British Journal of Educational Technology (2013)
doi:10.1111/bjet.12042

Usability testing of a multimedia e-learning resource for electrolyte and acid-base disorders

Mogamat Razeen Davids, Usuf Chikte, Karen Grimmer-Somers and Mitchell L Halperin

Razeen Davids is Associate Professor in the Division of Nephrology and Usuf Chikte is Head of the Department of Interdisciplinary Health Sciences at Stellenbosch University, Cape Town, South Africa. Karen Grimmer-Somers is Professor of Allied Health and Director of the International Centre for Allied Health Evidence, University of South Australia, and a visiting professor at Stellenbosch University. Mitch Halperin is Emeritus Professor in the Keenan Research Building, Li Ka Shing Knowledge Institute of St. Michael's Hospital and Division of Nephrology at the University of Toronto. Address for correspondence: Prof Mogamat Razeen Davids, Division of Nephrology, Stellenbosch University & Tygerberg Hospital, Cape Town 7505, South Africa. Email: mrd@sun.ac.za

Abstract

The usability of computer interfaces may have a major influence on learning. Design approaches that optimize usability are commonplace in the software development industry but are seldom used in the development of e-learning resources, especially in medical education. We conducted a usability evaluation of a multimedia resource for teaching electrolyte and acid-base disorders by studying the interaction of 15 medical doctors with the application. Most of the usability problems occurred in an interactive treatment simulation, which was completed successfully by only 20% of participants. A total of 27 distinct usability problems were detected, with 15 categorized as serious. No differences were observed with respect to usability problems detected by junior doctors as compared with more experienced colleagues. Problems were related to user information and feedback, the visual layout, match with the real world, error prevention and management, and consistency and standards. The resource was therefore unusable for many participants; this is in contrast to good scores previously reported for subjective user satisfaction. The findings suggest that the development of e-learning materials should follow an iterative design-and-test process that includes routine usability evaluation. User testing should include the study of objective measures and not rely only on self-reported measures of satisfaction.

Introduction

e-Learning is considered to be as effective as educational interventions delivered by traditional media (Chumley-Jones, Dobbie & Alford, 2002; Cook *et al*, 2008) and has rapidly become part of the medical education mainstream (Ellaway & Masters, 2008). Creative educators are increasingly using animation, simulations and virtual 3-D learning environments (Hansen, 2008) to create engaging learning resources for students and health-care professionals. Virtual patients, for instance, hold particular promise for assisting in the development of clinical reasoning ability (Cook & Triola, 2009).

Developing innovative e-learning materials can be expensive and time-consuming. A survey of virtual patient development at US and Canadian medical schools revealed that the cases took an average of 16.6 months to complete and that 85% of them cost over $10 000 (Huang, Reynolds & Candler, 2007). It is therefore important to maximize the educational impact of these

Practitioner Notes

What is already known about this topic

- The usability of computer interfaces may have a major influence on learning.
- While design approaches that optimize usability are common in the software development industry, this is not the case with e-learning, especially in the area of medical education.

What this paper adds

- Neglecting the evaluation of usability may lead to the implementation of e-learning materials with poor usability, with failure to achieve desired educational outcomes.
- The results of objective user testing do not correlate well with evaluations based on self-reported user satisfaction.

Implications for practice and/or policy

- e-Learning development should include routine usability evaluation and follow an iterative design-test-redesign approach.
- Usability evaluation should include observing typical end-users interacting with the system and not be based only on subjective ratings of user satisfaction.

resources. One aspect that has not been sufficiently emphasized in the implementation of effective e-learning is the usability of the technology interface. This has a major impact on learning and should be considered when designing e-learning resources (Sandars, 2010; Zaharias, 2009). Usability is a concept from the field of human–computer interaction that describes the ease with which a technology interface can be used. The International Standard, ISO 9241-11, defines it as the "*Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use*" (Abran, Khelifi, Suryn & Seffah, 2003). A user interface should be so intuitive and self-evident that even inexperienced users can accomplish tasks successfully (Krug, 2006).

High usability of learning resources is essential, though of course not sufficient, to achieving the desired educational impact (Sandars & Lafferty, 2010). This is likely to be especially relevant when the subject matter is complex and contains multiple interacting elements (Sweller, 2010). Such material presents a heavy intrinsic cognitive load in view of the limited capacity of working memory and is often perceived as difficult to learn. Poorly designed user interfaces can present an additional, extraneous cognitive load, as the user has to struggle with challenging content as well as with the technology interface. Reducing extraneous cognitive load has been shown to lead to large gains in learning efficiency (Mayer & Moreno, 2003; van Merrienboer & Sweller, 2010); optimizing the usability of e-learning resources therefore seems essential.

Iterative methodologies that include the routine evaluation of usability are common in the software development industry (Bygstad, Ghinea & Brevik, 2008; Holzinger, Errath, Searle, Thurnher & Slany, 2005; Mao, Vredenburg, Smith & Carey, 2005; Sohaib & Khan, 2010). As far back as the mid-80s, Gould and Lewis (1985) recommended the following design principles: an early focus on users and their tasks; empirical user testing starting early in the development process; and an iterative approach using cycles of design, testing and redesign until the application meets performance and usability goals. This approach is seldom used in the development and evaluation of e-learning resources, especially in medical education (Sandars, 2010). There are two main categories of usability evaluation techniques: *empirical user testing* involves studying typical end-users interacting with the application while *usability inspection methods* involves

experts evaluating the application against a set of rules or design principles (Dumas & Salzman, 2006).

Selecting which methods and measures to use when evaluating an e-learning resource remains difficult. For example, we can evaluate usability, learner interactions, learner perceptions or learning outcomes, we can collect subjective or objective data, qualitative or quantitative data, and we can make use of experts or we can involve typical end-users (Dyson & Campello, 2003). Even if the focus is on usability as in this study, different approaches are available and each will have their own resource requirements, examine a particular aspect of usability and detect different usability problems. A common recommendation is to combine methods whenever resources allow and to alternate between inspection by experts and end-user testing.

User testing usually involves participants being asked to think aloud as they interact with the system being tested. Evaluations may be conducted in settings ranging from sophisticated usability laboratories to informal settings employing paper prototypes (Snyder, 2003). User testing has been rated by usability professionals as having a greater impact on product development than inspection methods, although the latter is also very commonly used (Mao *et al*, 2005; Rosenbaum, Rohn & Humburg, 2000). Developers are less likely to question the validity of the results when usability problems are identified by real users rather than by experts (Dumas & Salzman, 2006). However, real users may be expensive and difficult to recruit and the recording, coding and analysis of testing sessions may also be expensive and time-consuming. Nielsen has popularized simpler methods, pointing out that any testing is better than not testing at all, and demonstrating that four to five users are sufficient for each cycle of testing (Nielsen, 2012). This "discount usability" approach (Nielsen, 2009) may be an efficient option for improving the process of developing e-learning materials.

Inspection methods are often less expensive because they involve fewer people and can detect many problems in a limited amount of time. Evaluators may also suggest solutions to the problems they find. The most commonly used technique is heuristic evaluation, in which expert evaluators find usability problems by examining an interface and judging its compliance with well-established usability principles, called heuristics. The process is influenced by the skills of the evaluators, with the ideal evaluators being "double experts" at usability and the domain of the application being evaluated (Nielsen, 1992). However, such individuals may be difficult to find or very expensive to employ. Evaluators may also have their own biases regarding interface design or may have insufficient domain knowledge, causing domain-specific problems to be missed. They may miss problems that affect real users or identify many low priority problems that hardly affect real users.

The raw data generated by an evaluation need to be transformed before it can be used to improve the user interface (Howarth, Andre & Hartson, 2007). Each occurrence of a usability problem encountered by a user or evaluator is a problem instance. All related instances must be recognized and consolidated into distinct problems, and the problems may then be categorized according to the interface elements involved, the severity of the problems or the design principles violated. See Table 1 for a set of widely used principles for guiding good interface design. Categorizing the problems in this way makes it easier to identify solutions to address them and also to prioritize them for fixing during the subsequent redesign process.

We have developed a web-based learning resource to help students and practicing clinicians acquire expertise in the complex area of electrolyte, water and acid-base disorders, an area of medicine that students and clinicians find particularly difficult to master (Dawson-Saunders, Feltovich, Coulson & Steward, 1990). Patients with these disorders are usually encountered by doctors working in the fields of internal medicine or pediatrics, or in subdisciplines of these fields such as nephrology, endocrinology and intensive care medicine. Our Electrolyte

Table 1: Principles of good interface design (heuristics). The first 10 are those proposed by Nielsen (2005), and the last is from Karat et al. *(1992)*

Heuristic	*Descriptors*
1. Visibility of system status; feedback	Keep users informed through timely appropriate feedback. They always know where they are, which actions can be taken and how they can be performed.
2. Match with the real world—language, conventions	Speak the users' language, use familiar terms and concepts; follow real-world conventions.
3. Consistency and conformity to standards	Words, situations and actions mean the same thing; application uses commonly accepted conventions and conforms to user expectations.
4. Minimize memory load; recognition rather than recall	Objects, actions and options accessed easily. The user should not have to remember information from one part of the application to another.
5. Aesthetic and minimalist design	No irrelevant information as it competes with relevant information and diminishes their relative visibility. Animation and transitions should be used sparingly.
6. Help and documentation	It is better if the system can be used without documentation. If required it should be concise, easy to search and task-centered.
7. User control and freedom	The user can control the direction and pace of the application. Clearly marked exits if they take wrong options by mistake. Support undo and redo.
8. Flexibility and efficiency of use	Users can modify the application to suit their individual capabilities and needs, for example, by using shortcuts.
9. Error prevention and tolerance	Careful design to prevent errors occurring. Despite user errors, the intended result may still be achieved by error correction or good error management.
10. Help users recognize, diagnose and recover from errors	Error messages should be in plain language (no codes or jargon) and suggest a solution.
11. Intuitive visual layout	Position elements on screen to be easily perceived and understandable, and visually attractive.

Workshop provides instruction and the opportunity to practice the treatment of electrolyte disorders through an interactive simulation. The application is freely accessible at http://www.learnphysiology.org/sim1/.

The underlying teaching approach and the initial development of the Electrolyte Workshop have been described previously (Davids, Chikte & Halperin, 2011). The application was built in Flash® and involved several iterations of development and review by the authors and the development team. This informal review process by content experts and experienced developers detected and corrected many usability problems with the application. Self-reported end-user satisfaction with the completed application was good as judged by positive comments and high ratings on the System Usability Scale (Brooke, 1996).

This paper reports on an evaluation that focuses on objective measures of usability obtained by observing, recording and analyzing the interaction of end-users with the application. The study did not address educational outcomes. Testing was conducted with doctors working in the field of internal medicine, our main target audience. The purpose was to determine how well our Electrolyte Workshop conforms to principles of good interface design and to inform further development. The study illustrates the importance of user testing in evaluating e-learning materials and, in particular, demonstrates the need to observe users and examine objective data rather than to rely solely on more easily obtained questionnaire data.

Methods

Ethics approval for the study was granted by the Committee for Human Research at the Faculty of Health Sciences of Stellenbosch University (project no. N08/05/158).

The e-learning resource

The Electrolyte Workshop is built in Adobe® Flash® and consists of case-based tutorials. There are two sections: cases in the WalkThru section present a clinical problem, then demonstrate how an expert would analyze the data and make decisions about treatment. Animation is used to illustrate changes in body fluid compartment sizes, brain cell size and plasma sodium concentrations. The concept is "look and learn," analogous to the use of worked-out examples in other disciplines (Renkl, 2005), which allows students to appreciate the underlying principles rather than being focused on finding solutions to the problem presented.

Cases in the second section, called the HandsOn section, are interactive and include a treatment simulation where users can select from a menu of therapies and receive immediate feedback via animations and text messages. The HandsOn cases have introductory ("lead-in") slides that set the scene for the treatment simulation. These slides contain important clinical and laboratory data that are needed to complete the treatment simulation. After successful completion of the simulation a summary slide is displayed containing several "take-home messages."

Currently the application contains only two cases, one in each section. The WalkThru case is that of a young girl with acute hyponatremia related to Ecstasy use, and the HandsOn case is that of chronic hyponatremia in a patient with Addison's disease.

Participants

User testing was conducted with 15 doctors at an academic department of medicine. The group included 10 doctors who were undertaking postgraduate training in internal medicine ("registrars") and 5 qualified specialists in internal medicine, nephrology and endocrinology. This group is typical of our target population. We considered that the specialists and registrars were likely to be different in terms of subject knowledge and experience, and therefore recruited 15 participants to allow us to include sufficient participants from both groups and also to improve the overall usability problem detection rate (Faulkner, 2003).

User testing equipment and procedures

The application was loaded onto two 15-inch laptop computers, each equipped with a mouse and a webcam with an integrated microphone.

To facilitate the capture and analysis of information from each testing session we installed a usability software tool on each computer. We selected Morae® (http://www.techsmith.com) for this purpose because it is widely used and suited our requirements in terms of data collection and analysis options, cost and ease of use. Running unobtrusively in the background, it records all user interactions with a website or computer application. This includes the user's voice, webcam video of facial expressions and video of all on-screen activity. It also captures data like mouse clicks and keyboard activity. Recordings are marked up to log the start and end of tasks, instances of usability problems, user comments and occasions when help was needed. Metrics like time, task completion rates, usability problem counts and mouse activity are readily generated.

Participants received written instructions. They were required to work through the WalkThru and HandsOn cases and look carefully at the different panels on each slide. They were encouraged to try different options in the treatment simulation and were also asked to look at the glossary. No time limits were set.

Measures of usability

For the purposes of evaluating usability the WalkThru case, the introductory slides of the HandsOn case, the treatment simulation of the HandsOn case, and the glossary were each regarded as a separate task.

Binary task completion rates and the detection of usability problems were recorded for each task as measures of effectiveness. Time on task and input device activity (mouse clicks and mouse movement) was recorded for each task as measures of efficiency.

Successful task completion in the WalkThru case and the introductory slides of the HandsOn case simply required that participants navigate through that section from beginning to end, viewing all the information available. For completion of the interactive treatment simulation in the HandsOn case participants had to treat their patient effectively by applying appropriate therapy at the correct dosages, and then exit the simulation to end with a summary "take home messages" slide. In the case of the glossary, participants were simply required to open it by clicking a text hyperlink on a slide or by using its navigation tab at the top of the screen.

The usability problems detected by participants as they worked through the tasks were categorized by severity, the interface element involved and the design principle (heuristic) violated. Our definition of a serious usability problem is based on that of Nielsen (1997), which takes into account the impact, frequency and persistence of the problem; it refers to a problem that may cause unacceptable delays or even task failure for the user and which needs to be fixed before an application is released. Table 1 lists the heuristics we considered when analyzing the usability problems detected. They are based on those proposed by Nielsen (2005) and as used by Karat, Campbell and Fiegel (1992). Each problem identified was mapped to one or more heuristic.

Statistical tests

Binary task completion rates are reported as proportions, usability problems as counts, and time on task (in minutes) and mouse activity (clicks and movement in pixels) as means ± SD. For the comparisons between specialists and registrars, and between those participants who completed a task successfully and those who did not, Fisher's exact test was used to compare proportions, and the Wilcoxon rank sum test was used to compare usability problem detection, time on task and mouse activity. The significance level was set at .05.

Results

User testing focused on measures of effectiveness and efficiency and yielded data that are described below and in Table 2. Although not the focus of this study, we also compared specialists

Table 2: Measures of effectiveness: successful task completion rates and counts of usability problems detected by participants. Where the same problem was encountered by multiple participants these instances were merged to provide a count of unique or distinct problems

	Task completion	*All problems*		*Serious problems*	
	Rate (%)	*Problem instances*	*Distinct problems*	*Problem instances*	*Distinct problems*
Task 1: WalkThru case	15/15 (100)	4	4	1	1
Task 2: HandsOn lead-in slides	8/15 (53)	16	5	10	2
Task 3: HandsOn treatment simulation	3/15 (20)	44	18	34	12
Total		64	27	43	15

with registrars, and participants who completed a task successfully with those who did not, and summarize these results at the end of this section.

Measures of effectiveness: task completion rates and usability problem detection (Table 2)

Task completion rates

Participants all completed the WalkThru case with ease. The lead-in section of the HandsOn case was completed successfully by eight participants (53% task completion rate) while the treatment simulation was completed successfully by only three participants (20%). The glossary was viewed by nine participants, none of whom experienced any usability problems while accessing this feature of the application. All of them opened the glossary by clicking its main navigation tab at the top of the screen and not via a text hyperlink on one of the slides.

Usability problem detection

A total of 27 distinct usability problems were identified, 15 of which were categorized as serious. A median of 4 problems were detected per participant, and in the case of the serious problems the median detection rate was 3 per participant. Table 3 contains a sampling of the serious usability problems detected, and lists the interface elements involved, the heuristics violated, as well as proposed solutions for addressing these problems.

In the WalkThru case four distinct usability problems were detected: these related to user information and feedback (two problems), user control and freedom (unclear navigation, one problem) and match with the real world (a problem with case accuracy, one problem). The only error categorized as serious was the last mentioned, which violated the heuristic of matching with the real world. An animation showed fluid moving out of the intracellular fluid compartment then simply disappearing and not appearing in the extracellular fluid compartment (see the first line of Table 3 for details and Multimedia Appendix S1 for a video clip).

In the lead-in section of HandsOn case a total of five distinct usability problems were identified (16 separate instances were recorded). They related to user information (one problem), the visual layout (two) and match with the real world (two). Two problems were categorized as serious: one was related to inadequate user information and the other to the heuristic of providing an intuitive visual layout. A sliding panel displaying important laboratory data opens on clicking its tab on the side of the screen (Figure 1). This sliding panel was completely missed by seven participants (47%). One of these participants worked through the case twice, and two others worked through it three times without discovering the panel (see line 2 of Table 3 for details and Multimedia Appendix S2 for a video clip).

In the treatment simulation of the HandsOn case a total of 18 distinct usability problems were identified (44 separate instances were recorded). These were related to user information and feedback (five problems), visual layout (three), match with the real world (one), user control and freedom (one), consistency and conformity to standards (two), error prevention and tolerance (five) and error management (one). Twelve of these 18 problems were graded as serious, based on their impact and the frequency of their occurrence.

The first serious usability problem identified in the treatment simulation related to the fidelity of the case and lack of clarity regarding the correct treatment (Table 3 line 3). Two participants, both experienced specialists, were not convinced of the need to apply any fluid therapy in this case of Addison's disease.

The most frequently encountered problem related to the heuristic of designing for error prevention and tolerance. There were repeated unsuccessful attempts by 10 participants (67%) to apply multiple treatments simultaneously (Table 3 line 4 and Multimedia Appendix S3). The simulation was designed to allow treatments to be applied sequentially, not simultaneously, so that feedback could be given after each step. Groups of treatment options are displayed in separate panels.

Table 3: Selected examples of serious usability problems detected with the interface element and heuristic involved, and proposed measures to address them. The first example is from the WalkThru case and the others from the HandsOn case. The number of participants encountering a particular problem is included in column 1. Quotes from participants are in italics

Examples of usability errors	*Interface element*	*Heuristic involved*	*Solution*
The sums don't add up; 1.8L moved out [of the intracellular fluid] but I don't see it in the extracellular fluid! (1/15)	WalkThru case: Case accuracy	Match with the real world: language, conventions, case accuracy	Revise the animation to show the extracellular fluid compartment increasing in volume as 1.8L of water moves into it from the intracellular fluid compartment.
Participants do not notice the sliding lab data panel; this panel contains important information on blood and urine chemistry (9/15)	Lead-in slides: lab data tab	Intuitive visual layout	Redesign the interface to avoid using the sliding panel—group all related data and display in plain view in the left panel.
In this patient with Addison's, why can't you start with only the mineralocorticoid then wait? (2/15)	Treatment simulation: case accuracy	Match with the real world: language, conventions, case accuracy	As the simulation is designed to provide practice at prescribing accurate fluid therapy, amend the case data so that the need for fluid treatment is clear.
Participants try unsuccessfully to select and apply multiple treatments simultaneously; the application is designed to have one treatment given at a time, with feedback supplied after each step. (10/15)	Treatment simulation: treatment selection	Error prevention and tolerance	Remove all panel covers from the treatment option groups so that users clearly see that only one option can be selected and applied at a time. Reinforce this in the information provided just before the simulation is attempted.
Clicking the slider rail does not indicate dose despite the "thumb" moving; have to drag the thumb or double-click on the rail. (2/15)	Treatment simulation: slider control	Consistency and conformity to standards	Reprogram the slider so that the thumb (also) moves with a single click on the rail.
Participants who do not understand the use of the slider control are applying treatments with a dose of zero. (5/15)	Treatment simulation: slider control	Error prevention and tolerance	Program an error message to pop up if the dose applied is zero; explain that the slider must be dragged to indicate the dose.
Error message "Please select a radio option" is not clear to all participants. (1/15)	Treatment simulation: error messages	Help users recognize, diagnose and recover from errors	Avoid jargon or provide links to the glossary.
Messages displayed are sometimes vague or unhelpful. (3/15)	Treatment simulation: error messages	Help users recognize, diagnose and recover from errors	Review the algorithms underlying the error messages; ensure that all messages are relevant and useful.
Some participants seem to struggle to end the HandsOn case—the ending is not clear especially if the treatment applied in the simulation was not successful. (2/15)	Treatment simulation: navigation	User control and freedom	Offer all users access to the summary "take home messages" slide, even when they have not applied treatment successfully.

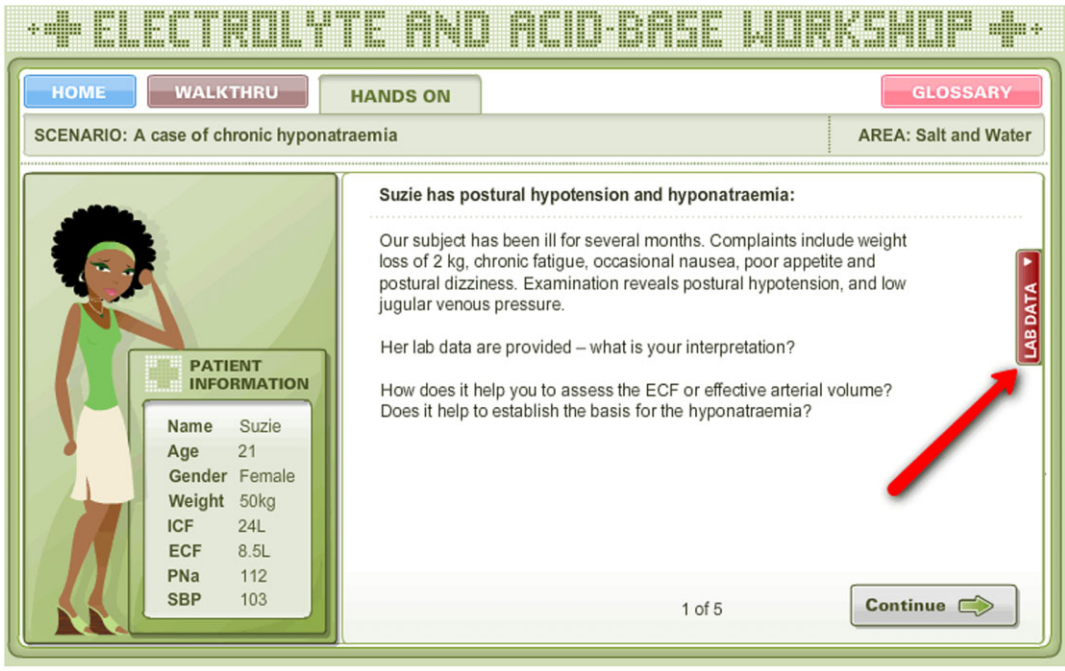

Figure 1: The lab data panel slides open on clicking its tab at the side of the screen (arrow). This was missed by several participants

Clicking on a panel cover causes it to slide open to reveal the options for that treatment group. When clicking on the panel for another treatment group, those options are revealed while the previous panel closes. This design led to much confusion and frustration. Most participants did not realize that a selected option was deselected once they clicked on another panel to try and add a second treatment.

A second serious problem, also related to error prevention and error management, was that some participants were unable to use the slider control (Table 3 lines 5 and 6). They would select a therapy but fail to indicate the dose by dragging the "thumb" along the rail of the slider control and would therefore apply a dose of zero (Figure 2 and Multimedia Appendix S4). As a result there was no change in plasma Na concentration or fluid compartment volumes. The impact of this usability problem was compounded by the display of poor feedback messages. For example, "*Your patient remains stable. Would you like to try something else?*" was vague and unhelpful, and contributed to participants' frustration. Additional usability problems related to the slider control are illustrated in Figure 3 and through video clips in Multimedia Appendices S5 and S6.

Problems with respect to user control and freedom were exposed when some participants appeared to have difficulty ending the simulation. The summary "take-home messages" slide was displayed only after successful completion of the simulation. After unsuccessful treatment attempts, participants were only offered the choice to try again, or to exit without any further feedback.

Measures of efficiency: time on task and mouse activity

Time on task

Participants spent a mean of 8.4 minutes on the WalkThru case, 6.8 minutes on the lead-in section of the HandsOn case and 9.9 minutes on the treatment simulation. The participants who accessed the glossary spent a mean of 1 minute on that part of the application.

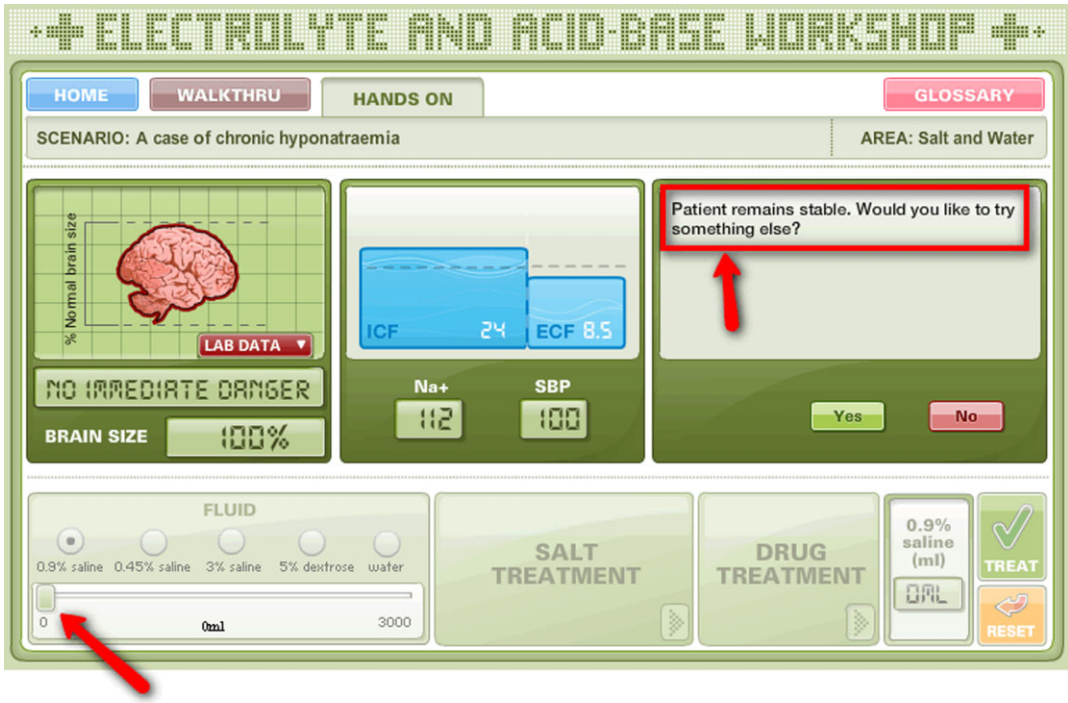

Figure 2: The participant has clicked "Treat" without using the slider to indicate the dose of 0.9% saline, and there is therefore no change in any patient parameter. The feedback message is unhelpful

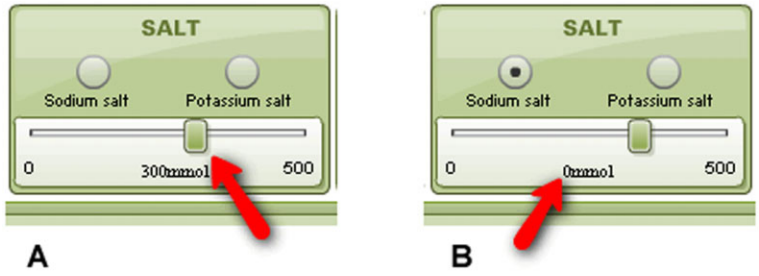

Figure 3: A. The participant has indicated a dose of 300 mmol without first having to select a treatment option by clicking one of the radio buttons. B. Clicking on the slider rail causes the thumb to jump to the point clicked but the dose indicated is still 0 mmol

Mouse clicks and mouse movement

As expected, the treatment simulation, being the most interactive section of the application, had the greatest mouse activity with a mean of 98.3 clicks and 38 884 pixels of mouse movement per participant.

Specialists versus registrars

There were no differences between the specialists and registrars with respect to task completion rates, or in the median number of total usability problems or serious usability problems detected.

The time spent on each task by specialists and registrars was similar. However, with regard to mouse activity for the treatment simulation, the specialists had a lower mean mouse click count

(47.4 vs. 123.8, $p = .010$) and less mouse movement (20 090 vs. 48 281 pixels, $p = .020$) than the registrars. Mouse activity for the other tasks was similar.

Successful task completion versus failed task completion

The two tasks that were not completed successfully by all participants were the lead-in section of the HandsOn case and the treatment simulation of the HandsOn case. Successful participants on the lead-in section had a higher mean mouse click count (20.4 vs. 11.7, $p = .042$) while those who completed the simulation successfully had a lower mouse click count (41.3 vs. 112.6, $p = .030$) and less mouse movement (14 859 vs. 44 890 pixels, $p = .030$). The successful participants on the lead-in section had a lower usability problem detection rate on this task ($p = .017$) while there was no difference on the treatment simulation.

Discussion

Despite having followed an iterative design-and-review process involving the authors and developers, this evaluation with typical end-users detected several serious usability problems that had not been exposed during the initial development. Almost all were related to the interactive HandsOn tutorial and, in particular, the treatment simulation. Based on this evaluation, our e-learning application fell short with respect to principles of good interface design and would have been unusable for a large proportion of users, thus severely limiting the potential educational impact.

This finding is in contrast to the satisfactory self-reported user feedback previously obtained and confirms the observation that subjective measures of users' perceptions are often poorly correlated with objective measures (Bangor, Kortum & Miller, 2008). Our participants were aware who had developed the system and might well have been less critical in their responses because of this. When the aim is to improve the usability of a product, it is clear that it is not sufficient to employ only subjective measures of user satisfaction. The problems detected by employing user testing have allowed us to compile a detailed list of suggested revisions for the next iteration of the application.

Employing specialized usability software provided us with a rich source of data in the form of video recordings and usability metrics, giving us unique insights into participants' experiences. This allowed us to appreciate the full impact of the usability problems detected. For example, the levels of frustration—visible on participants' faces as they struggled with the slider control and repeatedly applied dosages of zero with no change in any patient parameters—may have been missed without the webcam video data stream. Another example was where the recordings provided accurate quantitative data that helped us to evaluate the utility of the glossary. Only 60% of participants accessed this section of the application—a mere 1 minute was spent there by those who did—and not one participant reached the glossary by clicking on a text hyperlink to access an explanation or definition as was intended. It would seem that participants only opened the glossary because this was required by the written instructions provided. It is probable that our participants were familiar with the terminology and concepts used and hence had little need to consult the glossary. This type of user support might be of more value to undergraduate students.

Registrars tended to spend more time on each task and had more mouse activity, although this was statistically significant only for mouse activity in the treatment simulation. This might reflect them finding the content more unfamiliar and challenging as opposed to their senior colleagues but may also reflect a greater inclination to explore the application. As expected, fewer mouse clicks were recorded by participants who missed the sliding data panel in the lead-in section of the HandsOn case. In the treatment simulation, participants who could not complete the task successfully had much more mouse activity as they made one failed attempt after another.

While there were few usability problems detected in the WalkThru case, the interactive HandsOn section was effectively unusable for the majority of our participants. This was true for both experienced clinicians and their junior colleagues. It was therefore not possible for these participants to achieve the intended objective of improving their skill and confidence in treating hyponatremia through practice in a simulated environment.

The design flaws causing the poor usability violated a number of heuristics. The principle of ensuring visibility of system status means that users should always know what was happening through clear information and appropriate feedback. Our feedback messages were often unhelpful or irrelevant. The heuristic of error prevention and management was not well implemented, as evidenced by the problems with the slider control and the repeated attempts at multiple treatment selection, which was compounded by the unhelpful error messages. The sliding lab data panel that was missed by many participants indicated that we did not succeed in providing an intuitive visual layout. While user control and freedom was reasonably well ensured by the clear navigation and the self-paced nature of the application, several users appeared to be unclear how to exit the HandsOn case, as it did not display the closing summary slide unless treatment had been successful.

The question of how many users are sufficient to evaluate a technology interface has long been debated in the usability literature. Five users will, on average, uncover 80% of usability problems (Turner, Lewis & Nielsen, 2006). This well-known "five users is enough" approach is appropriate when the probability of each user discovering a given problem is around 0.3, when applications are not too large and complex, when testing is done at an early stage of development and when several cycles of design-and-test are envisaged (Turner *et al*, 2006). When the application is larger and complex or when later versions are tested after the most obvious problems are already fixed then the probability of problem detection will fall and five users will not be enough (Spool & Schroeder, 2001). When the application is designed for more than one target group, then users from each subpopulation will need to be recruited and once again a greater number of users will be required.

Faulkner (2003) found that while five users detected a mean of 80% of problems present, wide confidence intervals implied that a particular set of five users detected as few as 55% of the problems. With 10 users, the lowest percentage of problems detected was 80%, and with 15 users it was 90%. Faulkner recommended testing the maximum number of users that resources allow to increase the confidence that the problems that need to be fixed will be found (Faulkner, 2003). We followed this recommendation. Our group of 15 participants enabled us to include both specialists and registrars, who differed in subject knowledge and clinical experience. We believed that this might impact on the detection of usability problems; however, we found that the usability problem detection rates were similar in the two groups and thus independent of differences in expertise.

Large increases in key metrics have been documented using an iterative approach to improve the usability of websites and software applications (Marcus, 2005). At least two cycles of usability testing should be undertaken, starting in the early stages of development and using simple prototypes or wireframes. Another cycle of testing should be undertaken with the fully functional "live" version of the product. Additional testing is advisable since new problems may be introduced when fixing the old ones. Ideally, this should continue until no new problems of significance are detected, but this iterative process will often be cut short by practical considerations. We believe that our Electrolyte Workshop requires at least one more cycle of revision and evaluation before we will have a robust and well-designed e-learning resource.

Several lessons were learned in the course of doing the study. End-users need to be involved much earlier, ideally before or at the stage where simple prototypes or wireframes are being built. Even

experienced developers will not anticipate all the problems that a novice user may encounter, as was starkly demonstrated here. It is also clear that using only satisfaction ratings is insufficient, as these may correlate poorly with other, more objective, measures of usability.

Usability inspection methods, especially heuristic evaluation, may offer another efficient option in evaluating e-learning materials if an expert panel with the required experience and expertise can be assembled. We have learned that user testing can be resource intensive with suitable users difficult or expensive to recruit, and conducting, recording and analyzing testing sessions very time-consuming. It may therefore be most efficient to first use heuristic evaluation to find and fix the most obvious problems and then to undertake testing with a small number of end-users.

Our informal reviews during the initial development did not involve usability experts or the use of formal guidelines or checklists, and overlooked many serious problems. Inspection techniques as well as user testing can be used from the very early stages of the development process. If usability evaluation is only done at the end of the design cycle, changes to the interface are usually more costly and difficult to implement.

Conclusions

Our usability evaluation, which was facilitated by specialized usability software, allowed us to identify many problems that were missed during the initial development process. These problems would otherwise have gone undetected and we would have released a resource with very limited potential educational impact. Our findings will inform a careful revision of the application and guide further content development. Future studies will examine the effect of optimizing usability on measures of learning as well as on users' motivation and engagement with the application.

The design of e-learning materials, modules and programs for medical education should include routine usability evaluation and follow an iterative design-and-test process. This is essential if we are to exploit the full potential of the electronic medium and maximize learning outcomes for all users in our target populations. User testing should be employed from the earliest phases of development and should include the study of objective measures obtained by observing the interaction of users with the system being tested, and not rely only on subjective measures of user satisfaction.

Conflicts of interest

None.

Acknowledgements

We thank Justin Harvey of the Stellenbosch University Centre for Statistical Consultation for assistance with data analysis and Martin Schreiber for a critical review of the manuscript.

Funding

This work was supported by grants from the South African Universities Health Sciences IT Consortium and Stellenbosch University's Fund for Innovation and Research into Learning and Teaching.

References

Abran, A., Khelifi, A., Suryn, W. & Seffah, A. (2003). Usability meanings and interpretations in ISO standards. *Software Quality Journal, 11*, 325–338.

Bangor, A., Kortum, P. T. & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction, 24*, 574–594.

Brooke, J. (1996). SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & I. L. McClelland (Eds), *Usability evaluation in industry* (pp. 189–194). London: Taylor & Francis.

Bygstad, B., Ghinea, G. & Brevik, E. (2008). Software development methods and usability: perspectives from a survey in the software industry in Norway. *Interacting with Computers, 20*, 375–385.

Chumley-Jones, H. S., Dobbie, A. & Alford, C. L. (2002). Web-based learning: sound educational method or hype? A review of the evaluation literature. *Academic Medicine*, *77*, S86–S93.
Cook, D. A., Levinson, A. J., Garside, S., Dupras, D. M., Erwin, P. J. & Montori, V. M. (2008). Internet-based learning in the health professions: a meta-analysis. *Journal of the American Medical Association*, *300*, 1181–1196.
Cook, D. A. & Triola, M. M. (2009). Virtual patients: a critical literature review and proposed next steps. *Medical Education*, *43*, 303–311.
Davids, M. R., Chikte, U. M. E. & Halperin, M. L. (2011). Development and evaluation of a multimedia e-learning resource for electrolyte and acid-base disorders. *Advances in Physiology Education*, *35*, 295–306.
Dawson-Saunders, B., Feltovich, P. J., Coulson, R. L. & Steward, D. E. (1990). A survey of medical school teachers to identify basic biomedical concepts medical students should understand. *Academic Medicine*, *65*, 448–454.
Dumas, J. S. & Salzman, M. C. (2006). Usability assessment methods. *Reviews of Human Factors and Ergonomics*, *2*, 109–140.
Dyson, M. C. & Campello, S. B. (2003). Evaluating virtual learning environments: what are we measuring. *Electronic Journal of E-Learning*, *1*, 11–20.
Ellaway, R. & Masters, K. (2008). AMEE Guide 32: e-Learning in medical education Part 1: learning, teaching and assessment. *Medical Teacher*, *30*, 455–473.
Faulkner, L. (2003). Beyond the five-user assumption: benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, *35*, 379–383.
Gould, J. D. & Lewis, C. (1985). Designing for usability: key principles and what designers think. *Communications of the ACM*, *28*, 300–311.
Hansen, M. M. (2008). Versatile, immersive, creative and dynamic virtual 3-D healthcare learning environments: a review of the literature. *Journal of Medical Internet Research*, *10*, e26.
Holzinger, A., Errath, M., Searle, G., Thurnher, B. & Slany, W. (2005). From extreme programming and usability engineering to extreme usability in software engineering education (XP+ UE→ XU). Proceedings of the 29th Annual International Computer Software and Applications Conference (COMPSAC'05) (pp. 169–172). Washington, DC: IEEE Computer Society.
Howarth, J., Andre, T. S. & Hartson, R. (2007). A structured process for transforming usability data into usability information. *Journal of Usability Studies*, *3*, 7–23.
Huang, G., Reynolds, R. & Candler, C. (2007). Virtual patient simulation at U.S. and Canadian medical schools. *Academic Medicine*, *82*, 446–451.
Karat, C.-M., Campbell, R. & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. (pp. 397–404). Monterey, California, USA, ACM.
Krug, S. (2006). *Don't make me think! A common sense approach to Web usability*. Berkeley, CA: New Riders.
Mao, J. Y., Vredenburg, K., Smith, P. W. & Carey, T. (2005). The state of user-centered design practice. *Communications of the ACM*, *48*, 105–109.
Marcus, A. (2005). User interface design's return on investment: examples and statistics. In R. G. Bias & D. J. Mayhew (Eds), *Cost-Justifying usability: an update for the internet age* (2nd ed.). (pp. 17–39). San Francisco, CA: Elsevier.
Mayer, R. E. & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*, 43–52.
Nielsen, J. (1992). Finding usability problems through heuristic evaluation. Proceedings of the 1992 SIGCHI conference on Human Factors in Computing Systems (pp. 373–380). Monterey, CA: ACM.
Nielsen, J. (1997). Severity ratings for usability problems. Retrieved 2011/09/07, from http://www.useit.com/papers/heuristic/severityrating.html
Nielsen, J. (2005). Ten usability heuristics. Retrieved 31/08/2012, from http://www.useit.com/papers/heuristic/heuristic_list.html
Nielsen, J. (2009). Discount usability: 20 years. Retrieved 28/08/2012, from http://www.useit.com/alertbox/discount-usability.html
Nielsen, J. (2012). How many test users in a usability study? Retrieved 28 August 2012, from http://www.useit.com/alertbox/number-of-test-users.html
Renkl, A. (2005). The worked-out examples principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 229–245). Cambridge, UK: Cambridge University Press.
Rosenbaum, S., Rohn, J. A. & Humburg, J. (2000). A toolkit for strategic usability: results from workshops, panels, and surveys. Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 337–344). The Hague, The Netherlands: ACM.

Sandars, J. (2010). The importance of usability testing to allow e-learning to reach its potential for medical education. *Education for Primary Care*, *21*, 6–8.
Sandars, J. & Lafferty, N. (2010). Twelve Tips on usability testing to develop effective e-learning in medical education. *Medical Teacher*, *32*, 956–960.
Snyder, C. (2003). *Paper prototyping: the fast and easy way to design and refine user interfaces*. San Diego, CA: Morgan Kaufmann.
Sohaib, O. & Khan, K. (2010). Integrating usability engineering and agile software development: a literature review. International Conference on Computer Design and Applications (pp. 32–38).
Spool, J. & Schroeder, W. (2001). Testing web sites: five users is nowhere near enough. Proceedings of the CHI 2001 conference on Human Factors in Computing Systems (pp. 285–286). Seattle, Washington: ACM.
Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*, 123–138.
Turner, C. W., Lewis, J. R. & Nielsen, J. (2006). Determining usability test sample size. In W. Karwowski (Ed.), *International encyclopedia of ergonomics and human factors* (2nd ed.). (pp. 3084–3088). Boca Raton, FL: CRC/Taylor & Francis.
van Merrienboer, J. J. & Sweller, J. (2010). Cognitive load theory in health professional education: design principles and strategies. *Medical Education*, *44*, 85–93.
Zaharias, P. (2009). Usability in the context of e-learning. *International Journal of Technology and Human Interaction*, 5, 37–59.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Multimedia Appendix S1: This animation illustrates the movement of 1.8 L of fluid out of cells when hypertonic saline is used to treat acute hyponatremia. There is a problem with fidelity in that this fluid does not appear in the extracellular fluid compartment.

Multimedia Appendix S2: The lab data panel slides open on clicking its tab on the right side of the screen. This panel is easily missed and obscures on-screen text when open.

Multimedia Appendix S3: Participants tried unsuccessfully to select and apply multiple treatments simultaneously. Most participants did not realize that their first option was deselected once they clicked on another panel to try and add a second treatment.

Multimedia Appendix S4: Some participants failed to indicate the dose by dragging the "thumb" along the rail of the slider and therefore applied dosages of zero. The impact of this usability problem was compounded by the display of inappropriate feedback messages.

Multimedia Appendix S5: After a single click on the rail of the slider, the slider thumb jumps to the point clicked but the dose indicated is still 0 mmol. The dose is only registered when the thumb is dragged or the rail is double-clicked.

Multimedia Appendix S6: The slider is visible, and the participant is able to indicate a dose without first selecting the treatment to be applied.