Tool for Accurately Predicting Website Navigation Problems, Non-Problems, Problem Severity, and Effectiveness of Repairs

Marilyn Hughes Blackmon†, Muneo Kitajima‡ and Peter G. Polson†

†Institute of Cognitive Science
University of Colorado at Boulder
Boulder, Colorado 80309-0344 USA
+1 303 492 5063
{blackmon, ppolson}@psych.colorado.edu

‡ ‡National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1, Higashi, Tsukuba, Ibaraki 305-8566 Japan
+81 29 861 6650
kitajima@ni.aist.go.jp

Abstract

The Cognitive Walkthrough for the Web (CWW) is a partially automated usability evaluation method for identifying and repairing website navigation problems. Building on five earlier experiments [2,4], we first conducted two new experiments to create a sufficiently large dataset for multiple regression analysis. Then we devised automatable problem-identification rules and used multiple regression analysis on that large dataset to develop a new CWW formula for accurately predicting problem severity. We then conducted a third experiment to test the prediction formula and refined CWW against an independent dataset, resulting in full cross-validation of the formula. We conclude that CWW has high psychological validity, because CWW gives us (a) accurate measures of problem severity, (b) high success rates for repairs of identified problems (c) high hit rates and low false alarms for identifying problems, and (d) high rates of correct rejections and low rates of misses for identifying non-problems.

Categories and Subject Descriptors: H.5.2 [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces – Evaluation/methodology, Theory and methods, User-centered design; H.5.4 [Information Interfaces and Presentation (e.g., HCI)]: Hypertext/Hypermedia – Navigation, Architectures, Theory, User issues; H.1.2. [Models and Principles]: User/Machine Systems – Human information processing, Human factors;

General Terms: Design, Theory; Verification, Experimentation; Performance; Measurement; Human Factors

Keywords: Cognitive Walkthrough for the Web, CWW, CoLiDeS, cognitive model, user model, Latent Semantic Analysis, LSA, usability problems, repairs, usability evaluation method, information scent, heading labels, link labels

INTRODUCTION

This paper focuses on significant advances in the development of the Cognitive Walkthrough for the Web (CWW) [1,3,4]. CWW is a usability evaluation method (UEM) that identifies and repairs problems hindering successful navigation of large, complex websites. Our first paper on CWW [4] described how we transformed the original Cognitive Walkthrough [26] to create CWW and validated the CWW problem-identification process against data from three experiments. Our second paper [3] reported two additional experiments that demonstrated the effectiveness of CWW-guided repairs for improving user performance.

In the work reported here we have taken two large steps forward. First, we have developed a method for calibrating the severity of the usability problems identified by CWW. Our new measure of problem severity is the *predicted mean total clicks* that users will make to accomplish a particular task on a specific webpage. We could equally well describe our measure of *problem severity* as a measure of *task difficulty* that is based on a principled theory and computational model of differences in task difficulty. This prediction formula is applicable to any task. It isolates particular factors that can cause any task to be more difficult, identifies which of the factors are contributing to the difficulty of each particular task, determines what amount of difficulty each factor contributes to the overall measure of difficulty, and sums these contributions to produce the predicted mean total clicks.

Second, we have increased the level of automation for CWW and paved the way for its full automation. The more automated ACWW interface and tutorial are available at http://autocww.colorado.edu/~brownr/ACWW.php> and <http://autocww.colorado.edu/~brownr/>. The new ACWW interface cuts time to perform CWW analyses to about one-sixth of the time it takes to perform the same CWW analyses at <http://autocww.colorado.edu>.

Practitioners and researchers can confidently use our predicted mean total clicks measure of problem severity. We demonstrate that the CWW problem severity measure is both reliable and psychologically valid. We rigorously evaluate the accuracy of the predicted mean total clicks, adhering to the rigorous standards for assessing usability evaluation methods (UEMs) that have been advocated by Gray and Salzman [9,10] and Hertzum and Jacobsen [11]. The work reported here required three experiments beyond those reported in our first two papers on CWW [3,4]. The compiled dataset is very large both in terms of the number and diversity of tasks tested in the laboratory (228 total tasks), and in terms of the number of experimental participants who did each task (generally 38 or more). We will also show that the predicted number of clicks is highly correlated with the probability of task failure and with mean solution time to perform the task.

For practitioners it is crucial to have both the increased automation and the accurate measure of problem severity. Practitioners function under strict time constraints, and they must therefore prioritize repairing the most serious usability problems first, fixing other problems only if time permits. Potential pragmatic users of this tool include educators creating distance-learning materials. Educators can also apply the tool to build web-based enhancements for regular courses and to help students learn to successfully navigate websites to find information.

Researchers, too, will benefit from the increased level of automation and accurate measure of problem severity. In its current form, however, CWW is still limited to assessing the usability of texts used for the headings and links of the navigation system, and this is only one aspect of webpage and website usability evaluation. Other researchers will now find it more feasible to integrate CWW with other cognitive models and UEMs. For example, Miller and Remington [21] have used estimates of heading/link label quality to settle questions about the optimal information architecture and number of links per webpage.

THEORETICAL FOUNDATIONS OF CWW

CWW is a theory-based usability inspection method [22] for detecting and correcting design errors that interfere with successful navigation of a website [1,3,4]. CWW, like the original Cognitive Walkthrough [26], is derived from a goal-driven theory of website exploration, CoLiDeS [15].

CoLiDeS, an acronym for Comprehension-based Linked model of Deliberate Search, extends a series of earlier models [16] of performing by exploration and is based on Kintsch's [14] construction-integration theory of text comprehension and problem solving processes. CoLiDeS is part of a broad consensus among theorists and website usability experts [5,6,7,8,13,20,21,23,24,25] that problem solving processes determine users' information-seeking or search behaviors when exploring a new website or carrying out a novel task on a familiar website.

CoLiDeS and other models cited in the previous paragraph, agree on the assumption that users, at any step in a task, consider a set of actions and select the action they *perceive* to be most *similar* to their current goal. The term *action* refers to both mental operations and physical actions, e.g., clicking on a link or attending to a subregion of a webpage.

CoLiDeS assumes that it takes a two-step process to generate a physical action on a webpage (e.g., clicking a link, button, or other widget). Step one is an *attention* process that parses a webpage into subregions, generating descriptions of each subregion from heading texts and from knowledge of webpage layout conventions. CoLiDeS then attends to the subregion whose description is perceived to be most similar to a user's current goal.

Step two is an *action selection* process that selects and acts on a widget (e.g., a link) from the attended-to subregion. Using a comprehension-based process, CoLiDeS generates a description of each widget and selects a widget that is perceived to be most similar to the current goal. Then it generates a description of actions for the selected widget and selects an eligible one by considering knowledge of website interface conventions. The processes involved in generating descriptions of subregions and physical actions are assumed to be analogous to the processes of text comprehension, described by Kintsch's construction-integration theory of comprehension [14]. In the applications of CoLiDeS described below, it is assumed that heading texts determine the description of subregions and that link texts determine the descriptions of widgets (e.g., links).

Figure 1 shows schematically how the CoLiDeS attention and action selection processes work along with mental representations of an example webpage generated during the attention process. In this example the user first parses the entire webpage into seven subregions and attends to the content area. Then the user parses the attended-to content area subregion and probably focuses on either of two sub-subregions, the leftmost sub-subregion, International, that is the correct one, or the rightmost sub-subregion, Other Sites, that competes for the user's attention. On the assumption that the user selected the correct sub-subregion, the user proceeds to an action selection process. In Figure 1, the link labeled by the text "Oceania" is the correct link to accomplish the user's goal of wanting information about traveling to New Zealand and hiking the national parks on the south island. Unfortunately, even when users focus on the correct heading, they may not click the correct link, Oceania, because even college-educated users have little background knowledge about Oceania. Oceania is, thus an unfamiliar term for users with college-level general reading knowledge, and they may not realize that New Zealand is located in Oceania.

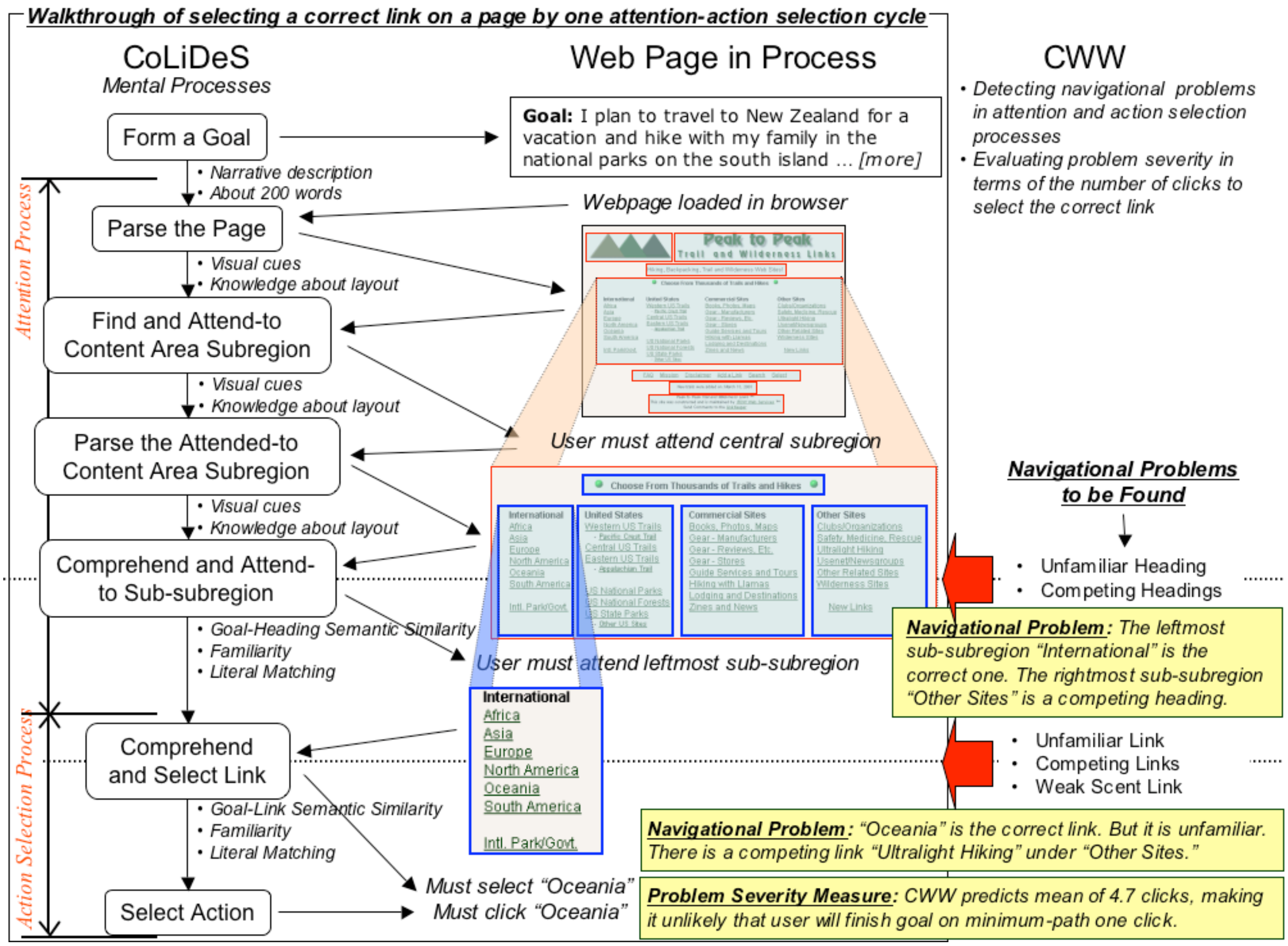

Figure 1. CoLiDeS model for how user accomplishes goal on Peak to Peak home page (<http://www.peaktopeak.net>) with CWW problem identification and prediction of 4.7 mean total clicks to accomplish user goal on this webpage

Performing a Task, Information Scent, and CWW

In the most straightforward case, the ideal case of pure forward search, performing a task (accomplishing the user's goal) involves making a series of *k* correct link choices that lead to a page that contains needed information or supports a desired action, such as purchasing a product. In the case of pure forward search CoLiDeS assumes that performing the task involves a sequence of *k* attention-action selection pairs, where on each page *both* the descriptions of the correct subregion and the correct link in that subregion are perceived to be most similar to the user's goal and are selected as a next move. A variety of alternative models of web navigation [5,6,7,8,23,25] describe the user's perceptions of similarity as *information scent* and the sequence of *k* pairs of perceptions as a scent trail. Successful completion of a task involves following a scent trail that leads a user to make correct choices at each step.

CWW [3,4] identifies usability problems derived from CoLiDeS's simulations of step-by-step user behavior for a given task on a particular webpage. CWW detects and corrects errors in the designs of webpages that can derail the simple scent-following process. For example, one or more incorrect alternatives may have equal or higher scent than the correct one and/or the correct alternative may have very weak scent. Following the two-step action selection process of CoLiDeS, CWW looks first for problems with headings and then with links nested under the headings.

Navigation Usability Problems CWW Detects

CoLiDeS predicts that users will encounter four types of usability problems while navigating websites to accomplish particular tasks (see Figure 1 for the relationship between the locations where the problem types occur and the corresponding CoLiDeS processes):

1. A *weak scent link* refers to the situation when a correct link is not semantically similar to the user goal and there are no other correct links that have moderate or strong similarity. CoLiDeS assumes that the user may never perceive the correct link as a useful target for action when it has weak scent. Users understand the text but perceive the link to be unrelated to their goals.

2. An *unfamiliar* problem occurs when typical users of the website lack sufficient background knowledge to comprehend a correct link or heading text. Unfamiliar problems happen when the topic is one that typical users know little about or when heading/link texts use technical terms or low frequency words that are novel for a particular user population. Unfamiliar texts have little or no meaning for typical users. Even if there is a strong objective similarity between the goal and the heading/link text, only users who comprehend the meaning can actually perceive the scent, not users who find the text unfamiliar.
3. A *competing headings* problem arises when any heading and its associated subregion is semantically very similar to the user goal but does not contain a correct link that leads to accomplishing the user goal. Competing headings problems are liable to be serious problems, because they divert the user's attention away from a correct heading that is on the solution path for that goal. CoLiDeS assumes that users will only attend to and click links in correct or competing subregions, ignoring links in other subregions.
4. A *competing links* problem occurs when a correct or competing subregion contains one or more links that are semantically similar to the user goal but not on the solution path. Competing links problems can occur even in the best-case scenario, when the user's attention has been first drawn to a semantically similar correct heading and its associated subregion. CWW now separately tallies the number of competing links that occur under competing headings and the number of competing links that occur under a correct heading.

Latent Semantic Analysis and Information Scent

CWW employs Latent Semantic Analysis (LSA) to compute similarities of goals with descriptions of subregions (headings) and possible physical actions in the attended-to subregion (link texts). Goals and descriptions are collections of words, and LSA can compute the similarity between any two collections of words.

LSA [17,18,19] is a machine learning technique that builds a semantic space representing a given user population's understanding of words, short texts (e.g., sentences, links), and whole texts. The meaning of a word, link, sentence or any text is represented as a vector in a high dimensional space, typically with about 300 dimensions. LSA generates the space from a very large collection of documents that are assumed to be representative of a given user population's reading experiences. While analyzing the distinctive characteristics of the particular user group, CWW evaluators choose the LSA semantic space whose corpus of documents best represents the background knowledge of the particular user group – the space built from documents that these users are likely to have read.

The CWW website (http://autocww.colorado.edu) currently offers a college level space for French and five spaces that accurately represent general reading knowledge for English at college level and at third-, sixth-, ninth-, and twelfth-grade levels. So far CWW researchers have tested predictions and repairs only for users with college-level reading knowledge of English, but they will soon expand to other reading levels and languages.

The degree of semantic relatedness or similarity between any pair of texts, such as the description of a user's goal and a link label on a webpage, is measured by the cosine value between the corresponding two vectors. Cosines are analogous to correlations. Each cosine value lies between +1 (identical) and -1 (opposite). Near-zero values represent two unrelated texts.

CWW uses LSA to compute the semantic similarities between user goals and subregion heading and link labels or descriptions of other widgets. CWW predicts that users attend to the subregion with the highest goal-heading (or goal-subregion) cosine value and the link or widget in the attended-to subregion with the highest goal-link (or goal-widget) cosine value. CWW represents user goals with realistic, narrative goal statements that are long enough for accurate LSA predictions (100-200 words).

Another important measure provided by LSA is term vector length, a measure that is correlated with word frequency, and that estimates how much knowledge about a word or phrase is embedded in the designated LSA semantic space. Words with low frequency in the corpus (e.g., specialized technical or scientific terms) have short term vector lengths. When a heading/link has a short term vector length, CWW predicts that users modeled by the semantic space will perceive it to be relatively meaningless, reducing the probability that users will attend to or click on it.

EXPERIMENTS 1 AND 2

The goals of Experiments 1 and 2 were to (a) replicate the findings of the foundational experiments [3,4], and (b) greatly enlarge the number and diversity of tasks tested in the laboratory, resulting in very large dataset sufficient for multiple regression analysis.

Replicate earlier experiments with 100 new tasks

The subjects (52 in Experiment 1 and 76 in Experiment 2) were undergraduates enrolled in an introductory psychology course, who completed the experiment as part of their course requirements. The procedure used a simulated encyclopedia website that presented a series of tasks. A countdown timer allowed subjects to spend no more than 150 seconds on each task. If the person found the correct encyclopedia article, clicking a link in a "correct item" box took the person on to the next item in the sequence. If the person did not find the correct encyclopedia article within the time limit, a time-expired page appeared. When the person clicked the link on the time-expired page, the next item appeared, just as it would have if the person had found the actual encyclopedia item.

Experiment 1 used 10 new tasks with CWW-identified goal-specific competing heading problems. Experiment 2 used 40 new tasks, half with competing headings problems and half with unfamiliar problems. In both experiments there were two webpage conditions, unrepaired and repaired. The webpages for Experiment 1 all displayed 74 links nested under 13 headings, presenting the goal statement at the top of the webpage. The webpages for Experiment 2 also displayed the goal at the top of the webpage but below the goal statement were 93 links nested under 9 headings. The repaired and unrepaired webpages for the same task looked identical, but the unrepaired webpage had only one correct link, the link that was correct on the online encyclopedia website being simulated. In contrast, the repaired webpage for competing heading problem tasks provided at least one correct link option under each and every competing heading. For unfamiliar problem tasks, there were two repairs: (a) substitution of a familiar link text for the unfamiliar link text, such as 'Paleontology and Fossils' in place of "Paleontology," and (b) addition of at least one correct link that was a familiar link, that was nested under the heading with the highest goal-heading cosine, and that was the link with the highest goal-link cosine compared to other links in that subregion.

For both experiments there were two groups of subjects, and everyone unknowingly alternated back and forth between doing a task on the repaired webpage condition and then doing a task on the unrepaired webpage condition. For each task it was possible to compare performance for the group that did the task in the repaired webpage condition, and the other group that did the task in the unrepaired webpage condition. To analyze the data we used Repeated Measures ANOVA, after first computing an average for each subject's performance on all tasks of the same problem type in the same webpage condition. We used different orders for presenting the items, but no order effects were found in either experiment, allowing us to ignore presentation order.

For Experiment 1 both subject groups had identical means of 5.62 clicks for the fives tasks done on the unrepaired webpages, and the between-group difference was very slight for the five tasks done on repaired webpages – 1.81 clicks compared to 1.96 clicks. There was a statistically significant difference between the repaired and unrepaired webpage conditions, $F\ (1,50) = 253.46$, $p < .0001$, replicating earlier findings [3].

For Experiment 2 the means for unrepaired vs. repaired competing heading problem tasks was 6.46 and 2.01, respectively. For unfamiliar problem tasks the means for unrepaired vs. repaired were 5.70 and 2.55, respectively. Technically we verified the hypothesis that the repaired webpage condition produced better performance by finding a significant interaction between group condition and the means for the four sets of 10 items: odd-numbered competing headings items, even-numbered competing headings items, odd-numbered unfamiliar items, and even-numbered unfamiliar items, $F\ (3, 222) = 244.67$, $p < .0001$. This pattern of results, showing highly significant differences between repaired and unrepaired webpage conditions, again replicates our earlier findings [3].

Compile dataset for multiple regression analysis

We used multiple regression analysis to derive a formula for predicting problem severity. Successful use of multiple regression requires a large dataset of genuinely comparable items. Experiments 1 and 2 provided 100 tasks for the regression analysis, and we were able to reuse data for 64 tasks from two webpage conditions (unrepaired and repairedNoExamples) from two earlier experiments [3].

The result was a dataset with a total of 164 tasks that compiled all the tasks done under experimental conditions that met specific criteria for inclusion. The criteria required pairs of tasks. For each pair of tasks, the goal was identical for two well-matched experimental groups, but one experimental group tried to accomplish the goal on an unrepaired webpage and a second group tried to accomplish the same goal on a repaired webpage. For the sample of tasks done in the unrepaired condition, the tasks manifested diverse combinations of competing headings, competing links, unfamiliar links, and weak-scent links.

The resulting compilation consisted of 82 pairs of tasks, 164 tasks altogether. For all 164 tasks we set a minimum 0.76 cosine between the actual webpage content article and the goal (summary of the article) shown to experimental participants, ensuring that experimental participants had an accurate representation of the complete article they were trying to find in the website. The experimental groups that met the criteria were drawn from four different experiments, and no tasks done by these experimental groups were excluded from the dataset for re-analysis.

Like any other type of problem-solving behavior, performance on these tasks exhibits a lot of between-subject variance, and 20 experimental participants per task is considered the minimum to produce stable means for problem-solving tasks. We far exceeded that minimum. To ensure stable estimates of mean performance for each of the 164 tasks, the mean clicks for 144 of the tasks were based on data from at least 38 experimental participants, and the means for the remaining 20 tasks were based on the data from at least 23 experimental participants.

Automatable rules solve reliability problem

We then developed a procedure for re-analyzing all 164 items. Towards that end, we iteratively rescored the set of 164 tasks until we had created a set of automatable rules for identifying competing headings, competing links, unfamiliar links, and weak-scent links. These rules and the accompanying rationale for each rule can be downloaded from <http://autocww.colorado.edu/~blackmon/Tutorials/AutomatableRules.doc>, and Figure 2 describes, step-by-step, the complex CWW procedure with the current edition of its parameters.

Automatable rules eliminate the subjective, time-consuming hand editing of LSA analyses that the CWW creators originally thought necessary [4], paving the way for more complete automation of CWW available in the ACWW interface at <http://autocww.colorado.edu/~brownr>. These automatable rules solve the reliability problem inherent in hand-edited LSA analyses by using completely objective rules to identify competing headings/links and weak-scent/unfamiliar links. These rules build on objective LSA measures of similarity and familiarity, avoiding the deficiency of low inter-rater agreement in UEMs [11].

The automatable rules are all written as if-then production rules, making it easy for a computer programmer to write code to fully automate the CWW problem-identification process. For example, a competing heading is a heading that pulls attention away from a correct heading, but the automatable rules specify two different sets of precisely defined conditions that can independently prompt classification of the heading as a competing heading. The first set has three conditions that must all be simultaneously met: (a) the goal-heading cosine must be greater than or equal to 0.8 times the highest goal-heading cosine of any correct heading, (b) the goal-heading cosine must be greater than or equal to 0.10 (i.e., not weak scent), and (c) the highest goal-link cosine for the links nested under the heading must be greater than or equal to 0.20.

Development of the prediction formula

Finally we developed a multiple regression model of task difficulty. For initial laboratory studies [3,4] CWW researchers had deliberately selected tasks that each epitomized one class of usability problems, either competing headings, competing links, or unfamiliar problems. In actual fact, however, few tasks are pure examples of just one of the four usability problems. Most tasks are afflicted by more than one type of usability problem, and some tasks are afflicted by all four of the CWW problems.

By doing a multiple regression analysis of the 164-item data set we tried to account for the variance in task difficulty, indexed by mean total clicks. For the full 164-item dataset the mean total clicks ranges from 1.0 click to 10.3 clicks with a mean of 3.7 clicks. The observed multiple regression weights evaluated how much each type of usability problem contributed to the overall difficulty level.

The multiple regression analysis resulted in a regression model of task difficulty that explains 57% of the variance in observed mean total clicks as a function of three independent variables, F (4, 160) = 74.22, $p < .0001$, adjusted $R^2 = 0.574$. All three independent variables are statistically significant: (a) whether or not the only correct link was unfamiliar, $t = 5.1$, $p < .0001$, (b) whether or not the only correct link was a weak scent link, $t = 5.8$, $p < .0001$, and (c) number of competing links nested under competing headings, $t = 10.8$, $p < .0001$. The intercept is also significant, $t = 14.0$, $p < .0001$.

1. Select the most appropriate semantic space to represent a particular user group.
2. Collect a set of user goals to represent what that user group is likely to want to accomplish on the website under analysis.
3. Simulate how the user will parse the webpage and identify all the individual subregions of the webpage.
4. Simulate the process of elaboration that occurs during comprehension of short heading and link texts.
5. Apply the LSA One-to-Many analysis to compare the goal statement with the elaborated headings and links and then sort the results first by headings vs. links, and then by decreasing cosine value. Then examine the sorted results and identify and mark the correct heading(s) and link(s), the ones that actually lead to accomplishing the goal in the actual online website being simulated.
6. Apply the automatable set of rules for distinguishing unfamiliar correct links, weak-scent correct links, competing headings, competing links under correct headings, and competing links under competing headings.
7. Examine the results and see how to repair the problems.
8. Apply the CWW Problem Severity Level formula for predicting the mean total clicks under both the repaired and unrepaired condition.

Figure 2. Procedure for evaluating problem severity

The minimum solution path for all 164 tasks was a single click, but the statistically significant intercept of 2.199 reveals that even the non-problem tasks took an average of over two clicks to complete. The intercept and unweighted regression coefficients give us a working formula for predicting the mean total clicks:

Mean total clicks = 2.199
+ 1.656 *if the correct link is unfamiliar*
+ 1.464 *if the correct link has weak-scent*
+ 0. 754 *times the number of competing links nested under competing headings*

Evaluation of prediction formula in same dataset

The next step after completing the multiple regression analysis was to apply the multiple regression formula to predict the mean total clicks for each of the 164 tasks, and Table 1 displays the accuracy of the predictions by comparing predicted and observed mean total clicks.

Table 1 sorts all 164 items into three groups, one for predicted non-problem items that the CWW formula predicted would be done in less than 2.5 mean total clicks, one for moderate problems (predicted to be between 2.5 and 5.0 clicks), and one for serious problems (5.0 or more clicks). These threshold values yielded three groups with similar numbers of tasks per group: 65 non-problem items, 55 moderate problem items, and 44 serious problem items, and for all three groups the observed values for mean total clicks are very close in value to the corresponding predicted values for mean total clicks.

Table 1. Comparison of observed and predicted scores for 164-item dataset.

Problem Severity Level	*Observed*	*Predicted*
No Problem (predicted clicks 1.0–2.5)	2.17	2.20
Moderate Problem (predicted clicks 2.5–5.0)	3.52	3.80
Serious Problem (predicted clicks 5.0 and up	6.43	6.17

To test whether these thresholds were reasonable for distinguishing non-problems from moderate problems and moderate problems from serious problems, we drew from the 164-item dataset the 100 items for which we have recorded percentages of task failure, i.e., percentages of experimental participants who did not complete the task in the allotted time (usually 130 seconds). We did a simple regression of the percentages of task failure per task on observed mean total clicks for the same task, finding a correlation of 0.93, F (1, 98) = 651.78, p < .0001, adjusted R^2 = .87.

We then used the regression formula (percent task failure = -.154 + 0.082 times observed mean total clicks) to estimate a task failure rate of 5% at 2.5 mean total clicks (operationally defined as the threshold between non-problem and problem items), 26% at 5.0 mean total clicks (operationally defined as the threshold between moderate and severe problems), 51% at 8.0 mean total clicks, and 76% at 11.0 mean total clicks.

We can provide internal validation of the multiple regression analysis another way, by subdividing the 164-task dataset into unrepaired and repaired tasks. For the 82 unrepaired tasks in the 164-task dataset the predicted and observed clicks were 5.02 vs. 5.22, and for the 82 repaired tasks the predicted and observed clicks were 2.29 vs. 2.09.

Rates of Hits vs. False Alarms

Even though Table 1 shows little discrepancy between predicted and observed mean total clicks, a more exacting standard is to examine hit rates vs. false alarm rates for the unrepaired tasks within the total dataset of 164 tasks. At the time the experiments were performed, 82 tasks were selected as unrepaired tasks. By then-defined criteria, these 82 unrepaired tasks were all predicted to be problems. The current, more accurate CWW procedure, however, diagnosed only 75 of these 82 tasks as problem tasks.

The overall hit rate for these 75 tasks in the unrepaired condition is 92% (69/75), and the overall false alarm rate is 8% (6/75). For the 46/75 tasks that had predicted serious problems (predicted mean clicks of 5.0 or higher), the hit rate was 100% and the false alarm rate was 0%. In other words, 46/46 (100%) tasks had observed mean clicks of 2.5 or greater. Of these 46 problems that were predicted to be serious problems 36/46 (78%) actually had observed mean clicks of 5.0 or higher (the other 10 had observed mean clicks between 2.5 and 5.0).

Success Rates for Repairs

Another important question concerns the success rate for CWW repairs of problems. A rigorous standard for a successful repair requires a statistically significant superiority in performance for experimental participants who performed the task on the repaired webpage compared to experimental participants who performed the task on the unrepaired webpage. The success rate, then, is the percent of all unrepaired problem tasks that meet this rigorous standard for successful repair. Of the 82 unrepaired tasks in the original dataset, 75 are predicted to be problems by the current criteria, and the overall success rate for repairs is 83% (62/75). For the 46/75 predicted problems that were predicted to be serious problems (predicted mean clicks 5.0 or more), however, the success rate was much higher – 43/46 (93%). This is important to practitioners, because it shows that they will reap solid benefits for investing effort in repairing the serious problems, problems that would cause high rates of task failure if left unrepaired.

As reported earlier [3], we found statistically significant differences (p <.0001) between repaired and unrepaired conditions in repeated-measures ANOVA analyses in our initial studies of CWW-guided repairs for 32 tasks. By current criteria only 25 of the 32 tasks are predicted to be problems, and our re-analysis of the initial study found that the success rate for repairs was only 16/25 (64%) for these 25 tasks. In subsequent experiments we added 50 additional tasks in both unrepaired and repaired conditions, and the success rate has risen to 46/50 (92%) for these tasks. While developing the formula for accurately predicting mean total clicks for both unrepaired and repaired tasks, therefore, we have also improved our success rate for repairs. As a result, we can now set a higher standard for evaluating the success of repairs. The higher standard is an important advance, because practitioners' need to know that the time they invest in repairing problems is wisely spent.

Among all 75/82 unrepaired tasks predicted to be problems, we have consistently found unfamiliar problems to be particularly challenging to repair. Unfamiliar problems are difficult to repair, because there is no easy way to compensate for users' low background knowledge of a topic. Out of the 75 tasks predicted to be problems, 28 have an unfamiliar correct link, and the success rate was only 21/28 (75%) for these tasks. For the other 47/75 tasks with no unfamiliar problem, the success rate was 41/47 (87%).

CROSS-VALIDATION EXPERIMENT

We ran a third experiment to see if the prediction formula replicated with a completely new set of tasks, and we included in the set 35 tasks predicted to be non-problems. The subset of 35 non-problems enabled us to measure rates of correct rejections vs. misses for the first time.

Subjects

The 113 experimental participants were all enrolled in the introductory psychology course and completed the experiment for course credit.

Table 2. Compare coefficients for 164-task dataset and 64-task cross-validation dataset (in parentheses)

Independent Variable	Unweighted Coefficient	Standard Coefficient
Intercept	2.199 (2.481)	2.199 (2.481)
Competing Links Under Competing Headings	.754 (.551)	.578 (.423)
Unfamiliar Correct Link	1.656 (2.040)	.264 (.330)
Weak-Scent Correct Link	1.464 (1.484)	.254 (.280)

Materials

We selected a total of 64 new items from the same online encyclopedia website from which the 164-item dataset was drawn for the development of the prediction formula. The simulated encyclopedia website has a total of 93 links nested under nine headings. For each of the 64 tasks, non-problem tasks as well as problem tasks, one and only one of the 93 links actually led to the target encyclopedia article. This created a conservative test of the ability of the prediction formula to distinguish between tasks that users can do easily compared to tasks associated with high rates of task failure. Using the prediction formula, we computed the predicted mean total clicks for all 64 items. For each of the 35 non-problem tasks the prediction was identical: 2.20 mean total clicks. For the 29 problem tasks the formula predicted a mean of 5.43 clicks, ranging from 3.66 to 8.19.

Procedure

Experimental participants first did five practice tasks and then alternated between non-problem and problem tasks, preventing discouragement from encountering two or more difficult items in a row. We divided the 64 tasks into two sets of 32 tasks with the same percentage of non-problem and problem tasks. There were two experimental groups, and each group did one set of 32 tasks after completing the five practice items. To control for order effects, we used three randomly ordered sequences for each set of 32 tasks.

The main webpages displayed the words "Find an encyclopedia article about X" followed by a summary (100-200 words) of the target encyclopedia article. This task description appeared at the top of each of the headings-links webpages, and below it appeared nine headings and up to one set of links associated with one of the headings. The summary was highly similar to the text of the actual encyclopedia (operationally defined as a minimum LSA cosine of 0.80). The minimum cosine ensures that experimental participants have an accurate representation of each article they were asked to find. The summary was available throughout the search time. A countdown timer limited search time for each task to 130 seconds.

Results

As a within-subject variable we computed the average clicks for the non-problem tasks and for the problem tasks. Both groups performed consistently better on non-problem tasks than on problem tasks, and the difference was significant, paired t-test (df = 52) = 17.44, $p < .0001$, and paired t-test (df = 59) = 28.23, $p < .0001$, respectively.

Table 3. Predicted vs. Observed Means for 64 Tasks Included in Cross-Validation Experiment

Type of Task (64 Tasks)	Mean Total Clicks	
	Predicted	Observed
Predicted Non-Problems	2.20	2.20
Predicted Problems	5.43	5.68

We repeated the multiple regression analysis with the cross-validation set of 64 new tasks and three independent variables explained 50% of the variance, F (3, 60) = 22.042, p <.0001. The three independent variables are identical to the three independent variables found for the 164-task original dataset. All three independent variables are statistically significant: (a) whether or not the only correct link was unfamiliar, t = 3.6, p = .0007, (b) whether or not the only correct link as a weak scent link, t = 3.0, p = .0037, and (c) number of competing links nested under competing headings, t = 4.5 p <.0001. The intercept is also significant, t = 9.7, p <.0001. The correlation coefficients for the cross-validation sample of 64 tasks were similar to the correlation coefficients for the 164-task dataset for the original multiple regression analysis, as shown in Table 2.

The observed and predicted mean total clicks for problem and non-problem items are shown in Table 3. We computed the predicted mean total clicks using the prediction formula derived from the 164-task dataset. For the 35 tasks with predicted clicks less than 2.5 (predicted non-problems), the predicted and observed clicks were both 2.20. For the 29 tasks predicted to be problems, the predicted mean total clicks averaged 5.43, just under the observed value of 5.68. It is important to remember that all 64 tasks in the cross-validation study – both predicted non-problems and predicted problems – each had only one correct link out of 93 total links. The CWW prediction formula was highly successful in distinguishing non-problem tasks for which people would locate and click the one and only correct link within two or three attempts, compared to problem tasks for which it would take an average of over five clicks to find the correct link.

For the 29 tasks predicted to be problems, the hit rate was 26/29 (90%), and the false alarm rate was 3/29 (10%). That is, 26/29 had observed mean total clicks greater than or equal to 2.5, and 3 had observed mean total clicks less than 2.5. These rates of hits vs. false alarms for the cross-validation dataset are similar to the rates of hits vs. false alarms for the 164-task dataset.

Even more important, the cross-validation dataset provided valid rates of correct rejections vs. misses, a statistic for which we have previously not had adequate data. For the 35 tasks that were predicted to be non-problem items (i.e., tasks with predicted mean total clicks <2.5), 24/35 (69%) had observed mean clicks of less that 2.5, the rate of correct rejections. Although 11/35 (31%) were misses, only 4/35 (11%) had observed mean clicks of greater than 3.5. and none was a serious problem (i.e., none had observed mean clicks equal or greater than 5.0).

Discussion

What we have shown is that we can use a multiple regression model derived from our earlier studies to fully replicate the multiple regression analysis and accurately predict performance on a new group of diverse tasks with new participants. Unlike the experiments from which the 164-task dataset were drawn, the experiment from which the cross-validation tasks were drawn was a closer simulation of the actual online website. Nevertheless, the simulated website was an online encyclopedia that was the same for both the 164-task original dataset and the 64-task cross-validation dataset. The advantage of online encyclopedias is that they cover all possible topics and use familiar categories that were learned in school. Although it is reasonable to assume that the results from this diverse sample of online encyclopedia tasks will generalize to any informational website, this cross-validation study provides no evidence to confirm that assumption.

CONCLUSIONS

Progress Toward Automating CWW

A truly useful CWW has to be automated, and the work reported here represents an important advance in that direction. The automatable rules for problem-identification and the new prediction formula, combined with LSA, together pave the way to full automation of the CWW method. Ivory-Ndiaye (2003) reviews currently available automated tools for usability evaluation of websites and comes to a similar conclusion about their necessity.

Reliability of CWW

We have solved the reliability problems inherent in most UEMs [11], including the original Cognitive Walkthrough. From the outset CWW has largely eliminated reliance on subjective human judgments by substituting LSA to objectively estimate similarity and familiarity. The work reported here – development of automatable rules for problem identification and the new formula for predicting problem severity – free us almost completely from relying on subjective human judgments.

Validity of CWW and Measure of Problem Severity

The psychological validity of CWW for college-educated user populations is demonstrated by our high rates of hits vs. false alarms, correct rejections vs. misses, high success rates for repairs, and the accuracy of our new measure of problem severity. Gray and Salzman [9,10] have criticized previous UEM evaluation studies for failure to report such statistics, and Gray and Salzman, along with Hertzum and Jacobsen [11], found that UEMs typically do a very poor job of rating usability problem severity.

Our problem severity measure has the highest hit rate for identifying serious navigation usability problems, and for the most serious problems we also find that CWW-guided repairs of navigation usability problems have the highest rates of statistically significant performance improvements. With the help of this tool, therefore, usability experts and web developers can find and repair the problems most worth repairing.

Why Cross Validation Is Critical

The cross validation study of 64 new tasks was a balanced sample of diverse non-problem and problem tasks. As shown in Table 2, this study successfully replicated the prediction formula derived from the initial dataset. The excellent agreement of the estimated regression coefficients is in part due to the large samples of subjects and tasks. The current parameters of the prediction formula are relatively stable and consistent with the CoLiDeS cognitive model, but it is likely that these parameters will require fine-tuning when extended to much larger samples of tasks done on a wide variety of informational websites.

Limitations of CWW

The principal limitations of CWW are a consequence of the tasks used in most of our experiments to date: searching for experimenter-specified articles on a simulated encyclopedia website, using experimental webpages that feature textual headings and links in simplified webpage layouts with no graphics (see experiments at <http://autocww.colorado.edu/~blackmon>). The current version of CWW nevertheless dovetails with research on other aspects of website usability. For example, Miller and Remington's [21] simulation results demonstrate how the structure of a site interacts strongly with patterns and variances of scent values on a webpage.

Our experiments have so far [3,4] been limited to testing predictions of heading and link selection for college-educated users. A driving motivation in our work has been our hypothesis that we can successfully extend the CWW to evaluating websites for user groups who speak any language at any level of general reading knowledge. Using a variety of LSA semantic spaces, we expect to be able to soon extend CWW beyond college-educated populations and make reliable, psychologically valid judgments of diverse user populations, including users of all ages and levels of background knowledge and members of other language and cultural groups.

Contribution and Benefit

The most important contribution of this paper is the development of a measure of problem severity (task difficulty) that is theoretically sound, psychologically valid, reliable, and automatable. This measure of problem severity benefits both researchers and practitioners. As the size and diversity of the dataset continues to grow over time, we can recalibrate the parameters of the formula for higher accuracy. We will, therefore, periodically post updates in the formula at <http://autocww.colorado.edu/~brownr> and will also track expanded CWW functionality at <http://auto.colorado.edu/~blackmon>.

ACKNOWLEDGMENTS

This work was supported by an NSF grant, EIA-0137759 to the first author.

REFERENCES

1. Blackmon, M. H. Cognitive Walkthrough. In W. S. Bainbridge (Ed.), *Encyclopedia of Human-Computer Interaction*, 2 volumes. Great Barrington, MA: Berkshire Publishing, 2004.
2. Blackmon, M. H., Kitajima, M., Mandalia, D. R., & Polson, P. G. Automating Usability Evaluation: Cognitive Walkthrough for the Web Puts LSA to Work on Real-World HCI Design Problems. In T. K. Landauer, D. S. McNamara, S. J. Dennis, and W. Kintsch (Eds.), *LSA: A Road to Meaning*. Mahwah, NJ: Erlbaum (to appear, 2006).
3. Blackmon, M. H., Kitajima, M., & Polson, P.G. (2003) Repairing usability problems identified by the Cognitive Walkthrough for the Web. *Proc. of CHI 2003*, ACM Press (2003), 497–504.
4. Blackmon, M. H., Polson, P. G., Kitajima, M., & Lewis, C. Cognitive Walkthrough for the Web. *Proc. of CHI 2002,* ACM Press (2002), 463–470.
5. Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. Using information scent to model user information needs and actions and the Web. *Proc. of CHI 2001,* ACM Press (2001), 490–497.
6. Chi, E., Pirolli, P., & Pitkow, J. The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a website. In *Proceedings of CHI 2000*, ACM Press (2000), 161–168.
7. Chi, E. H., Rosien, A., Supattanasiri, G., Williams, A., Royer, C., Chow, C., Robles, E., Dalal, B., Chen, J., & Cousins, S. (2003). The Bloodhound Project: Automating discovery of web usability issues using the InfoScent™ Simulator. *Proc. of CHI 2003,* ACM Press (2003), 505–512.
8. Furnas, G. W. Effective view navigation. *Proc. of CHI'97*, ACM Press (1997), 367–374.
9. Gray, W. D., & Salzman, M. C. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human–Computer Interaction, 13 (1998)*, 203–261.
10. Gray, W. D., & Salzman, M. C. Repairing damaged merchandise: A rejoinder. *Human–Computer Interaction, 13* (1998), 325–335.
11. Hertzum, M, & Jacobsen, N.E. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human–Computer Interaction, 15* (2003), 183–204.
12. Ivory-Ndiaye, M. Y. An empirical approach to automated website evaluation. *Journal of Digital Information Management, 1* (2003), 75–102.
13. Katz, M. A., & Byrne, M. D. Effects of scent and breadth on use of site-specific search on e-commerce websites. *ACM Transactions on Computer-Human Interaction*, *10* (2003), 198–220.
14. Kintsch, W. *Comprehension: A paradigm for cognition*. Cambridge, U.K. & New York: Cambridge University Press, 1998.
15. Kitajima, M., Blackmon, M. H., & Polson, P. G. (2000). A Comprehension-based model of Web navigation and its application to Web usability analysis. *Proc. of HCI 2000*. Springer-Verlag (2000), 357–373.
16. Kitajima, M., & Polson, P. G. A comprehension-based model of exploration. *Human–Computer Interaction, 12* (1997), 345–389.
17. Landauer, T. K. Learning and representing verbal meaning: Latent Semantic Analysis theory. *Current Directions in Psychological Science*, 7 (1998), 161–164.
18. Landauer, T. K. & Dumais, S. T. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104* (1997), 211–240.
19. Landauer, T. K., Foltz, P., & Laham, D. An introduction to Latent Semantic Analysis. *Discourse Processes*, 25 (1998), 259–284.
20. Larson, K., & Czerwinski, M. Webpage design: Implications of memory, structure and scent for information retrieval. In *Proceedings of CHI'98*, ACM Press (1998), 25–32.
21. Miller, C. S., & Remington, R. W. Modeling Information Navigation: Implications for Information Architecture. *Human–Computer Interaction, 19* (2004), 225–271.
22. Nielsen, J., & Mack, R. L. *Usability Inspection Methods*. New York: John Wiley & Sons, Inc., 1994.
23. Pirolli, P. The use of proximal information scent to forage for distal content on the World Wide Web. In Kirlik, A. (Ed.), *Working with technology in mind: Brunswikian resources for cognitive science and engineering,* in press.
24. Pirolli, P., & Card, S. K. Information foraging. *Psychological Review, 106* (1999), 643–675.
25. Pirolli, P. L., & Fu, W. SNIF-ACT: a model of information foraging on the World Wide Web (9th International Conference on User Modeling). *Lecture Notes in Artificial Intelligence, 2702* (2003), 45–54.
26. Wharton, C., Rieman, J., Lewis, C., & Polson, P. The cognitive walkthrough method: A practitioner's guide. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 105–140). New York: John Wiley, 1994.