technol* 20:403–410. DOI: 10.1007/s10639-013-9291-7

[19] Jonathan Grudin. 1988. Why CSCW applications fail: problems in the design and evaluationof organizational interfaces. In Proc. CSCW '88. ACM, pp.85–93. DOI: 10.1145/62266.62273

[20] Y. Guiard and M. Beaudouin-Lafon (eds.), 2004. Fitts' law fifty years later: Application and contributions from human-computer interaction. A special issue of the *International Journal of Human-Computer Studies*, 61 (6).

[21] C. Heath and P. Luff. 1991. Collaborative activity and technological design: task coordination in London underground control rooms. In *Proceedings of ECSCW'91*. Kluwer Academic Publishers, USA, 65–80.

[22] S. Henry, 2007. Accessibility in User-Centered Design: Example Scenarios. *Just Ask: Integrating Accessibility Throughout Design*. Lulu.com. http://www.uiaccess.com/accessucd/scenarios_eg.html

[23] Interaction Design Foundation, 2019. The Principles of Service Design Thinking - Building Better Services. (accessed 29/1/2020). https://www.interaction-design.org/literature/article/the-principles-of-service-design-thinking-building-better-services

[24] N. Iivari, Marianne Kinnula, Leena Kuure, and Tonja Molin-Juustila. 2014. Video Diary as a Means for Data Gathering with Children - Encountering Identities in the Making. *International Journal of Human-Computer Studies* 72, 5: 507--521

[25] V. Kaptelinin and B. Nardi, 2012. *Activity Theory in HCI: Fundamentals and Reflections*. Morgan and Claypool

[26] H. Khalid and A. Dix, 2010. The experience of photologging: global mechanisms and local interactions. *Pers Ubiquit Comp*ut 14:209–226. DOI: 10.1007/s00779-009-0261-4

[27] A. Kidd, 1994. The marks are on the knowledge worker. In *Proceedings of CHI '94*. ACM, 186–191. DOI:https://doi.org/10.1145/191666.191740

[28] S. Kujala, V. Roto, K. Väänänen-Vainio-Mattila, and A. Sinnelä, 2011. Identifying hedonic factors in long-term user experience. In *Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces (DPPI '11)*. ACM, Article 17, pp. 1–8. DOI: 10.1145/2347504.2347523

[29] C. Lallemand, 2012. Dear Diary: Using Diaries to Study User Experience. *User Experience magazine*, August 2012. User Experience Professionals Association (UXPA) https://uxpamagazine.org/dear-diary-using-diaries-to-study-user-experience/

[30] I. Leite, C. Martinho and A. Paiva , 2013. Social Robots for Long-Term Interaction: A Survey. *International Journal of Social Robotics*, 5:291–308

[31] C. Lewis, P. Polson, C. Wharton and J. Rieman, 1990. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proc. CHI'90*. ACM, 235–242. DOI:10.1145/97243.97279

[32] R. Kohavi and S. Thomke, 2017. The Surprising Power of Online Experiments. *Harvard Business Review*, September 2017, pp.74–82.

[33] I. S. MacKenzie, 2003. Motor behaviour models for human-computer interaction. In J. M. Carroll (ed.) *HCI models, theories, and frameworks: Toward a multidisciplinary science*, pp. 27-54. San Francisco: Morgan Kaufmann.

[34] Y. Malhotra. 1998. Business Process Redesign: An Overview. *IEEE Engineering Management Review*, 26(3), Fall 1998.

[35] J. McCarthy and P. Wright, 2007. *Technology as Experience*. MIT Press.

[36] L. Myers, |1987. *Proposed Military Standard for Task Analysis*. Technical Memorandum 13-87. U.S. Army Human Engineering Laboratory, Maryland US.

[37] J. Nielsen and R. Mack (eds), 1994. *Usability Inspection Methods*, John Wiley & Sons Inc

[38] L. Nielsen, 2013, Personas. In: *The Encyclopedia of Human-Computer Interaction, 2nd Ed.* M. Soegaard and R. Dam,(eds.). The Interaction Design Foundation. http://www.interaction-design.org/encyclopedia/personas.html

[39] D. Norman, 1988. *Psychology of Everyday Things* (later The Design of Everyday Things). New York: Basic Book, 1988.

[40] *OpenBadges*. (accessed 27/1/2019). https://openbadges.org/

[41] T. Palmer, 2019. *The 2019 Design Tools Survey*. https://uxtools.co/survey-2019

[42] F. Paternò, C. Mancini an S. Meniconi 1997. ConcurTaskTrees: A Diagrammatic Notation for Specifying Task Models. In: Howard S., Hammond J., Lindgaard G. (eds) *Human-Computer Interaction INTERACT '97*, Springer.

[43] F. Paternò, 2000. *Model-Based Design and Evaluation of Interactive Applications*. Springer.

[44] *Peer 2 Peer University*. (accessed 27/1/2019). https://www.p2pu.org/

[45] P. Polson, C. Lewis, J. Rieman and C. Wharton, 1992. Cognitive walkthroughs: A method for theory-based evaluation of user interfaces. *International Journal of Man–Machine Studies*, 36:741–73,.

[46] D. Ramduny-Ellis, A. Dix, P. Rayson, V. Onditi, I. Sommerville and J. Ransom, 2005. Artefacts as designed, artefacts as used: resources for uncovering activity dynamics. *Cogn Tech Work* 7:76–87 (2005). https://doi.org/10.1007/s10111-005-0179-1

[47] Y. Rogers and P. Marshall, 2017. *Research in the Wild.. Synthesis Lectures on Human-Centered Informatics*. Morgan and Claypool. Doi:10.2200/S00764ED1V01Y201703HCI037

[48] J-J Rousseau, 1762. *The Social Contract*.

[49] A. Shepherd (1998) HTA as a framework for task analysis, *Ergonomics*, 41:11, 1537-1552, DOI: 10.1080/001401398186063

[50] R. Spencer, 2000. The streamlined cognitive walkthrough method, working around social constraints encountered in a software development company. In *Proc. CHI'00*. ACM, 353–359. DOI: 10.1145/332040.332456

[51] R. Thomas, 1998. *Long Term Human-Computer Interaction - an exploratory perspective*. Springer.

[52] Usability.gov, 2019. *Heuristic Evaluations and Expert Reviews*. (accessed, 26/1/2019). https://www.usability.gov/how-to-and-tools/methods/heuristic-evaluation.html

[53] Websitetips.com, 2019. *Lorem Ipsum... Who?* (accessed 27/1/2019). http://websitetips.com/articles/copy/lorem/

[54] C. Wharton, J. Rieman, C. Lewis and P. Polson, 1994. The cognitive walkthrough: a practitioner's guide. In *Usability Inspection Methods*. John Wiley, New York.