Usability evaluation of an interactive decision support system for user-guided design of scenarios of watershed conservation practices

Adriana D. Piemonti, Kristen L. Macuga and Meghna Babbar-Sebens

ABSTRACT

This paper evaluated user interaction with the graphical user interface of WRESTORE, an environmental decision support system (EDSS) for watershed planning that utilizes user ratings of design alternatives to guide optimization. The following usability metrics were collected for stakeholders as well as surrogates (who are often used for DSS development due to time or cost constraints): task times across sequential sessions, percentage of time spent and of mouse clicking in different areas of interest, and trends in self-reported user confidence levels. Task times conformed to theoretical models of learning curves. Stakeholders, however, spent 15% more time and made 14% more mouse clicks in information gathering areas than surrogates. Confidence levels increased over time in 67% of stakeholders, but only in 29% of surrogates. Relationships between time spent or mouse clicking events and confidence level trends indicated that confidence ratings increased over time for users who conducted more information gathering. This study highlights the importance of increasing user interactions within information gathering areas of an interface to improve user-guided search, and suggests that developers should exercise caution when using surrogates as proxies for stakeholders. It also demonstrates a quantitative way to evaluate EDSS that could assist others developing similar tools.

Key words | decision support system, genetic algorithms, human–computer interface, participatory design, planning, usability

Adriana D. Piemonti (corresponding author)
Meghna Babbar-Sebens
School of Civil and Construction Engineering,
Oregon State University,
101 Kearney Hall,
Corvallis,
OR 97331,
USA
E-mail: *debora.piemonti@gmail.com*

Kristen L. Macuga
School of Psychological Science,
Oregon State University,
Reed Lodge, 2950 SW Jefferson Way,
Corvallis,
OR 97331,
USA

INTRODUCTION

Decision support systems (DSSs) are commonly used to support the design of watershed plans for sustainable management of water resources and landscapes (Jakeman *et al.* 2008; Lavoie *et al.* 2015). In recent years, stakeholders' participation in the planning and design process has been recommended as an approach to improve stakeholder adoption of watershed plans, and for successful restoration of watershed ecosystems impacted by land-use changes, climate change, etc. (Gregory 2000; Jakeman *et al.* 2008; Voinov & Bousquet 2010; Babbar-Sebens *et al.* 2015). The goals of these watershed restoration plans are to improve the hydrological and ecological functions of the land, without deterioration of the existing agricultural and recreational services provided by the watershed (Kelly & Merritt 2010). Diverse best management practices (BMPs), such as wetlands, filter strips, grassed waterways, cover crops, no-till practices, among others, have been proposed as solutions to prevent or reduce pollutant loads in water bodies, and for mitigation of flood events through runoff control and peak flow reduction (Ice 2004; Arabi *et al.* 2007; Artita *et al.* 2008).

doi: 10.2166/hydro.2017.017

Achieving an optimal distribution and selection of BMPs that is also acceptable to the stakeholder community is an inherently complex process. Multiple researchers have investigated coupled simulation models and optimization algorithms to identify the optimal distribution of BMPs in a watershed (e.g., Arabi *et al.* 2006; Kelly & Merritt 2010; Lethbridge *et al.* 2010; Tilak *et al.* 2011; Kaini *et al.* 2012). In these studies, simulation models have been used to estimate the landscape responses and evaluate related design objectives, whereas optimization algorithms have been used to find design alternatives in large decision spaces that enhance one or more objectives in the watershed. While such analytical techniques have the benefits of generating optimized scenarios of design alternatives with respect to quantifiable goals, they are limited in their ability to incorporate diverse subjective, unquantifiable or unquantified criteria (such as personal or social values, beliefs, interests, biases, local knowledge, etc.) and preferences of stakeholders. This inability has likely contributed to the unsuccessful adoption of 'optimal' design alternatives, and because of this the use of optimization algorithms in watershed planning has been criticized by some (Mendoza & Martins 2006).

In an attempt to find more effective ways to include stakeholders in the design process and potentially increase adoption of BMPs, researchers have begun to explore approaches, including those that use information technology, to provide platforms where the affected stakeholders can express their unique socio-economic and subjective constraints during the design process. For example, analytical approaches such as systems dynamics models (Metcalf *et al.* 2010), Bayesian networks (Castelletti & Socini-Sessa 2007; Zorrilla *et al.* 2010), fuzzy sets and cognitive maps (de Kok *et al.* 2000), agent-based models (Barreteau & Abrami 2007), and human-guided/interactive optimization (Babbar-Sebens & Misnker 2011) combined with graphical user interfaces (GUIs) have all been used to create participatory decision support platforms aimed at improving the engagement of stakeholders. These semi-structured and interactive DSS elicit data, knowledge, and feedback from stakeholders, and then use that information to generate design alternatives. A recent compilation by Jakeman *et al.* (2008) and McIntosh *et al.* (2011) exposed relevant aspects that should be considered when developing environmental DSS (EDSS) for watershed planning and management. McIntosh *et al.* (2011) emphasized the importance of the evaluation of human factors such as needs, goals, fatigue, learning, etc., in the design of interfaces used for EDSS. Similarly, a growing number of studies in water management have indicated interest in improving EDSS usability to enhance the user's understanding of the issues, as well as the design alternatives proposed for solving the problem (Johnson 1986; Power & Sharda 2007; Kirchoff *et al.* 2013; Babbar-Sebens *et al.* 2015).

ISO Standard 9241 defines usability as the 'extent to which a product can be used by the specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use'. Estimating usability attributes for different EDSS can potentially help developers to identify users' interactions with their system, and gain a better understanding of their behaviors, preferences, and needs. Some researchers, such as Haklay & Tobon (2003), Slocum *et al.* (2003), Jankowski *et al.* (2006), Nyerges *et al.* (2006), Boisvert & Brodaric (2012), Lavoie *et al.* (2015), and Sadler *et al.* (2016), have examined the factors that influence the development and/or usability of EDSS. However, appropriate measurements and evaluations for EDSS usability, as well as standardized techniques, have not been extensively studied, applied, or developed for most of the EDSS.

For example, in Slocum *et al.* (2003), the authors used a combination of cognitive walkthrough methods, think aloud protocols, and pluralistic inspection while developing software intended to visualize the uncertainties of global water balance. The cognitive walkthrough methods allow developers and human factors specialists to evaluate the steps of tasks-scenarios. The think aloud protocol permits users to verbalize their thought processes as they perform the tasks, while pluralistic inspection allows users with different expertise and backgrounds to contribute towards the improvement of the system. However, such approaches do not present quantitative results, but rather qualitative comments centered on the expertise of each of the tested groups. Similarly, Lavoie *et al.* (2015) used think aloud protocols and recorded interviews to determine the usability of their ATES tool (Aménagement du territoire et eau souterraine, or land planning and groundwater). The usability tests were performed for three different users, who were all associated with the land planning field.

Other studies by Jankowski *et al.* (2006) and Nyerges *et al.* (2006) evaluated usability of a collaborative spatial EDSS that sought to create a consensus for solutions/options for surface water allocation and groundwater pumping rates, using qualitative and quantitative methods. The qualitative methods involved non-standard usability questionnaire data, while quantitative methods provided information on task times for different activities, but they did not consider any measurement of the GUI's usability. Jankowski *et al.* (2006) and Nyerges *et al.* (2006) provided communication results of two different groups (of ten stakeholders each) that collaborated together to propose scenarios of regulation laws for water allocation and groundwater pumping rates. However, their study examined the analysis and evaluation of collaborative interaction rather than the evaluation of EDSS based on individual stakeholder's interaction and user's behavior. Despite these advances, there are still gaps in the use of quantitative evaluation methods to determine the usability of such participatory EDSS.

Recently, Tullis & Albert (2013) outlined quantitative measures such as task time, number of mouse clicks, and percentage of time spent in specific areas of interest (AOIs) of a web interface, as potential approaches that researchers can use for quantifying the usability of software and webpages. In this study, we used these metrics to evaluate a new, web-based participatory EDSS called WRESTORE (watershed restoration using spatio-temporal optimization of resources; http://wrestore.iupui.edu/). WRESTORE has been recently developed with the goal of allowing stakeholders, policy-makers, and planners to participate in the design of a spatially distributed system of BMPs in a watershed. WRESTORE uses a modification of the IGAMII (interactive genetic algorithm with mixed initiative interactions – Babbar-Sebens & Misnker 2011) algorithm to engage users in a human-guided search (HS) process that is targeted towards identification of user-preferred alternatives. In WRESTORE, users are shown multiple design alternatives and are asked to provide a qualitative rating of the candidate designs based on their preferences and subjective criteria. This user rating is then used as an additional objective function in the search algorithm to search for similar or better alternatives that would have design features liked by the user. In summary, WRESTORE optimizes five objective functions: four objective functions are related to the physical performance of the BMPs within the watershed (specifically, maximize peak flow reduction, maximize reduction in sediment load, maximize reduction in nitrate load, and minimize economic cost), and one objective function (maximize user ratings) is related to the user's qualitative rating of the newly found design alternatives. Babbar-Sebens *et al.* (2015) have extensively explained how the user ratings for design alternatives are obtained via the GUI. Because these incoming user ratings are used in real time by the underlying optimization algorithm to search for new solutions/design alternatives that agree with user's personal qualitative preferences (such as land management, biases towards practices and their implementation, individual local constraints and needs, etc.), evaluating the usability of the GUI is critical for interactive optimization algorithm performance. In addition to the user ratings, the interface is also used to collect confidence levels from the users. The self-reported confidence levels indicate how confident a participant was regarding his/her own rating of a candidate design alternative (Babbar-Sebens & Misnker 2011). The confidence levels, along with the user's interaction, offer potential insights into the quality of the user's input. This information is also valuable for assessing the reliability of a user-guided search.

Usability testing was conducted for WRESTORE, developed by Babbar-Sebens *et al.* (2015), and this article summarizes the findings on the nature of user behavior observed for two different types of users (stakeholders and surrogates) who interacted with the search tool's GUI. In the section immediately below, we present the objectives and research hypotheses investigated in this work. The next section includes a description of the methodology and the experimental design, followed by two sections that contain the analysis and discussion of the results. Finally, we present some concluding thoughts, future work and recommendations that may address the usability testing of WRESTORE and similar EDSS.

OBJECTIVES

The aim of the study was to use an observational approach to determine how participants used the GUI of the WRESTORE EDSS, as they gathered information and

made decisions in order to rate different design alternatives. This kind of system, that includes direct participation of stakeholders in the optimization process, is relatively new in the field of watershed planning and management, and also in the field of EDSS. As mentioned earlier, since the GUI is the primary mechanism to collect the user ratings from end users, it is important to determine if it can support the necessary functions for a user to be able to easily: (a) gather information about the candidate design alternatives, (b) conduct comparisons between candidate design alternatives, and (c) provide meaningful feedback with improved confidence in his/her own evaluation of candidate design alternatives. Analyses of the user's data will help to better characterize his/her interaction with the tool, enabling improvements in the efficiency of the underlying search algorithm used by WRESTORE (i.e., IGAMII). We tested the tool's performance with two groups: stakeholders (watershed end users) and surrogates (non-stakeholder volunteers) because surrogate users' data are often used for DSS prototype development of interfaces due to time and cost constraints. Surrogate testing is an essential aspect of DSS development and allows designers to get any major usability problems solved prior to potentially wasting stakeholders' time and/or losing them in the participatory process. However, as a result, it becomes critical to know and understand the extent to which such volunteers can be used as proxies for future end users.

The study examined the following research hypotheses:

H1. Over time, users should become more efficient in using WRESTORE's GUI. As users learn how to navigate and use the GUI's features, overall task times should decrease across repeated sessions. This decrease should follow the theoretical learning curves (Yelle 1979; Newell & Rosenbloom 1981; Estes 1994), specifically, DeJong's learning formula (Jaber 2011).

H2. As stakeholders are directly affected by the issues and implementation decisions related to the watershed, we expect that they will use the tool more effectively, and spend a greater percentage of their time utilizing the information gathering areas of the interface than surrogate users will.

H3. Similarly, we expect that stakeholders will focus their attention on the more informative areas of the interface and therefore have a higher percentage of mouse clicks than surrogates in the information gathering areas of the web-interface.

H4. As users gain experience by interacting with the tool and develop a better understanding of the performances of different design alternatives, we expect that their overall self-reported confidence levels will increase over time, resulting in a positive trend.

H5. A comparison among confidence level trends, time spent, and mouse clicking events should show that when users spend more time and make more mouse clicks in information gathering areas, their confidence levels increase over time.

A final goal was to set a basis for evaluating the interactions between human factors and GUIs in participatory decision support tools. Such protocols will allow tool developers to test the usability of these systems for their user community, determine what improvements should be made based on specific user populations, and learn how to facilitate the participation of stakeholders in similar watershed design tools.

METHODOLOGY

Case study site

The web-based WRESTORE tool was developed by Babbar-Sebens *et al.* (2015) to enable watershed communities to engage in participatory design efforts and influence the spatial design of conservation practices (specifically, filter strips and cover crops in the experiments reported for this paper) on their landscape. WRESTORE software is currently being tested at the study site of Eagle Creek Watershed, Indiana, where multiple researchers (Tedesco *et al.* 2005; Babbar-Sebens *et al.* 2013; Piemonti *et al.* 2013) have conducted watershed investigations. A calibrated watershed model of the study site, based on the Soil and Water Assessment Tool (SWAT) (Arnold *et al.* 2001, 2005; Neitsch *et al.* 2005), was used to simulate the impact of a candidate plan of practices on the watershed and calculate values for objective functions for the period 2005–2008. The watershed was divided into 130 sub-basins, out of which 108 were suitable for the proposed practices. Values

of design variables indicated whether and how a practice was implemented in one of the suitable sub-basins. Based on the values of the design variables found collaboratively by the user and the micro-IGA, SWAT model parameters were modified to represent the design decisions.

Web-tool WRESTORE evaluation

Participants

Twenty-three participants volunteered for the study. We divided them into two groups based on their affiliation with the watershed: stakeholders and surrogates. Each group was treated as a predictor variable. The stakeholders group contained eight stakeholders from the Eagle Creek watershed who work at federal and state agencies and non-governmental organizations in programs that support the implementation of conservation practices in the watershed. The surrogates group contained 15 non-stakeholder volunteers with science and engineering backgrounds, who were not directly involved in the watershed.

Three participants' data sets were excluded from the analysis. One participant (a stakeholder) quit the study before finishing, leaving an incomplete set of answers. The other two participants (a stakeholder and a surrogate user) failed to follow the instructions. Therefore, 20 participants' data sets were analyzed (six stakeholders – five males, one female; 14 surrogates – six males, eight females).

We also want to note that for seven participants in the surrogates group, the tool used a different underlying hydrology model (used by the optimization algorithm to calculate the four physical objective functions). However, since the GUI features were identical for all participants, it can be reasonably assumed that the underlying differences in the physics of the hydrological model did not affect the user–interface interaction and experience.

Design procedure

While the study was designed to test user interaction with the tool and its ability to generate designs that agreed with the user's subjective criteria, this article only reports the findings on the user–tool interaction. Babbar-Sebens *et al.* (2015) provide a detailed explanation of the WRESTORE tool and its performance in multiple user experiments.

The experiment began with a registration process, and selection of two conservation practices (cover crops and filter strips) and four physical goals (maximize peak flow reduction, minimize economic cost, maximize sediment reduction, and maximize nitrate reduction). The selection of specific conservation practices and goals was controlled and prescribed by the researchers in order to preserve homogeneity across the different participants. After the registration, user-instructions on how to perform the test were provided via in-person meetings and/or workshops, and also displayed on a pop-up window of the GUI's main feedback interface. Once the participant had read through the instructions, he/she closed the pop-up window to transition to the main design and feedback interface (Figure 1). In this interface, the participants could:

1. View the spatial distribution of the conservation practice(s) using the alternative maps (component (iii) in Figure 1). These maps indicate the location where the practice(s) should be implemented.
2. Select which conservation practice and its distribution were shown in the map (using the legend – component (ii) in Figure 1).
3. Compare the effectiveness of the design via the four different goals estimated using the underlying cost and hydrologic models (component (v) in Figure 1).
4. Provide a user rating for each design via a Likert-type scale with three options (I like it, Neutral, and I do not like it) – component (iv) in Figure 1.
5. State the level of confidence (using the scale bar in component (iv) in Figure 1) in their personal ratings of each design.
6. Record their answers and move to the next pages to view additional design alternatives generated by the tool (component (vii) in Figure 1).
7. Submit their ratings of the design alternatives to the underlying interactive optimization algorithm to generate a new set of design alternatives for the next session (component (vi) in Figure 1).

To measure interface interaction, we tracked the time spent and number of mouse clicks in three main AOIs.

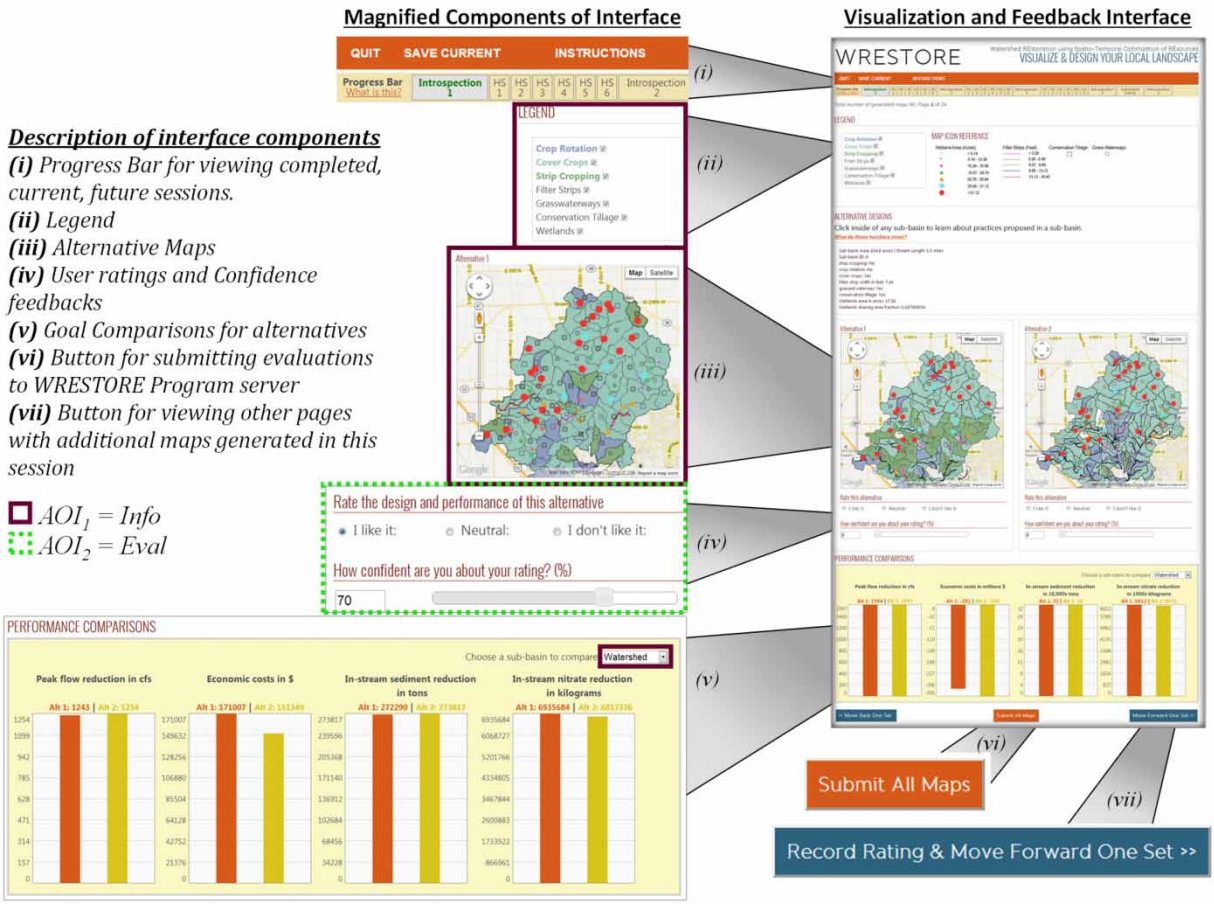

Figure 1 | Components of WRESTORE tool's main design and feedback interface.

The first AOI was associated with information gathering ($AOI_1 = Info$) and included user interactions with the legend (component (ii)), the alternative maps (component (iii)) and/or the drop-down (sub-drop in component (v)) menu. The drop-down menu allowed users to compare the performance of two design alternatives at watershed scale or at a local sub-basin scale (component (v)). The second AOI was associated with the evaluation process ($AOI_2 = Eval$), and included user interactions with the Likert-scale user ratings and the confidence sliders (component (iv)). The final AOI ($AOI_3 = Other$) included all other user activity outside the areas covered by AOI_1 and AOI_2, i.e., if the participant clicked in an area outside of the solid (Info) and dashed (Eval) boxes shown in Figure 1.

Users' interactions with these AOIs were tracked across all WRESTORE sessions. Each session had, on average, 20 different designs that participants were required to examine and evaluate. The progress bar on top of the webpage (Figure 1, component (i)) shows the participant's progress and how many more sessions were left to complete. Participants were asked to complete all of the sessions through Introspection 3 (I3), although they had the option to continue beyond I3 if they desired.

The participants went through two types of sessions: Introspection (I) sessions and HS sessions. Introspection session 1 (I1) was the starting point for all participants. This session showed 20 designs selected from a pool of non-dominated design alternatives found by a multi-objective genetic algorithm that used only the four physical objective functions to guide the search process (Babbar-Sebens *et al.* 2015). In this session, all participants saw the same 20 designs. After I1, participants transitioned into HS sessions (i.e., six HS sessions – HS1_1 to HS6_1) that had one-to-one correspondence with the six iterations

(or generations) of a micro-interactive genetic algorithm (micro-IGA; Babbar-Sebens & Misnker 2011; Babbar-Sebens *et al.* 2015). In each of these HS sessions, 20 new designs (population of micro-IGA), generated by the micro-IGA, were presented to the user. After the micro-IGA was over (i.e., at the end of HS6_1), designs that had high user ratings and/or were on the best non-dominated front were saved in to a case-based memory (CBM). Next, the participant revisited 20 random designs from the CBM in Introspection session 2 (I2), where he/she was provided the chance to modify his/her previous user ratings and confidence levels based on any newly formed preferences and perceptions. This revisiting technique aimed to improve the participants' skills in evaluating the performance of the design alternatives using the information gathered in the HS sessions. If the user modified his/her evaluation of the revisited design in the Introspection session, the changes were updated in the CBM. At the end of I2, a sub-set of designs from the CBM were again inserted in the initial population of the next micro-IGA, and a new set of six HS1_2 to HS6_2 sessions ensued. It is worth noting here that the WRESTORE interactive search tool was originally developed with an additional set of alternating I and HS sessions and an Automated Search session, all occurring after I3. The Automated Search session uses a machine learning model to learn from the participant's user ratings collected in the earlier HS sessions, and then conducts an exhaustive search on behalf of the participant by using the machine learning model as a simulated user. Interested readers are encouraged to refer to Babbar-Sebens *et al.* (2015) for details on these sessions. In order to limit human fatigue and keep the workload the same for all users, researchers requested that participants complete all alternating HS and I sessions, at least until the end of I3. Several participants in the surrogates group, however, did choose to complete sessions beyond I3. Nevertheless, these additional sessions were not considered as part of the primary analyses conducted to examine the hypotheses above.

After the participant had submitted his/her user ratings for design alternatives, 10–15 minutes of waiting time was needed for the underlying optimization algorithm and hydrologic model to generate and evaluate new design alternatives. To allow participants to complete the experiment at their own pace, an email notification system was used to send them an automated reminder whenever the next session was available for user interaction. The automated e-mail included a log-in link to the WRESTORE webpage, and when users clicked on the link to log into their account, a set of instructions on how to conduct the ongoing experiment was redisplayed on the screen (Babbar-Sebens *et al.* 2015).

MEASURES OF USER INTERACTION WITH THE INTERFACE

This section introduces the different usability metrics, or response variables that were employed to evaluate the user's interaction and usability of the WRESTORE tool.

Overall task times

We assessed the participant's ability to navigate and learn the tool by evaluating the mean task times across successive 'I' sessions and also across successive 'HS' sessions, during which design alternatives were presented to the user via the GUI. Task times for each session were recorded and used to infer how quickly participants learned to use the tool interface efficiently. When the participants were not using the tool (e.g., taking a break), they were instructed to press the 'Save all' button, so that we could consider these off-task time intervals as outliers and exclude them from the task time analyses. These events were saved in a database as 'Save all maps', and removed during the post-processing analysis, along with 'Quit' events. However, there were some occasions when some participants failed to click the 'Quit' or 'Save all' buttons, resulting in excessively long task times. To remove these outliers, we excluded the task time values that were greater than two standard deviations from the mean task time across all considered sessions, for each participant.

Mean percentage of time spent in different AOIs

In order to determine where users were focusing their attention, we compared time spent in each AOI (Info, Eval, and Other described in the section 'Web-tool WRESTORE evaluation' and sub-section 'Participants'). The percentage of time spent ($pts_{i,j,m}$) was calculated using:

$$pts_{i,j,m} = \frac{\sum_{h=1}^{H}\sum_{k=1}^{L}\Delta t_{i,j,m,k,h}}{\sum_{h=1}^{H} ttotal_{j,m,h}} * 100 \quad (1)$$

where $pts_{i,j,m}$ is the percentage of time spent, i is an index that goes from one to three and represents each AOI, j is an index that goes from one to five and represents I or HS sessions. Data from HS1_1 to HS6_2 sessions were grouped together into blocks of HS sessions in order to conduct a temporal analysis that was based on alternating session types. Therefore, if j is an even number it represents an Introspection session and the variable H is equal to one (because Introspection sessions are not evaluated in blocks), but if j is an odd number it represents a HS session and H is equal to six (representing each of the iterations in the micro-IGA). The index for each participant is represented by m. The variable $\Delta t_{i,j,k}$ is the interval of time between two events (k), L is the total number of events associated with the AOI_i, and $ttotal_{j,m}$ is the participant's (m) total time in each session (j).

The mean time spent was calculated by adding the percentage of time spent ($pts_{i,j,m}$) per AOI_i per participant (m). This mean was grouped by session type and by group. Therefore, to calculate the mean percentage of time spent by AOI we used:

$$MPTS_{i,j} = \frac{\sum_{m=1}^{N} pts_{i,j,m}}{N} \quad (2)$$

where $MPTS_{i,j}$ is the percentage of time spent in each AOI_i and N is the total number of participants in the group.

Mean percentage of mouse clicking events by AOI

The same AOIs described in the section 'Web-tool WRESTORE evaluation' and sub-section 'Participants' (Info, Eval, and Other) were used to track the mouse clicking events, and monitor user interactions with different areas of the interface. As the total number of clicking events varied between participants, the percentage of clicking events in each session was calculated for each participant, and for each group (surrogates and stakeholders).

For each participant, we used the following general formula:

$$PMC_{i,j,m} = \frac{\sum_{h=1}^{H} NC_{i,j,m,h}}{\sum_{h=1}^{H} TC_{j,m,h}} * 100 \quad (3)$$

where $PMC_{i,j,m}$ is the percentage of mouse clicks per participant per AOI, $NC_{i,j,m}$ is the number of clicks per i^{th} AOI, per j^{th} session, per m^{th} participant, and $TC_{j,m,h}$ is the total number of clicks per j^{th} session per m^{th} participant. As in the mean percentage of time spent, H will vary with the session type. Therefore, for j equal to an even number (representing I sessions), the variable H is equal to one, and for j equal to an odd number (representing HS sessions) the variable H is equal to six. This allowed us to compare the clicking interactions of individuals within each group.

The percentage for each group was calculated using the mean value of all the participants within the group. Therefore:

$$PG_{i,j} = \frac{\sum_{m=1}^{N} PP_{i,j,m}}{N} \quad (4)$$

where $PG_{i,j}$ is the percentage per group per i^{th} AOI, and N is the total number of participants per group.

Confidence levels

Confidence level indicates how confident the participant felt about his/her own user rating (I like it, Neutral, or I do not like it). User's confidence in his/her user ratings were indicated via the confidence level slider bar (component (iv) in Figure 1) that ranged in its scale from 0 to 100, and could be modified by the user during the session. However, changes in the confidence levels for the same designs that appeared multiple times during the experiment could not be tracked, as new changed values replaced previous values in the archive database.

We classified participants by confidence level trends in the following manner. First, we conducted a nonparametric Mann–Kendall hypothesis test (Helsel & Hirsch 2002) via a Matlab script (Burkey 2006) to assess whether the trends in average values of confidence levels (estimated from data on

confidence levels in each session) were monotonically increasing or decreasing. This test indicated whether or not participants presented a trend, at a significance level alpha of 0.1, and the Sen's slope (S value) determined if the trend was positive or negative. Since the main focus of this test was to minimize the risk of not detecting an existing trend (i.e., Type II error), a larger alpha value was chosen. Participants were thus separated into three confidence level trend groups (positive, negative, and no trend). Finally, we attempted to identify similarities and differences between participants who showed positive, negative, and no trend in confidence levels, and relate them to their interface behavior (i.e., time spent and mouse clicks) concerned with information gathering.

Relationships between confidence levels, time spent, and mouse clicking events

We fitted trend curves to usability data (i.e., time spent and number of mouse clicks) in order to determine any underlying patterns and relationships for participants within each confidence level trend (positive, negative, or no). We also compared the differences in these trends for Info and Eval type events. To select the best trend curve model, we used a combination of the coefficient of determination, and three versions of the Akaike information criterion (*AIC*).

The *AIC* provides the relative quality of a proposed model in a given data set, dealing with the goodness of fit and the complexity of the model. A lower *AIC* value indicates higher preference for a model. Three approaches for calculating AIC were used, based on the following equations:

$$AIC = -2(log(L)) + 2K \tag{5}$$

$$AIC_{Res} = n(log(\sigma^2)) + 2K \tag{6}$$

$$AIC_C = -2(log(L)) + 2K + \frac{2K(K+1)}{(n-K-1)} \tag{7}$$

where L is the likelihood of the model, K is the number of parameters, n is the sample size, AIC is the standard equation for the Akaike's information criterion, AIC_{Res} is based on the least square regression (assuming normal distribution), and AIC_C is the second order AIC that includes a penalization for small sample sizes (cited after Mazerolle 2004).

RESULTS

In this section, we present the results from the data analysis for the task time, percentage of time spent in each AOI, percentage of mouse clicks in each AOI, and confidence level trends, as well as the relationships between these variables.

Overall task times

We first assessed participants' mean task times for each session. The mean for each group (surrogates and stakeholders) was calculated and the results were separated according to the type of sessions (i.e., I or HS). Results for surrogates showed that in earlier sessions (for both I and HS) the mean task time and the standard errors were greater than in later sessions. Similar results were found for stakeholders in the I sessions. However, stakeholders' mean task times for HS sessions were somewhat more variable.

Figure 2 presents the mean task times for surrogates and stakeholders for all completed I sessions, where the main task of participants was to evaluate initial designs and/or

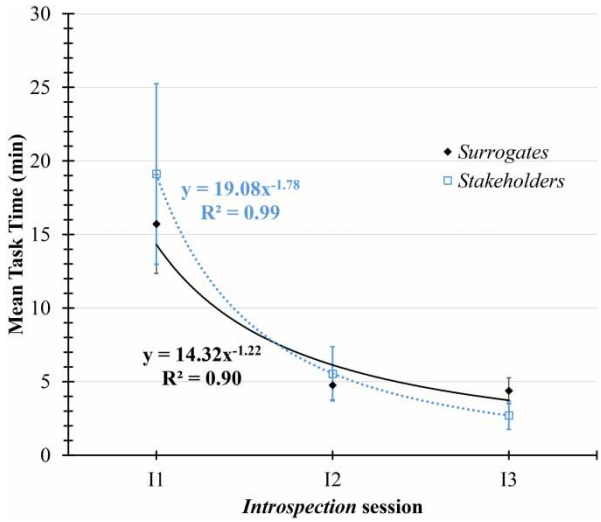

Figure 2 | Power function representing the learning curves of the mean task times for the surrogates and stakeholders groups across I sessions. Error bars represent the standard error of the mean.

re-evaluate their previous user ratings. Despite the limited number of Introspection sessions, both groups show a good estimation of the theoretical power learning curve typically used to represent learning processes (Yelle 1979; Newell & Rosenbloom 1981; Estes 1994; Jaber 2011), with a coefficient of determination >0.9 for both groups. However, we observed that the stakeholders' learning curve started at a higher value of mean task time than that of the surrogates. Stakeholders' mean task times decreased by 85% from I1 to I3, while surrogates' mean task times decreased by 74% from I1 to I3. This effect shows that the task might have seemed more challenging for the stakeholders in the first I session when they were still getting used to the interface, but as time progressed, their mean task time decreased to a value lower than the surrogates', in the last I session. Variability also decreased over time for both groups.

We also compared the task times of HS sessions for the two groups (Figure 3), where the main task of participants was to compare and evaluate newly generated design alternatives, and provide feedback on these designs to help guide the search algorithm in creating new design alternatives for the next HS session. Figure 3 shows two different learning curves for surrogates and stakeholders. Surrogates showed continuous learning with a decrease in average time across HS sessions. Surrogates' behavior fit the power learning curve with a coefficient of determination of $R^2 = 0.87$. The mean task times for stakeholders, although generally higher than surrogates, poorly fit the theoretical power learning curve, resulting in a coefficient of determination of $R^2 = 0.23$. This fit is substantially lower than the R^2 value previously obtained for the stakeholders' I sessions, primarily due to the higher variability of the mean task times within HS sessions. On average, the standard error for stakeholders in the HS sessions was 62% larger than that for surrogates.

Mean percentage of time spent in different AOIs

To determine which areas of the interface were capturing users' attention and encouraging interaction, we analyzed the mean percentage of time spent within the different AOIs for each group. Figure 4 shows a pie chart table that compares surrogates and stakeholders for all sessions. On average, stakeholders spent 15% more time on information gathering than surrogates, while surrogates spent more time in the Eval and Other AOIs (8% and 7%, respectively, greater than stakeholders).

Table 1 summarizes the data in Figure 4, and presents the overall mean percentages of time spent (averaged across sessions) and 95% confidence intervals (CIs) for surrogates and stakeholders in each AOI.

Mean percentage of clicking events in different AOIs

To determine which areas of the interface were capturing users' attention and encouraging interaction, we also analyzed

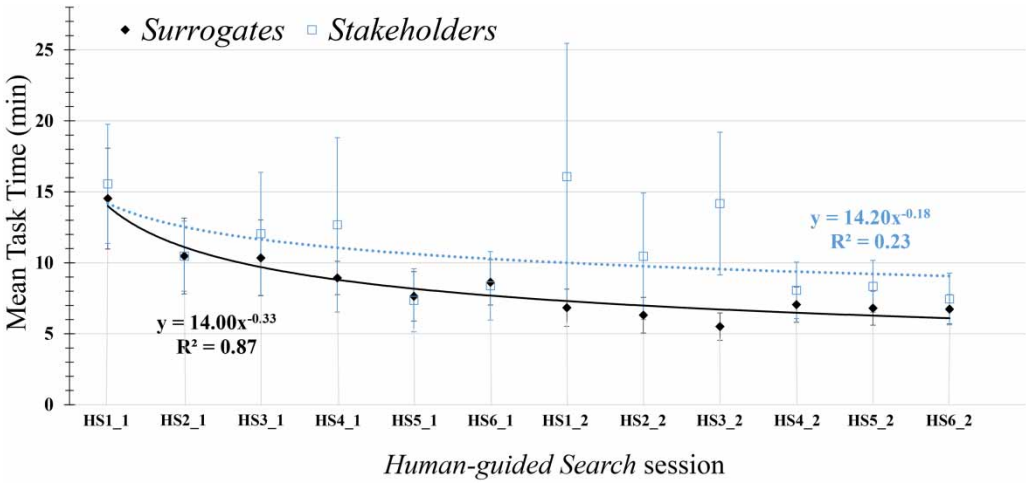

Figure 3 | Power function representing the learning curves of the mean task times for the surrogates and stakeholders groups across HS sessions. Error bars represent the standard error of the mean.

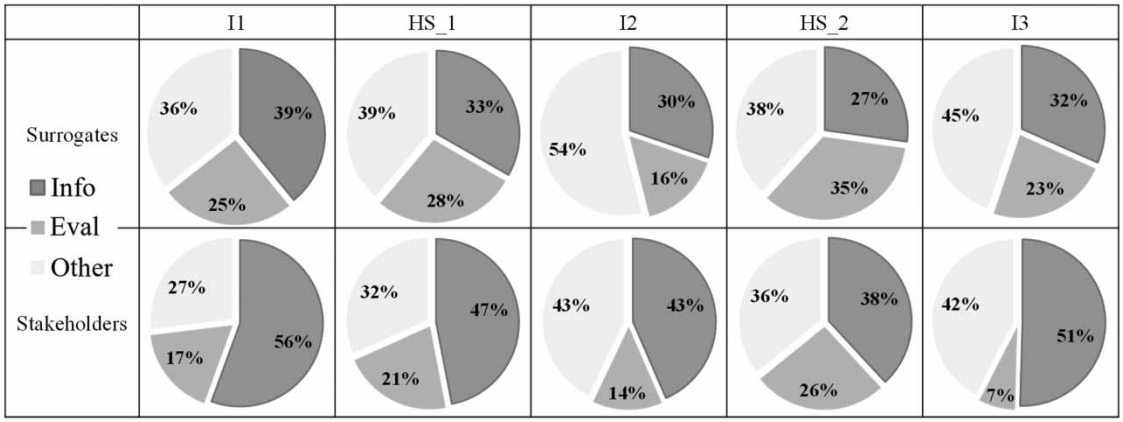

Figure 4 | Mean percentage of time spent in each AOI for surrogates and stakeholders. Refer to the section 'Web-tool WRESTORE evaluation' and sub-section 'Participants' and Figure 1 for descriptions of each AOI.

the mean percentage of clicking events in different AOIs. Similarly to Figure 4, Figure 5 displays a summary pie chart table with the mean percentages of mouse clicks in the Info, Eval, and Other AOIs for the two groups across the different types of sessions. On average, stakeholders clicked 14% more in Info areas than surrogates, while surrogates had a greater percentage of clicks in the Eval and Other AOIs in comparison to stakeholders. The percentage of clicks in the Eval AOI was 11% greater for the surrogates group, and the percentage of clicks in the Other AOI was 3% greater for surrogates.

We calculated 95% CIs for overall mean mouse clicking events, similar to the analyses in Table 1 for time spent. Table 2 presents the overall mean percentages of mouse clicks (averaged across sessions) for surrogates and stakeholders in each AOI.

Confidence levels

For each participant, and for each of the rating classes (i.e., I like it, Neutral, and I do not like it), a Mann–Kendall trend test was performed to identify if there were monotonic trends in the mean confidence levels across consecutive sessions. The results were separated according to positive, negative, or no trends, based on the results over time. A positive trend indicated that users were becoming more confident over time about the ratings they provided for the designs. On the other hand, a negative trend indicated that they were becoming less confident over time. In summary, 40% of all of the participants showed a positive trend in at least one of the rating classes. For the stakeholders group, 67% of the participants showed a positive trend, while just 29% of the participants in the surrogates group showed a positive trend. It is, however, important to mention here that these trends were calculated for the sessions that lasted until the end of I3. It is possible that if participants continued beyond I3 then they could change their trends, especially if the additional engagement during the interactive search process improved their reasoning process and led to a change in their confidence levels.

Table 1 | CIs for the mean percentages of time spent in each AOI across all sessions by group

AOI	Group	Mean	95% CI
Info	Surrogates	32	[25, 39]
	Stakeholders	47	[36, 57]
Eval	Surrogates	25	[21, 30]
	Stakeholders	17	[13, 22]
Other	Surrogates	42	[36, 48]
	Stakeholders	36	[27, 45]

Relationships between confidence levels, time spent and mouse clicking events

We also associated the mean percentage of mouse clicks for information gathering and the mean percentage of time spent for information gathering (Table 3) with each of the trends in mean confidence levels reported in the previous sub-section. Results showed that participants with a positive trend in mean confidence levels also had 13% more mouse

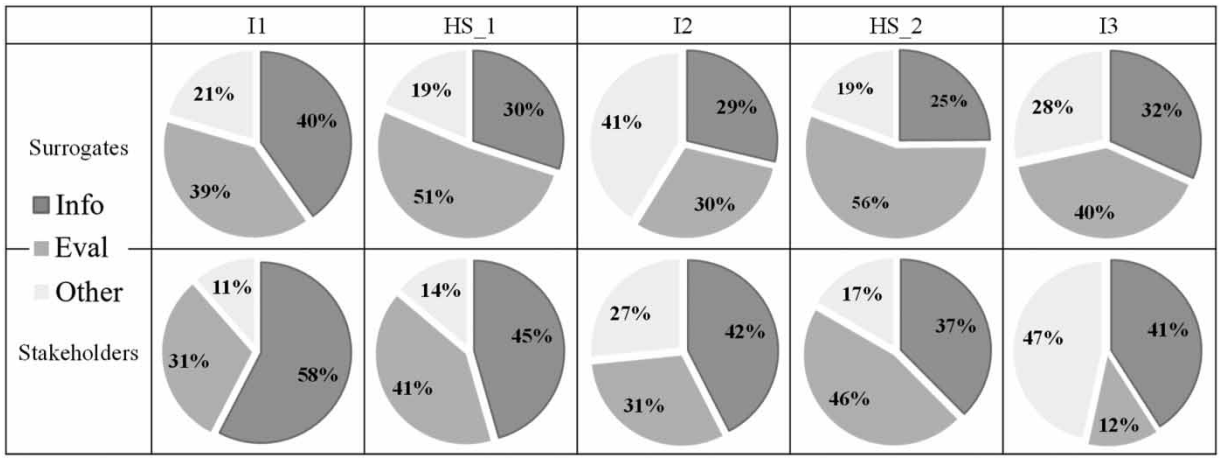

Figure 5 | Mean percentage of mouse clicking events for surrogates and stakeholders within each AOI. Refer to the section 'Web-tool WRESTORE evaluation' and sub-section 'Participants' and Figure 1 for descriptions of each AOI.

clicks than participants with a negative trend, and 9% more mouse clicks than participants with no trend in AOIs related to information gathering. Participants with a positive trend in mean confidence levels spent 1% more time than participants with a negative trend, and 12% more time than participants with no trend in AOIs related to information gathering.

As was suggested earlier, it is possible for participants to experience changes in the trends of the means of their confidence levels, during the course of their interaction with the tool. Therefore, we identified participants who had interacted with the WRESTORE tool beyond the I3 session, and re-evaluated the trends in their interaction data by including the data from the additional sessions after I3. Figure 6 shows four plots that relate time spent vs. number of mouse clicking events for each participant in the Info and Eval AOI. Figure 6(a) and 6(b) show the relationships for the data collected from sessions completed by all participants until the end of I3. Figure 6(c) and 6(d) show the results for all of the available data from each participant, which includes data from additional sessions for participants who progressed beyond I3. It can be seen that while the classification of trends in mean confidence levels remained the same for the majority of participants (see Figure 6(a) and 6(c)), the trend for Participant 2, however, changed from no trend (for sessions from I1 to I3) to positive trend (for sessions from I1 to I5). This indicates that when Participant 2 engaged actively with the tool longer than I3, she/he was eventually able to improve her/his confidence in the user ratings provided during the experiment. Results also showed that, for the Eval AOI (Figure 6(b) and 6(d)), there is no clear separation between the responses for each trend, irrespective of how long the experiment lasted.

Table 4 shows the comparison of the R^2 and three different AIC values obtained for four different approximations, generated using the results of the sessions

Table 2 | CIs for the mean percentages of mouse clicking events within each AOI across all sessions by group

AOI	Group	Mean	95% CI
Info	Surrogates	31	[24, 38]
	Stakeholders	45	[34, 56]
Eval	Surrogates	43	[38, 49]
	Stakeholders	33	[24, 42]
Other	Surrogates	26	[21, 30]
	Stakeholders	22	[14, 30]

Table 3 | Classification of confidence levels, mean percentage of clicking events, and mean percentage of time spent across participants with the same trend

Trend	% of total participants	Mean % of clicking events in Info AOI	Mean % of time spent in Info AOI
Positive	40	42	41
Negative	30	29	40
No	30	33	29

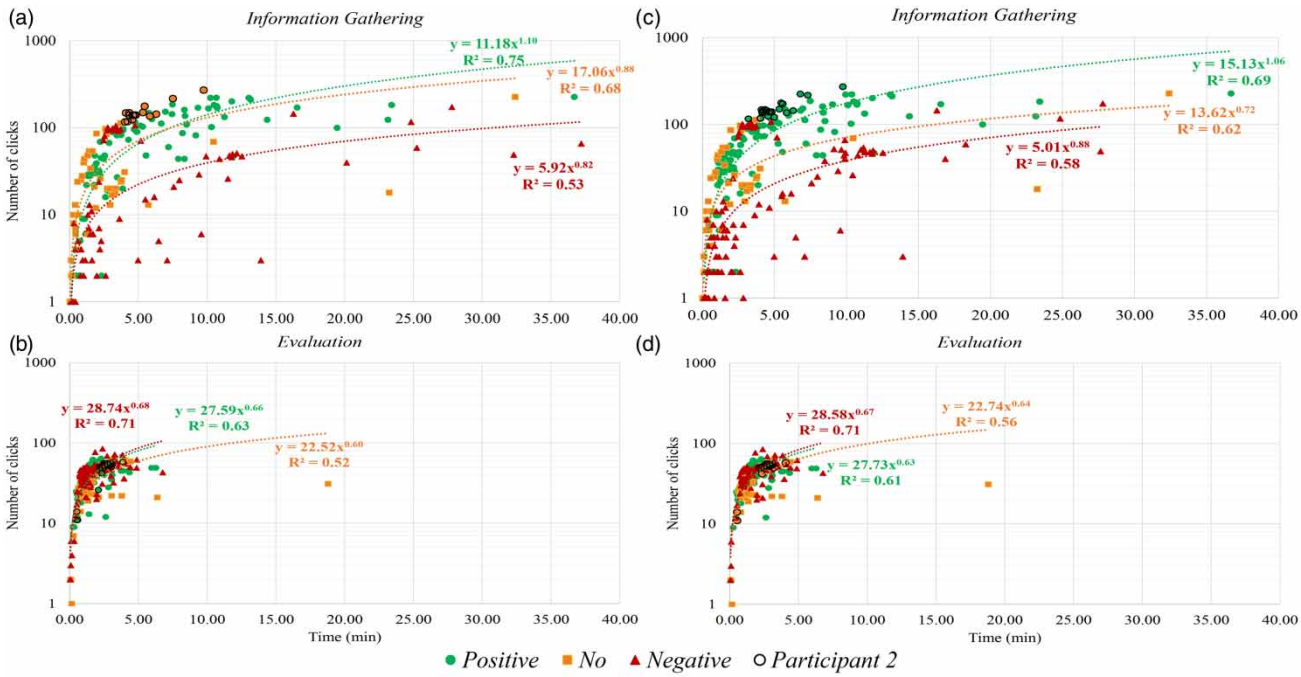

Figure 6 | Time spent vs. number of clicks per participant per trend, for Info (upper) and Eval (lower) AOIs. Parts (a) and (b) show the results for sessions I1 to I3. Parts (c) and (d) show the results for all completed sessions. The black bordered circles show the data for sessions of Participant 2, who had a 'No' trend until I3, but later ended up having a 'Positive' trend after eight additional sessions that she/he volunteered to complete.

Table 4 | Coefficient of determination and *AIC* values for the different tested models

		Sessions I1 to I3				All available sessions			
Trend	**Model**	R^2	*AIC*	AIC_{Res}	AIC_c	R^2	*AIC*	AIC_{Res}	AIC_c
Positive	Linear	0.38	−1.64	302.55	−1.59	0.17	−1.65	478.38	−1.62
	Quadratic	0.67	0.18	281.16	0.32	0.57	0.14	442.11	0.23
	Third	0.73	2.11	277.28	2.39	0.67	2.06	430.93	2.24
	Power	0.75	−1.85	321.37	−1.80	0.69	−1.89	502.60	−1.86
No	Linear	0.06	−1.13	219.08	−1.06	0.29	−0.44	145.04	−0.35
	Quadratic	0.45	0.46	207.22	0.67	0.31	1.42	146.48	1.69
	Third	0.60	2.40	205.28	2.83	0.66	2.99	149.38	3.54
	Power	0.68	−1.38	215.91	−1.31	0.62	−0.77	141.38	−0.68
Negative	Linear	2.35E-3	−0.69	242.36	−0.64	0.22	−0.71	325.21	−0.67
	Quadratic	0.14	1.07	239.52	1.24	0.25	1.19	325.49	1.30
	Third	0.17	2.83	250.73	3.17	0.35	3.08	321.65	3.31
	Power	0.53	−0.73	239.70	−0.68	0.58	−0.55	325.90	−0.52

The shaded boxes indicate the preferred model for a specific model selection criterion.

completed by all the participants. These R^2 and *AIC* values indicate that a power approximation represents the best-fit model for all of the trends, except for the negative trend in 'All available sessions' where a linear function seems to have a better fit.

DISCUSSION

In this paper, we analyzed usability metrics for a novel, web-based tool, WRESTORE, which supports decision-makers in the search and design of alternatives for allocating

conservation practices in agricultural watersheds. This kind of tool is relatively new in the field of watershed planning, necessitating protocols that will enable developers to assess what types of improvements should be made, based on quantitative user feedback. Furthermore, because surrogates are often a necessary component of prototype development due to time and cost constraints, it is critical to know to what extent their data can be generalized to the stakeholder population of interest.

Four main usability metrics were considered for the analysis: task time evaluation, percentage of time spent in different AOIs, percentage of mouse clicking events in different AOIs, and trends in mean confidence levels.

Overall task times

We evaluated overall task times in order to assess participants' efficiency in using the tool interface. Overall task times decreased across repeated sessions following a power learning curve. The results are consistent with hypothesis **H1**, where it was stated that over time, users should become more efficient in using the interface as they learn how to navigate and use the different features. I sessions followed a power learning curve for both surrogates and stakeholders, but stakeholders showed a greater decrease in overall task time for I sessions than surrogates.

HS sessions also followed a power learning curve for the surrogates. However, the mean task times for stakeholders were much more variable, resulting in a poor fit to the expected learning curve. The differences in mean task times across stakeholders may be due to the following potential reasons: (1) some abnormally large task times that were still within two standard deviations from the mean and hence were not excluded as outliers, possibly related to a re-learning process due to the lag time between sessions, (2) small sample size of stakeholders, or (3) existence of two learning curves instead of one overall learning curve as seen by an apparent increase in the time spent again at the start of the second block of HS sessions (i.e., HS1_2 in Figure 3).

Overall task times can provide estimates of the tool's efficiency to developers. Information on how fast the users learn via the GUI can be expected to assist the tool designers in the re-evaluation of the users' interactions during the sessions, and of the tool's ability to use participation in guiding the search process. Further, an insightful understanding of time needed to complete goals and provide useful feedback is also important for eliminating extra sessions, which could increase user fatigue and introduce noise without meaningful benefit to the search algorithm.

Mean percentage of time spent in different AOIs

The analysis of these data provided us with insightful evidence about where users were focusing their attention, as measured by the percentage of time each group spent in each of the AOIs. The groups showed fairly consistent behavior across sessions. However, stakeholders expended more time in the Info AOI, while surrogates expended more time in Eval and Other AOIs. This supports hypothesis **H2**, where we predicted that because the stakeholders are directly affected by issues and actions in the watershed, they will use the tool more effectively and their percentage of time spent in information gathering areas of the interface would be greater than for the surrogates.

Mean percentage of clicking events in different AOIs

These results help to understand how different groups were using the interface to perform the tasks. The higher percentage of mouse clicking events for stakeholders vs. surrogates in Info AOIs supports hypothesis **H3** and is consistent with the results reported above for mean percentage of time spent. Surrogates and stakeholders behave differently regarding information gathering. On average, the majority of stakeholders tend to make more mouse clicks to gather information from the user interface in order to make their decisions. Surrogates, on the other hand, do not explore the information gathering areas of the interface as much, making fewer mouse clicks in these regions. This could be a consequence of lack of interest in the task, or a lack of information about the tool's goals.

Confidence levels

Previous work using confidence levels, showed a positive trend as experimental time progressed (Babbar-Sebens & Misnker 2011). A positive trend indicates that mean

confidence levels increase over time as users gained experience with the tool. However, we did not find that mean confidence levels increased for all of the users. Therefore, hypothesis **H4** was only partially supported.

Our results showed that just 40% of participants exhibited a positive trend in mean confidence levels over time. Nevertheless, there was a clear distinction between surrogates and stakeholders in relation to these trends. Approximately 67% of the total stakeholder participants presented a positive trend, while just 29% of the total surrogate participants presented a positive trend.

Relationships between confidence levels, time spent, and mouse clicking events

The section examines the results on the relationships between trends in mean confidence levels, time spent, and mouse clicking events. The results in Table 3 help to support hypothesis **H5** that states 'when users spend more time and make more mouse clicks in information gathering areas, their mean confidence levels increase over time'. An analysis of the relationship between time spent, clicking events, and mean confidence level trends indicates that participants with a positive trend had a larger number of clicking events per unit time spent in information gathering areas, compared to participants with a negative trend. This may indicate that if a participant interacts with the interface to gather more information on the design alternatives for the same amount of interaction clock time (i.e., the time spent), then there exists a probability for them to either improve their self-confidence over time or maintain a steady value over time. In addition to the slope, the fitted value of the exponent in power curves was also lower for the negative trend than for the positive trend in mean confidence levels. In the Eval AOI, no clear difference across confidence level trends was observed.

CONCLUSIONS AND FUTURE WORK

This research adapted and applied usability techniques to a set of data collected with the WRESTORE tool to evaluate its performance and usability. Quantitative usability evaluations of EDSS are not typically conducted for these types of EDSS systems. However, this approach can offer valuable insights into the ways that these tools will be used. Overall, this work provided two substantial contributions: (1) determination and validation of usability and metrics for participatory design tools based on information technologies and (2) evaluation of differences between how surrogates (volunteers) and stakeholders (end users) use such interactive design tools.

WRESTORE was developed for the Eagle Creek Watershed, in Indianapolis, IN, with the goal of providing a more democratic venue for stakeholders to engage with the watershed community in the design of alternatives for spatial allocation of conservation practices. The tool was initially tested by surrogates who were not intimately involved with the issues and concerns in the watershed. Their feedbacks were recorded and saved for later analysis. Then, the tool was tested with stakeholders (i.e., potential end users) to determine if the findings from surrogates held true for the actual end users.

As the majority of usability tests are performed by students or volunteers, we wanted to track possible differences in responses between the tested group (surrogates) and the end users (stakeholders), and analyze to what extent the results from surrogates would be reflected in the behaviors of stakeholders. From the overall task time analysis, we concluded that the participants of both groups became more efficient as they learned how to navigate and use the tool's features. Generally, overall task times decreased across repeated sessions. Therefore, surrogates can potentially be used as proxies for stakeholders for overall task time analysis and improvements. However, they differ in other potentially important regards, as discussed below.

As for how users focused their attention in terms of time spent across the different AOIs, results showed that stakeholders expended more time than surrogates in gathering information about the performance of the different design alternatives. This could be a result of the motivation of stakeholders in creating a designed distribution of BMPs that better suits their interests. Surrogates that were not involved with the watershed may have lacked this motivation. Similarly, as we predicted, a higher percentage of mouse clicks were made on the information gathering areas by stakeholders vs. surrogates.

We also noticed that the majority of the stakeholders showed an increase in their mean confidence levels over time, while the surrogates did not. A comparison among trends in mean confidence levels showed that positive and no trends were associated with more information gathering activity. Surrogates that did less information gathering were more likely to show a decrease in their confidence levels. As observed, for one of the participants (Participant 2), extensive information gathering over repeated sessions can also lead to a later positive change in the trend of mean confidence levels (see Figure 6).

A potential limitation of this initial investigation of the tool's usability is that it only involved a specific group of stakeholders, who work at federal and state agencies and non-governmental organizations in programs that support the implementation of conservation practices in the watershed, but future plans include experimentally testing the tool on a wider range and a greater number of stakeholders, including farmers. Additionally, we specifically chose quantitative usability metrics as this type of analysis had been previously neglected in the literature, but in future work we also plan to include follow-up questionnaires to further investigate user reasoning and evaluate user self-reports along with the quantitative data.

This quantitative analysis of usability metrics has led to the following suggestions for improving the WRESTORE tool, which may also be of interest to other researchers developing EDSS:

1. Decrease the time that the user has to expend for giving feedbacks, particularly reducing the number of sessions they need to go through in order to avoid user over-fatigue (Butler & Winne 1995), when it is apparent that the user has become efficient at using the tool; overall task times can be a useful indicator for this.
2. Motivate the use of the AOIs related to information gathering to focus users' attention on more informative aspects of the interface and also increase the confidence levels of the users. This motivation could be achieved through interface development that emphasizes the exploration of areas where users have the opportunity to gather more information by clicking through menus, graphs, maps, and improved data visualizations to allow better comparisons among design alternatives.
3. Provide a final 'summary' session that recapitulates the findings and designs of desirable alternatives found by the users.

ACKNOWLEDGEMENTS

We would like to acknowledge the funding agency – National Science Foundation (Award ID #1014693 and #1332385). We would also like to thank all of our collaborators from the different agencies and institutions: Empower Results LLC team for facilitating user testing, Dr Snehasis Mukhopadhyay and Mr Vidya B. Singh for computational support, and all workshop participants.

REFERENCES

Arabi, M., Govindaraju, R. S. & Hantush, M. 2006 Cost-effective allocation of watershed management practices using a genetic algorithm. *Water Resour. Res.* **42**, W10429.

Arabi, M., Frankenberger, J. R., Engel, B. A. & Arnold, J. G. 2007 Representation of agricultural conservation practices with SWAT. *Hydrol. Process.* **22**, 3042–3055.

Arnold, J. G., Moriasi, D. N., Gassman, P. W., Abbaspour, K. C., White, M. J., Srinivasan, R., Santhi, C., Harmel, R. D., van Griensven, A., Van Liew, M. W., Kannan, N. & Jha, M. K. 2001 SWAT: Model Use, Calibration and Validation. *Trans. ASABE* **55** (4), 1491–1508.

Arnold, J. G., Potter, K. N., King, K. W. & Allen, P. M. 2005 Estimation of soil cracking and the effect on surface runoff in a Texas Blackland prairie watershed. *Hydrological Processes* **19** (3), 589–603.

Artita, K., Kaini, P. & Nicklow, J. W. 2008 Generating alternative watershed-scale BMP designs with evolutionary algorithms. In: *World Environmental and Water Resources Congress*, Ahupua'A, HI, pp. 1–9.

Babbar-Sebens, M. & Misnker, B. S. 2011 Interactive Genetic Algorithm with Mixed Initiative Interaction for multi-criteria ground water monitoring design. *Applied Soft Comp.* **12**, 182–195.

Babbar-Sebens, M., Barr, R. C., Tedesco, L. P. & Anderson, M. 2013 Spatial identification and optimization of upland wetlands in agricultural watersheds. *Ecol. Eng.* **52**, 130–142.

Babbar-Sebens, M., Mukhopadhyay, S., Singh, V. B. & Piemonti, A. D. 2015 A web-based software tool for participatory optimization of conservation practices in watersheds. *Environ. Modell. Softw.* **69**, 111–127.

Barreteau, O. & Abrami, G. 2007 Variable time scales, agent-based models, and role-playing games: the PIEPLUE river basin management game. *Simulation Gaming* **38** (3), 364–381.

Boisvert, E. & Brodaric, B. 2012 Groundwater Markup Language (GWML) – enabling groundwater data interoperability in spatial data infrastructures. *J. Hydroinform.* **14** (1), 93–107.

Burkey, J. 2006 *A Non-Parametric Monotonic Trend Test Computing Mann–Kendall Tau, Tau-b, and Sens Slope Written in Mathworks-MATLAB Implemented Using Matrix Rotations*. King County, Department of Natural Resources and Parks, Science and Technical Services Section, Seattle, WA, USA. Available from: http://www.mathworks.com/matlabcentral/fileexchange/authors/23983.

Butler, D. L. & Winne, P. H. 1995 Feedback and self-regulated learning: a theoretical synthesis. *Rev. Educ. Res.* **65**, 245–281.

Castelletti, A. & Socini-Sessa, R. 2007 Bayesian networks and participatory modelling in water resource management. *Environ. Modell. Softw.* **22** (8), 1075–1088.

de Kok, J., Titus, M. & Wind, H. G. 2000 Application of fuzzy sets and cognitive maps to incorporate social science scenarios in integrated assessment models. A case study of urbanization in Ujung Pandang, Indonesia. *Integrated Assessment* **1** (3), 177–188.

Estes, W. K. 1994 *Classification and Cognition.* Oxford University Press-Clarendon Press, New York, NY.

Gregory, R. 2000 Using stakeholder values to make smarter environmental decisions. *Environment* **42** (5), 34–44.

Haklay, M. & Tobon, C. 2003 Usability evaluation and PPGIS: towards a user-centred design approach. *Int. J. Geogr. Inform. Sci.* **17** (6), 577–592.

Helsel, D. R. & Hirsch, R. M. 2002 *Statistical Methods in Water Resources Techniques of Water Resources Investigations.* Book 4, chapter A3. US Geological Survey, 522 pp. Available from https://pubs.usgs.gov/twri/twri4a3/.

Ice, G. 2004 History of innovative best management practice development and its role in addressing water quality limited waterbodies. *J. Environ. Eng.* **130** (6), 684–689.

Jaber, M. Y. 2011 *Learning Curves: Theory, Models, and Applications.* CRC Press, Boca Raton, FL.

Jakeman, A. J., Voinov, A. A., Rizzoli, A. E. & Chen, A. H. 2008 *Environmental Modelling, Software and Decision Support.* Elsevier B.V., The Netherlands.

Jankowski, P., Robischon, S., Tuthill, D., Nyerges, T. & Ramsey, K. 2006 Design considerations and evaluation of a collaborative spatio-temporal decision support system. *Trans. GIS* **10** (3), 335–354.

Johnson, L. E. 1986 Water resource management decision support system. *J. Water Resour. Plann. Manage.* **112** (3), 308–325.

Kaini, P., Artita, K. & Nicklow, J. W. 2012 Optimizing structural best management practices using SWAT and genetic algorithm to improve water quality goals. *Water Resour. Manage.* **26**, 1827–1845.

Kelly, R. A. & Merritt, W. S. 2010 The Role of Decision Support Systems (DSS) in planning for improved water quality in coastal lakes. In: *Decision Support Systems in Agriculture, Food and the Environment: Trends, Applications and Advances* (B. Manos, M. Matsatsinis, K. Paparrizos & J. Papathanasiou, eds). IGI Global, Hershey, PA, pp. 48–73. doi: 10.4018/978-1-61520-881-4.ch003.

Kirchoff, C. J., Lemos, M. C. & Dessai, S. 2013 Actionable knowledge for environmental decision making: broadening the usability of climate sciences. *Annu. Rev. Environ. Resour.* **38**, 393–414.

Lavoie, R., Joerin, F. & Rodriguez, M. 2015 ATES: a geo-informatics decision aid tool for the integration of groundwater into land planning. *J. Hydroinform.* **17** (5), 771–778.

Lethbridge, M. R., Westphal, M. I., Possingham, H. P., Harper, M. L., Souter, N. J. & Anderson, N. 2010 Optimal restoration of altered habitats. *Environ. Modell. Softw.* **25** (6), 737–746.

Mazerolle, M. J. 2004 *Mouvements et reproduction des amphibiens en tourbieres pertubees* [Movement and Reproduction of Amphibians in Disrupted Wetlands]. PhD dissertation, Dept. Forest. and Geomat., Univ. du Quebec, Quebec, Montreal, Canada.

McIntosh, B. S., Ascough II, J. C., Twery, M., Chew, J., Elmahdi, A., Haase, D., Harou, J. J., Hepting, D., Cuddy, S., Jakeman, A. J., Chen, S., Kassahun, A., Lautenbach, S., Matthews, K., Merritt, W., Quinn, N. W. T., Rodriguez-Roda, I., Sieber, S., Stavenga, M., Sulis, A., Ticehurst, J., Volk, M., Wrobel, M., van Delden, H., El-Sawah, S., Rizzoli, A. & Voinov, A. 2011 Environmental decision support systems (EDSS) development – Challenges and best practices. *Environ. Modell. Softw.* **26** (12), 1389–1402.

Mendoza, G. A. & Martins, H. 2006 Multi-criteria decision analysis in natural resource management: a critical review of methods and new modelling paradigms. *Forest Ecol. Manage.* **230**, 1–22.

Metcalf, S. S., Wheeler, E., BenDor, T., Lubinski, K. S. & Hannon, B. M. 2010 Sharing the floodplain: mediated modelling for environmental management. *Environ. Modell. Softw.* **25** (11), 1282–1290.

Neitsch, S. L., Arnold, J. G., Kiniry, J. R. & Williams, J. R. 2005 *Soil and Water Assessment Tool*. Theoretical Documentation Version 2005. Grassland, Soil and Water Research Laboratory, Agricultural Research Service and Blackland Research Center, Texas Agricultural Experiment Station, Temple, TX.

Newell, A. & Rosenbloom, P. S. 1981 *Mechanisms of Skill Acquisition and the Law of Practice. Cognitive Skills and Their Acquisition.* Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pp. 1–51.

Nyerges, T., Jankowski, P., Tuthill, D. & Ramsey, K. 2006 Collaborative water resource decision support: results of a field experiment. *Ann. Assoc. Amer. Geogr.* **96** (4), 699–725.

Piemonti, A. D., Babbar-Sebens, M. & Luzar, E. J. 2013 Optimizing conservation practices in watersheds: do community preferences matter? *Water Resour. Res.* **49** (10), 6425–6449.

Power, D. J. & Sharda, R. 2007 Model-driven decision support systems: concepts and research directions. *Decision Support Systems* **43** (3), 1044–1061.

Sadler, J. M., Ames, D. P. & Khattar, R. 2016 A recipe for standards-based data sharing using open source software and low-cost electronics. *J. Hydroinform.* **18** (2), 185–197.

Slocum, T. A., Cliburn, D. C., Feddeman, J. J. & Miller, J. R. 2003 Evaluating the usability of a tool for visualizing the uncertainty of the future global water balance. *Cart. Geogr. Inform. Sci.* **30** (4), 299–317.

Tedesco, L. P., Pascual, D. L., Shrake, L. K., Casey, L., Hall, B. H., Vidon, P. G. F., Hernly, F. V., Barr, R. C., Ulmer, J. & Pershing, D. 2005 *Eagle Creek Watershed Management Plan: an Integrated Approach to Improve Water Quality.* CEES Publication 2005-07, IUPUI, Indianapolis, IN.

Tilak, O., Babbar-Sebens, M. & Mukhopadhyay, S. 2011 *Decentralized and Partially Decentralized Reinforcement Learning for Designing a Distributed Wetland System in Watersheds.* IEEE International Conference on Systems, Man, and Cybernetics (SMC), Anchorage, Alaska.

Tullis, T. & Albert, B. 2013 *Measuring the User Experience: Collecting, Analyzing and Presenting Usability Metrics.* Elsevier, Inc., Waltham, MA.

Voinov, A. & Bousquet, F. 2010 Modelling with stakeholders. *Environ. Modell. Softw.* **25** (11), 1268–1281.

Yelle, L. E. 1979 The learning curve: historical review and comprenhensive survey. *Decision Sciences* **10** (2), 302–328.

Zorrilla, P., Carmona, G., De la Hera, A., Varela-Ortega, C., Martinez-Santos, P., Bromley, J. & Henriksen, H. J. 2010 Evaluation of Bayesian networks in participatory water resources management, Upper Guadiana Basin, Spain. *Ecol. Soc.* **15** (3). http://www.ecologyandsociety.org/vol15/iss3/art12/.

First received 9 February 2016; accepted in revised form 13 April 2017. Available online 18 May 2017