Feedback and Task Analysis for E-Commerce Sites

Paula Kotzé, Karen Renaud, and Tobias van Dyk
University of South Africa,University of Glasgow

Abstract: Developers of e-commerce applications are often sceptical about Web-site usability guidelines. User testing is also usually not carried out because it is expensive in terms of time and expertise. The spectacular usability and commercial failure of some sites attest to the folly of such practices. The main reason for developers neglecting current evaluation practices is that they are often vague, and in the case of user testing, too difficult to do effectively. This paper therefore offers an alternative. The e-commerce shopping process has been analysed from a task-based point of view, and a set of task-weighted metrics to be used by developers in evaluating their sites has been proposed. These metrics have been applied to four sites and the results of the evaluation are given.

Key words: Feedback, Task analysis, Usability, E-commerce, Evaluation, Task-weighted metrics.

1. INTRODUCTION

Feedback during the user task can be used to assist in supporting understanding of the functionality and requirements of an E-commerce (EC) application and can be effectively harnessed to minimize website abandonment. In a previous paper [13] we chose to discuss one particular aspect of web-site usability, namely that of feedback. We now continue to advocate the extensive use of feedback to increase the user's task support of systems, thus enhancing the ease of use of these systems. To support an understanding of the EC shopping process the purchasing part of the shopping cycle was analysed and two distinct and dissimilar phases were identified. Phase-specific evaluation metrics were applied to these. These

metrics provided a first attempt at defining an evaluation mechanism, which can be used by developers to flag problem usability and feedback areas.

This paper extends the previous work by proposing a combination of a task-based and metric-based approach to enhancing Web usability through effective feedback. This is a viable yet novel technique for providing inexperienced developers with a tool that can be used to improve the quality of their sites. Section 2 will briefly reiterate previous findings on the nature of feedback in EC applications [13]. Before a methodology for evaluating proposed EC web site can be provided it is necessary to understand the nature of the EC shopping experience, and this will be discussed in Section 3. Sections 4 and 5 propose a task-weighted evaluation methodology for EC systems that incorporates a number of essential feedback loops. Section 6 is an explanation of the method used and Section 7 discusses the results of the application of the metrics to on four large EC sites. Section 8 concludes.

2. FEEDBACK

The OED defines feedback as: *signifying a response*, *modifying the behaviour of the user* and *promoting understanding*. The traditional role of feedback in human-computer interaction is often seen exclusively as pertaining to the first use. The extension of the feedback concept to include all the above-mentioned features will enable EC sites to give better and more helpful feedback to users. Feedback serves a behavioural purpose in the interaction between users and computers, with the computer fulfilling the same conversational role as a conversational participant [11]. Only by means of feedback can participants in a conversation detect faults in the understanding of what is said [7]. The success of the human-computer 'conversation' will depend on the user being able to gauge the 'knowledge' of the application. Feedback must make the 'knowledge' of the application, based on previous inputs, tangible and accessible in order to fulfil its role adequately in the face of an untutored and unknown user population.

Most users of EC systems will not have been trained in their use. The user interface will therefore have to be designed with great care so that the user can discover everything that is task-supportive from the system, based on the feedback showing the perceptible system state. The designer of the user interface must be sure to bestow rational behaviour on the application – ensuring that the application behaves in a way that is reasonable and intelligible. By concentrating on the EC task the developer can move closer to a system that the user can use intuitively.

The conversational model of user interaction, with respect to the current computer usage paradigm of recognition rather than recall [4], leads us to consider users as reacting according to the way they interpret the state of the

system. The quality of the feedback provided by the system can assist in enabling an understanding of the state of the system and becomes very important when the system is prone to long response times, a common occurrence in EC systems. It is necessary to consider the purpose of any feedback, and the way a user can be expected to make use of such feedback as is provided.

3. THE E-COMMERCE PURCHASING TASK

Singh, Jain, and Singh [14] break up the EC process into three activities: identifying and finding a vendor, purchasing and tracking. We will examine only one of their processes – namely the *purchase* task, which can divided into two distinct phases, as shown in Figure 1.

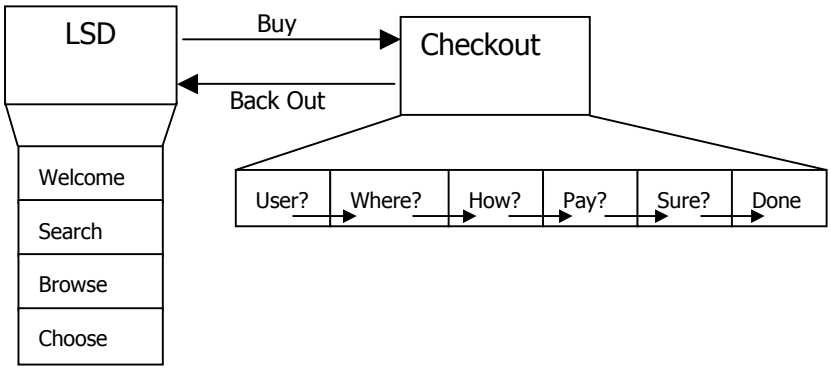

Figure 1. The two phases and ten stages of the purchase task [13].

(A) *Look, See and Decide (LSD):* This stage will typically be used to look at available products, compare them, and to make a decision about whether or not to purchase products. This may be done one or more times until the consumer has found products that satisfy his or her needs. This phase is intensely user-driven because the user is looking at and assimilating information continuously. It is composed mostly of searching and browsing, discussed extensively by Belew [2]. It has the following substages, which can be traversed iteratively and in varying sequences:

Welcome; Search; Browse; Choose.

(B) *Checkout:* When users trigger this stage they have made their choice of offered products and have decided to make a purchase. They now have

to provide certain details, such as their address and credit card details. This stage is system-driven and changes the paradigm of the interaction process from user initiative to system initiative. Feedback is of critical importance during this stage. Users who feel that they have lost control can simply leave the site without any embarrassment, unlike a user who is standing at a checkout till in a supermarket. This stage is typically composed of at least the following steps, which should be navigated in a logically sequential fashion:

User? → Where? → How? → Payment? → Sure? → Done

Some Websites will have all these stages integrated into one page (e.g. www.amazon.co.uk) but the implied functionality is the same – each of these categories of information must be provided so that the transaction can be carried out. Brinck, Gergle and Wood [3] combine UML use case analysis with hierarchical task analysis into a powerful technique. They identify two use cases for a book purchasing scenario namely "Buy Book" and "Complete Order", which coincides with the LSD and Checkout phases as identified above.

The following section will report on the results of a task analysis carried out on the EC purchasing task, which reveals some important differences and insights. This task analysis facilitated the setting up of the required evaluation metrics.

4. E-COMMERCE TASK ANALYSIS

Task analysis is a valuable technique for refining and improving a user interface. A simple computer-operating model may serve as an effective basis for an understanding of the feedback-guided and goal-directed nature of an EC task execution. This model (Figure 2) can also serve to further highlight the tasking difference between the two phases of the EC shopping process. A definition for task analysis that is suitable within the context of this application domain is that offered by Dix *et al.* [6] – they describe task analysis as the *identification and description of the interactive system user's problem space, in terms of domain, goals, intentions, and tasks.*

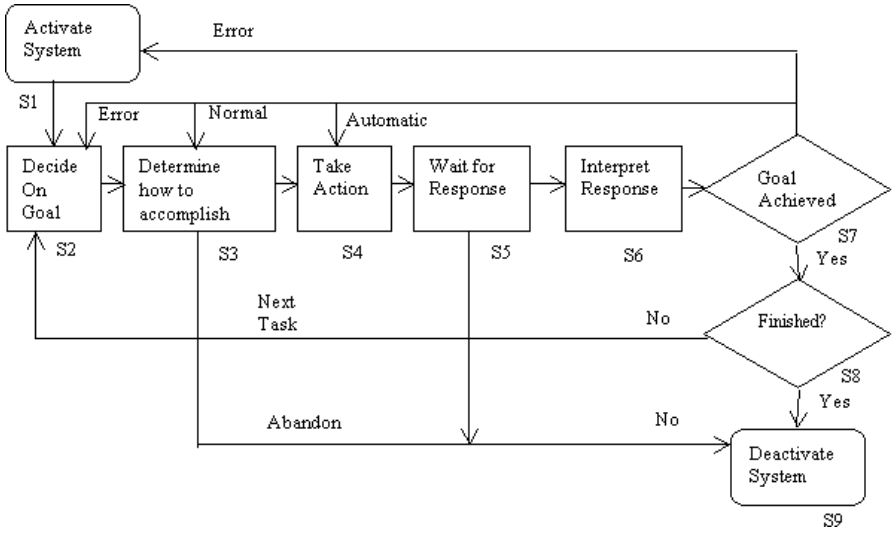

Figure 2: A Feedback scheme for Task Analysis based on a Simple Iterative Computer Operating Model (adapted from [16])

The nature of the shopping task differs significantly during the LSD (Look-See-Decide) and Checkout phases. The LSD phase is, in essence, a *user-driven* iterative browsing and selection task with (possibly) less well-defined goals and a larger number of possible actions and feedback options. The Checkout phase is a *system-driven* pre-defined, linear task with well-defined goals and sub-goals, and with a smaller number of predefined actions and well-defined sets of feedback loops. Most Web-design guidelines do not take these diametrically opposing operating paradigms into account, even though the principle of dialog initiative and system versus user pre-emption is well established [6].

During the LSD phase there will be three types of goals:

- browsing (searching), for the shopping object,
- categorizing (searching-result), the shopping objects, and
- specifying (selecting) a shopping object for the shopping cart.

The nature of the interaction is such that the customer should be kept interested in the results of the search-type goals, thus retaining them on-site – the focus is on discouraging user dropouts through abandonment of the goal or by linking them off-site [5]. System errors and poor response times during this phase are perceived to be less serious by the user (but not by the Website owner), since they may result in shopper abandonment.

The checkout phase has a single goal – completing the financial transaction (as defined by the contents of their shopping cart and their shipping preferences) as quickly and securely as possible with the minimum

of disruption and roadblocks [W1]. Accordingly this phase has a set of linear and intentionally fairly rigid sub-goals. The focus here is not on user entertainment but on completing the transaction rapidly and securely – before shoppers change their minds about their shopping carts and the related cost [W1]. This implies that response times, and clear feedback on reasons for delays are more important here than during the LSD phase.

Because errors could have a more serious (security and financial), impact, a well-designed user help function and clear explanatory sub-system are required. It is also advisable to provide the user with obvious and intuitive navigation clues as to where they are in the process by using progress or stage indicators. The trend should be to strive for the minimum number of pages or stages – rather have the user scrolling moderately than clicking through to a larger number of small pages [W1]. This is in stark contrast to having as much as possible of the relevant information immediately visible in the LSD stage [15]. Simplifying this process will ensure that there will be a smaller incidence of user dropout and shopping cart abandonment during this phase – provided additional costs such as shipping are shown as soon as possible. The effect of an error at this stage will affect the sequence of the sub-goals, and will make this phase non-linear (i.e. 'loopy').

When these aspects are applied to the model as presented in Figure 2 the following should be noted for the two phases:

1. More effort will be required for system activation (S1) in the LSD phase when compared to the checkout phase – for example, the customer has to have an established Internet session. The customer also needs to know about the site. This part is well understood by marketing professionals and sites are often well advertised in the media. Unfortunately this level of attention is often not paid to other aspects of the EC experience.
2. Goal formulation (S2) may be less clear in the LSD phase as compared to the checkout phase – the customer may want to re-evaluate options and re-formulate goals based on the range, price, and availability of the shopping objects during the outcome of a set of search results.
3. The intermediate stages (S2 – S6) are less proscribed for the LSD phase, and there will be a natural tendency to loop back to S2 during this phase – for example, if the response (at S5) takes too long.
4. The S3 stage is often trickier for the user to formulate in the LSD phase. The user may have some vague idea of an item he or she needs, but may have difficulty formulating a query. For example, the user may have heard about a popular autobiography by an Irish teacher who grew up in Limerick. The user types in many different search criteria – "Irish", "teacher", "Limerick" before perhaps finding the book *Angela's Ashes*

by Frank McCourt by browsing through the list of available autobiographies.

5. Interpretation of the response (S6) will be more difficult during the LSD phase when compared to the checkout phase – the customer may be presented with a range of shopping objects from which to choose compared to the linear progression during the checkout phase.
6. Measuring the success of the task at S5 will be more difficult for the LSD phase – the customer is dealing with a goal achievement based on an electronic description rather than confirmation of a familiar financial transaction as in the checkout phase.
7. The result of an error (which may be at S7 → S1, or S4 → S2) will be deemed to be less serious during the LSD phase than during the completion of a transaction in the checkout phase.
8. The transition S7 → S2 may be traversed during the LSD phase without an error having been made – it could happen as a result of a reformulated goal.
9. The transition S7 → S8 could be made as a result of abandonment even though the goal has not been achieved – since the site may not stock the required object.
10. The activity distance S2 → S7 should be as short as possible for the checkout phase with achievement of S7 always clearly visible, perhaps by means of stage indicators.
11. The S3 → S9 and S5 → S9 feedback loops should be minimised by ensuring good site usability.

It is necessary to translate the discussion of this section into some set of recommendations so that developers have guidelines to follow in order to ensure that adequate feedback is provided. Veen [15] refers to the difficulty of evaluating websites. Developers using traditional usability-testing methods are often faced with an iterative and time-consuming evaluation process involving a number of users – until every perceivable problem is solved. The provision of a set of easily applied metrics should make it easier for EC site developers to profit from accumulated research results and make the evaluation process a little less daunting.

5. TASK-WEIGHTED EVALUATION METRICS

The previous section discussed the differences between the two different phases of the E-Commerce task. It is fitting for the two phases to have different evaluation metrics as well – as befitting their different paradigms and needs – and for these metrics to be weighted according to their impact on usability.

We previously reported on a set of suitable evaluation metrics for measuring feedback-related usability of three E-Commerce websites [13]. The evaluation criteria used were equally weighted. In certain cases, however, it may be advantageous to prioritize some of the criteria by means of a selective, unequal, weighting. Examples of how to do this may be found in Levi and Conrad [9]. They describe the application of Nielsen and Mack's [10] usability guidelines to the evaluation of a set of Web pages. After the evaluation they modify the list based on feedback from their two different (HCI and Web developers) evaluation teams, and produce a new list by assigning severity ratings to each usability violation found on a five-point scale. In addition they also prioritize on the basis of the frequency of occurrence of the usability problem. Their scale varies from 0=Not a usability problem; 1=Cosmetic; 2=Minor; 3=Major; to 4=Catastrophic problem. They produce a list of usability violations, which contains both frequency and severity information. Nielsen [W4] produced a guideline for severity ratings based on the same scale, but then expanded on this by noting that the severity of a usability problem was a combination of four factors: The frequency with which the problem occurs, the impact of the problem on users if it occurs, the persistence of the problem, and the market (product popularity) impact of the problem.

Along the same lines, Bastien and Scapin [1] refer to the *amount and importance of usability problems found*. Another technique applies a *strength of evidence* scale to a set of evaluation criteria [W2]. These criteria are based on the type and number of research experiments that may support, or discount, the specific criterion. W3.org [W3] prioritizes in terms of (accessibility) guidelines that *must* be applied, *should* be applied, or *may* be applied.

We propose a novel approach, which assigns a *task weighting* to each of these previously equally-weighted criteria scores. This task weighting has two components:

1. A *task repetition* component (R). The task repetition component is an indicator of how often this task or activity will be encountered during the interaction. A weight of 0.1 will indicate it to be of low occurrence – implying that it only happens in exceptional cases, whereas a factor of 0.2 means that it happens very seldom and a weight factor of 1.0 will indicate that this type of activity occurs regularly during the interaction. A value of 0 indicates absence of the activity – indicating that it should not play a role in the calculation of the overall score.
2. A *task complexity* component (C). The task complexity component reflects the inherent degree of difficulty in executing the task or activity. A weight factor of 1.0 indicates that the activity is highly complex, requiring extensive background and operational knowledge, or requiring

a high degree of complex interaction. A weight factor of 0.1 indicates that the task is simple, with low interactivity. A weight value of zero is not possible because it indicates that no interaction is required.

Each of these two components either amplifies or attenuates the contribution of a specific usability feedback criterion to the overall score. The overall value should be a more faithful reflection of the website's overall usability than that rendered by non-weighted metrics.

In support of our view it should be noted that Brinck *et al*. [3] distinguish between the frequency and priority of tasks and that they (correctly) note that the starting point for their HTA (Hierarchical Task Analysis), should be important tasks that occur frequently. Similarly Nielsen's [W4] 'impact and persistence of the problem' could make the *problem context* a high-priority task. Lastly, Bohmann [W5] has developed and tested task metrics for a quantitative usability evaluation. These metrics make it possible to calculate the usability effect of redesign efforts. The two main metrics are: (1) Task Time – time to complete a task or set of tasks and (2) Task Errors – number of errors per task.

The next section will evaluate four E-Commerce book sellers' websites using this technique.

6. METHOD

It is difficult for developers, given a list of guidelines, to know which to follow. For example, developers are told to have the most important information visible to the user without scrolling. They are also told to provide the user with enough information to keep them interested in the site thus increasing the chance of a sale. Which of these is more important? We therefore propose the use of a set of *metrics* that can be used by developers to evaluate each page of a Website. Additionally we propose an approach of metric weighting that involves the use of the *task complexity* and *task repetition* components as discussed previously.

The following section will describe how these metrics were applied to a selection of four E-Commerce sites, and comment about the efficacy of the proposed evaluation mechanism. In order to evaluate E-Commerce Web pages, a raw score is given for each of the questions (metrics) as follows (Ravden and Johnson, [12]):

- Never (0) - the feature is never available.
- Sometimes (1) - the feature is seldom there.
- Mostly (2) - the feature is usually there.
- Always (3) - the feature is universally available.

The first step in the evaluation is the raw scoring of the usability metrics. The scores are determined per E-Commerce site, per phase (LSD and Checkout), per stage within the phase, and also per page, as a ratio to the maximum score. The scores for each metric in each stage are calculated by adding up the score for each page making up the stage and awarding a total for each particular metric feature. The scores for each feature are then totalled to arrive at a percentage per site per purchasing stage to arrive at a raw score. It is important to note that the evaluator should not feel constrained by the list of metrics given here – these were adapted and selected from a much larger list (11 sections and 179 metrics) developed and extensively tested by Ravden *et al*. [12]. It is likely that differences in E-Commerce sites may require the evaluator to re-visit this more comprehensive list and add to (or subtract from) the list of metrics given here. The activity checklist as presented by Kaptelinin, Nardi and Macaulay [8], which is based on a broadened view of task context (Activity Theory), may also yield additional evaluation metrics. It is strongly recommended that more than one evaluation (and evaluator) be used to arrive at the raw metric scores – three data sets can be considered to be the absolute minimum. The individual scores from the data sets should be averaged as an input to the second step.

The second step assigns values for the two task weight components (R+C), based on the evaluators' experiences with the site during the metric scoring step. These values are designed be have little effect initially on the raw scores until the evaluator develops more confidence in applying the correction. The natural tendency will be to choose median values of close to 0.5 for both the task repetition and complexity values which will imply that the weighting adjustment will effectively be 1, (R+C) = (0.5+0.5). Thus initially the adjustment will be no worse than the unadjusted raw metric scores. Ultimately as experience is gained in the use of the weight factors, the weight adjustments could realistically have a large influence on the metric score – consider a low complexity, low repetition value of 0.5 compared to a high repetition high complexity value of 1.5. The metric adjusted by the first would only contribute one third as much to the overall usability score for the E-Commerce site when compared to the second metric.

A third step would then be to eliminate those metrics with particularly low (R+C) values (for example if (R+C)<0.6), from the evaluation – this would partly alleviate the problem of a tendency towards an average of 1 for all (R+C) values when a large number of metrics are used.

To arrive at the final usability coefficient (UC) for the metric we apply this formula:

$$UC = (Score/MaxScore) * ((R+C)/TotalR+C)$$

To complete tables 3 and 4 we need to calculate an overall usability score for each phase per site in order to facilitate a comparison between sites. The % usability score based on raw (non-task-weighted) scores is:

$$\text{Raw} = (\sum \text{Score})/(\sum \text{MaxScore}) * 100$$

The percentage usability, based on task-weighted scores, is calculated as:

$$\text{Task-weighted} = (\sum \text{UC}) * 100$$

The normalised ratios make it easier to compare usability scores and are expressed as ratios relative to the highest scoring site. For sites other than the top site the score is calculated as follows:

Task-weighted score / Top-Site's Task-weighted score

A suitable list of evaluation metrics similar to previously reported results (Renaud *et al.* [13]) is shown in Tables 1 and 2 for the LSD and Checkout phases respectively. Typical task weighting factors for each metric are as indicated. To use these two tables the original (raw) score for each metric is multiplied by the factor given in the table for the metric. Of particular interest would be criteria with associated task or activities that have either high combined (i.e. repetition and complexity (R+C)), weight factors, or very low combined weight factors. This could imply that these task components are proportionally either more, or less important to the usability evaluation.

LSD Phase: Metrics for the evaluation of User Task support		**Task Weighting Factors (Repetition+Complexity)**
S3	Is it clear how the user must search for a product?	(0.8+0.5)/8.5 = 15% (>1)
S3	Are different types of information clearly separated?	(0.8+0.9)/8.5 = 20% (>1)
S4	Is it clear what needs to be done to select a product?	(0.1+0.6)/8.5 = 8% (<1)
S5→S9	Does the system inform the user of reasons for delays?	(0.8+0.5)/8.5 = 15% (>1)
S7→S2	Does the search engine offer alternatives if the search fails?	(0.5+0.8)/8.5 = 15% (>1)
S7→S2	Can the user undo a product selection?	(0.1+0.5)/8.5 = 7% (<1)
S7→S3	Does the system allow the user to explicitly check on previous searches?	(0.5+0.5)/8.5 = 12% (1)
S8→S9	Is it clear how the transition to checkout can be made?	(0.1+0.5)/8.5 = 7% (<1)
Average Task Weighting Factor for this Phase:		8.5/8.0 = 1.06

Table 1: User task metrics for the LSD phase

Checkout Phase: Metrics for the evaluation of User Task support		**Task Weighting Factors (Repetition+Complexity)**
S2,S3	Are possible actions clear?	(0.8+0.7)/16 = 9% (>1)
S3	Are instructions and messages concise, clear and unambiguous?	(1+0.8)/16 = 11% (>1)
S3,S5→S9	Can the user easily back out of the process?	(0.1+0.8)/16 = 6% (<1)
S4	Is the required format of user actions clearly indicated?	(0.5+0.8)/16 = 8% (>1)
S5→S9	Does the system inform the user of the reasons for delays?	(1+0.5)/16 = 9% (>1)
S6	Are user actions linked to changes in the interface?	(0.8+0.5)/16 = 8% (>1)
S6	Is there always an appropriate response to user actions?	(0.8+0.8)/16 =10% (>1)
S6	Does the user explicitly confirm the final purchase?	(0.1+0.4)/16 = 3% (<1)
S6	Does the system indicate the current stage?	(0.3+0.5)/16 = 5% (<1)
S6	Can users check on inputs provided during the process?	(0.1+0.5)/16 = 4% (<1)
S7→S8 or S7→S2,S3,S4	Does the system inform the user of the success or failure of their actions?	(1+0.5)/16 = 9% (>1)
S7→S2	Do error messages indicate the what, where and why and how to recover?	(0.5+1)/16 = 9% (>1)
S7→S8	Is it clear what the user must do to complete the task?	(0.5+0.7)/16 = 8% (>1)
Average Task Weighting Factor for this Phase:		16/13 = 1.23

Table 2: User task metrics for the checkout phase

7. EVALUATION

As a representative illustration of the technique, Tables 3 and 4 list the results of four evaluations on three different book sellers on the Internet namely Amazon (2001 and 2002, http://www.amazon.com), Kalahari (2001, http://www.kalahari.net), and Books Online (2001, http://www.uk.bol.com).

Evaluation of User Task support: LSD Stage	**Amazon (New)**	**Amazon (Old)**	**Kalahari**	**BOL**
Is it clear how the user must search for a product?	6/9 * F1 = 0.102	6/9 * F11 = 0.108	6/9 * F11 = 0.108	7/9 * F11 = 0.126
Does the search engine offer alternatives if a search fails?	9/9 * F1 = 0.153	9/9 * F12 = 0.125	3/9 * F12 = 0.042	0/9 * F12 = 0.000
Does the system inform the user of the reasons for delays?	5/9 * F1 = 0.085	5/9 * F13 = 0.090	3/9 * F13 = 0.054	3/9 * F13 = 0.054
Are different types of information clearly separated?	8/9 * F1 = 0.178	7/9 * F14 = 0.155	6/9 * F14 = 0.133	9/9 * F14 = 0.200
Is it clear what needs to be done to select a product?	9/9 * F1 = 0.082	9/9 * F15 = 0.075	6/9 * F15 = 0.050	6/9 * F15 = 0.050
Can the user undo a product	9/9 * F1	9/9 * F16	6/9 * F16	8/9 * F16

selection?	= 0.071	= 0.075	= 0.050	= 0.066
Is it clear what must be done to make the transition to Checkout?	7/9 * F1 = 0.055	6/9 * F17 = 0.050	0/9 * F17 = 0.000	9/9 * F17 = 0.075
Does the system allow users to explicitly check on previous searches?	6/9 * F1 = 0.078	0/9 * F18 = 0.000	6/9 * F18 = 0.083	0/9 * F18 = 0.000
Percentage: Raw Task-weighted	 59/72=81.9 0.804=80.4	 51/72=70.8 0.678=67.8	 30/72=41.7 0.520=52.0	 42/72=58.3 0.574=57.4
Normalized Ratio: Raw Task-weighted	 1.0 1.0	 0.865 0.843	 0.509 0.647	 0.712 0.714

Table 3: Applying the task metrics to the LSD phase

Note: F1 Refers to the corresponding values as given in table 1. For example for "Is it clear what a user must do to search for a product?" F1 = (0.8+0.5)/8.5. For the three older websites F11 to F18 have the following values: F11=(0.8+0.5)/8, F12=(0.5+0.5)/8, F13=(0.8+0.5)/8, F14=(0.8+0.8)/8, F15=(0.1+0.5)/8, F16=(0.1+0.5)/8, F17=(0.1+0.5)/8, F18=(0.5+0.5)/8.

Evaluation of User Task: Checkout Stage	**Amazon–New (3 stages)**	**Amazon–Old (6 stages)**	**Kalahari (3 stages)**	**BOL (5 stages)**
Are instructions and messages concise, clear and unambiguous?	8/9 * F2 = 0.100	12/18 * F21 = 0.073	6/9 * F21 = 0.073	11/15 * F21 = 0.080
Are possible actions clear?	7/9 * F2 = 0.073	12/18 * F22 = 0.065	5/9 * F22 = 0.054	12/15 * F22 = 0.078
Is the required format of user inputs clearly indicated?	8/9 * F2 = 0.072	15/18 * F23 = 0.076	7/9 * F23 = 0.071	10/15 * F23 = 0.061
Are user actions linked to changes in the interface?	7/9 * F2 = 0.063	13/18 * F24 = 0.057	6/9 * F24 = 0.053	12/15 * F24 = 0.063
Is there always an appropriate response to user actions?	6/9 * F2 = 0.067	12/18 * F25 = 0.065	6/9 * F25 = 0.065	13/15 * F25 = 0.084
Does the system inform the user of the success or failure of their actions?	8/9 * F2 = 0.083	14/18 * F26 = 0.071	6/9 * F26 = 0.061	11/15 * F26 = 0.067
Does the system inform users of the reasons for delays?	7/9 * F2 = 0.073	11/18 * F27 = 0.056	3/9 * F27 = 0.030	5/15 * F27 = 0.030
Do error messages indicate the what, where, and why, and how to recover?	4/9 * F2 = 0.042	9/18 * F28 = 0.045	4/9 * F28 = 0.040	11/15 * F28 = 0.067
Is it clear what the user has to do to complete the task?	7/9 * F2 = 0.058	11/18 * F29 = 0.048	5/9 * F29 = 0.044	14/15 * F29 = 0.074
Does the system indicate the current stage?	8/9 * F2 = 0.044	17/18 * F30 = 0.046	3/9 * F30 = 0.016	15/15 * F30 = 0.048
Can the user easily back out of the process?	8/9 * F2 = 0.050	10/18 * F31 = 0.030	2/9 * F31 = 0.012	3/15 * F31 = 0.011
Does the user explicitly confirm the final purchase?	9/9 * F2 = 0.031	18/18 * F32 = 0.036	9/9 * F32 = 0.036	15/15 * F32 = 0.036
Can users check on inputs provided during the process?	7/9 * F2 = 0.029	9/18 * F33 = 0.018	6/9 * F33 = 0.024	3/15 * F33 = 0.007
Percentage: Raw Task-weighted	 94/117=80.3 0.785 =78.5	 163/234= 69.7 0.686 = 68.6	 68/117=58.1 0.579 = 57.9	 135/195=69.2 0.706 = 70.6
Normalized Ratio: Raw Task-weighted	 1.0 1.0	 0.868 0.874	 0.724 0.738	 0.862 0.899

Table 4: Applying the task metrics to the checkout phase

Note: F2 Refers to the corresponding values as given in table 2. For example for "Are possible actions clear?" F2 = (0.8+0.7)/16. For the three older websites F21 to F33 have the following values: F21=(1+0.8)/16.5, F22=(0.8+0.8)/16.5, F23=(0.5+1)/16.5, F24=(0.8+0.5)/16.5, F25=(0.8+0.8)/16.5, F26=(1+0.5)/16.5, F27=(1+0.5)/16.5, F28=(0.5+1)/16.5, F29=(0.5+0.8)/16.5, F30=(0.3+0.5)/16.5, F31=(0.1+0.8)/16.5, F32=(0.1+0.5)/16.5, F33=(0.1+0.5)/16.5.

7.1 Discussion of results

The results from Table 3 show that:

1. Applying the task weighting has decreased the overall usability difference between the best site – Amazon (2002), and the worst site – Kalahari. Small changes were observed for BOL. Elimination of low value (R+C) task metrics will result in larger differences between weighted and un-weighted results.
2. For the lowest usability site (Kalahari), applying the task weighting results in a significant increase in its overall usability score. This would imply that Kalahari does focus on better usability for important tasks compared to other metrics which have lower (R+C) values.
3. Applying the weighting factors has emphasised the usability differences between the new and old Amazon sites. The new Amazon has improved considerably on its usability score for the LSD phase – this is mainly due to higher scores for content layout, information presentation, the provision of a history function, and more obvious navigation to next stages in the book purchase task.
4. High (R+C) criteria include activities associated with the presentation of information, and instruction-oriented actions.
5. Low (R+C) criteria include product selection and de-selection actions, and undo facilities.
6. The raw scores also yield useful information by themselves. They provide an evaluation mechanism that can be used by developers to flag problem feedback areas. For example the old Amazon website did not have a search history facility and as a result scored 0 for metric 8. This was corrected in the new Amazon site.

The results from Table 4 show that:

1. The task weighting has improved the score of BOL, but decreased that for all three other sites. The smaller changes when compared to Table 3 are in part due to the larger number of metrics used in this table when compared to Table 3 – i.e. there is to some extent an averaging around the mean of the (R+C) value. This could be avoided by eliminating all metrics with low (R+C) values (for example values < 0.6) from the scoring.

2. The new Amazon has improved very noticeably on its usability score for the Checkout phase – this is in part due to the folding of six previous stages onto three from 2001 to 2002, and also because of higher scores for layout of user options, condensed information presentation, and again more intuitive navigation to the next stage in the book purchase task.
3. High (R+C) criteria include user guidance, appropriate responses and the clarity of interaction messages and information presentation.
4. Low (R+C) criteria include the confirmation of the purchase and abort facilities.
5. Red-flagged (problem) areas based on the raw (un-weighted) scores from Table 4 are the lack of meaningful error messages, unexplained delays, and no intuitive undo facilities.

The results, and especially the approach adopted, namely that of prioritizing certain criteria over another set tailors the metrics to the nature of the task. On an intuitive level, it is clear that repetition of a task should make it more important (i.e. increase its weight); that the level of interaction required should also increase its weight; that the task duration should increase its weight; and that the level of knowledge required for the task should also increase its contribution to the website's overall usability score. The method used here for obtaining task repetition values is easily implemented since it simply counts the occurrence of these during a typical (shopping and browsing) interaction session.

8. CONCLUSION

Feedback can be used to assist the user in understanding the functionality and requirements of an EC application and can be effectively harnessed to ensure that users do not abandon websites. Additionally E-Commerce applications are task-oriented and goal-directed Web-based interactions, thus they lend themselves to the use of feedback-supported structured task analysis approaches. This study offers additional perspectives on Web-based tasks, and also introduces a way of using task-analysis to improve Web-based usability through a task-based weighting scheme during evaluation. This extended evaluation metric scheme and provides a more finely tuned mechanism for assisting developers to improve usability of E-Commerce websites, since the user's task is included in the formulation of the guidelines. This mechanism makes use of a novel usability metric prioritising scheme to yield information that can be used during both the design and maintenance phases.

The approach as outlined here needs to be applied to a larger sample of EC sites. It is currently designed for websites that fit the LSD and Checkout model, but the principle of scoring and then weighting usability metrics will make it suitable for other types of E-Commerce sites such as Internet-based banking. This will require fortifying or changing the metrics, by replacing and re-designing some of the items. Developing a faster questionnaire that can be delivered via the Web, and to which users rather than experts can respond, will also be beneficial, and may facilitate the partial automation of the questionnaires. It would also be necessary to obtain reliability information on the results obtained, by for example comparing the results with those obtained by a heuristic evaluation, or by comparing with the results from other approaches such as Web log analytics or shopper simulation models.

9. REFERENCES

[1] J. M. C. Bastien and D. L. Scapin. *A Validation of Ergonomic Criteria for the Evaluation of Human-Computer Interfaces.* International Journal of Human-Computer Interaction. 4(2), 1992.

[2] R. Belew. *Finding Out About*. Cambridge University Press. USA, 2000.

[3] T. Brinck, D. Gergle and S. D. Wood. *Usability for the Web: Designing Web Sites that Work*. Morgan Kauffmann, 2002.

[4] A. J. Dix. Closing the loop: modelling action, perception and information. In M. F. C. T. Catarci, S. Levialdi, and G. Santucci, editors, AVI'96 – Advanced Visual Interfaces, pages 20-28. ACM Press, 1991.

[5] A. Dix. *Design of User Interfaces for the Web*. Paton N. W. and Griffiths T. (Eds.) Proc. User Interfaces to Data Intensive Systems (UIDIS99). 5–6 September. IEEE Computer Society Publishers, Edinburgh, Scotland, 1999.

[6] A. Dix, J. Finlay, G. Abowd and R. Beale. *Human-Computer Interaction*. Hemel Hampstead: Prentice Hall International. UK. 2nd Edition, 1998.

[7] F. L. Engel and R. Haakma. Expectations and Feedback in User-System Communication. International Journal of Man-Machine Studies, 39(3):427-452, 1993.

[8] V. Kaptelinin, B. A. Nardi and C. Macaulay. The activity checklist: a tool for representing the space of context. Interactions, 6(4), July-August 1999.

[9] M. D. Levi and F. G. Conrad. *A Heuristic Evaluation of A World Wide Web Prototype Interactions*. Interactions. July-August, pages 50–61,. 1996.

[10] J. Nielsen J. and R Mack (Eds.). *Usability Inspection Methods*. John Wiley and Sons, 1994.

[11] M. A. Perez-Quinones and J. L. Sibert. Negotiating User-Initiated Cancellation and Interruption Requests. In Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems, volume 2 of SHORT PAPERS: Models, pages 267-268, 1996.

[12] S. J. Ravden and G. I. Johnson. Evaluating Usability of Human-Computer Interfaces: A Practical Method. John Wiley and Sons, 1989.

[13] K. Renaud, T. van Dyk and P. Kotzé. *A Mechanism for Evaluating Feedback of E-Commerce Sites*. The first IFIP conference on E-Commerce, E-Business, E-Government. I3E. Zurich, Switzerland, 4-5 October, 2001.

[14] M. Singh, A. K. Jain, and M. P. Singh. E-Commerce over communicators: Challenges and solutions for user interfaces. In Proceedings of the ACM Conference on Electronic Commerce (EC-99), pages 177-186, N.Y., Nov. 3-5. ACM Press. , 1999.
[15] J. Veen. *The Art and Science of Web Design.* Indianapolis, Indiana, New Riders, 2001.
[16] W. E. Woodson, B. Tillman and P. Tillman. *Human Factors Design Handbook.* McGraw Hill, 2nd edition, 1992.

Web References

[W1] Dack.com. Best practices for designing shopping cart and checkout interfaces. http://www.dack.com/Web/shopping/cart.html. February 2001.
[W2] National Cancer Institute (NCI) USA. Research-based Web Guidelines.. http://www.usability.gov/guidelines/intro.html. February 2001
[W3] Checklist of Checkpoints for Web Content Accessibility Guidelines. World Wide Web Consortium. http://www.w3.org/TR/WAI-WEBCONTENT/full-checklist.html. March 2001.
[W4] J. Nielsen. Severity Ratings for Usability Problems. http://www.useit.com/papers/heuristic/severityrating.html. March 2001.
[W5] K. Bohmann. User Performance Metrics. http://www.bohmann.dk/articles/user_performance_metrics.html. December 2001