Final accepted manuscript. To be published in Applied Ergonomics (2016).

Combining situated Cognitive Engineering with a novel testing method in a case study comparing two infusion pump interfaces

Authors: R. Schnittker[a1], M. Schmettow[a], F. Verhoeven[c] and J.M.C. Schraagen[a,b]

a. Department of Cognitive Psychology and Ergonomics, Faculty of Behavioural, Management and Social Sciences, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; e-mail raphaela.schnittker@monash.edu; m.schmettow@utwente.nl; j.m.c.schraagen@utwente.nl;

b. Department of Human Behaviour and Organisational Innovation, TNO Earth, Life and Social Sciences, P.O. Box 23, 3969 ZG Soesterberg, The Netherlands; e-mail: jan_maarten.schraagen@tno.nl

c. Institute for Engineering and Design, Utrecht University of Applied Sciences, P.O. Box 182, 3500 AD Utrecht, The Netherlands; e-mail: fenne.verhoeven@hu.nl

Corresponding author:
Jan Maarten Schraagen, PhD
TNO Earth, Life and Social Sciences
P.O. Box 23
3769 ZG Soesterberg
The Netherlands
E-mail: jan_maarten.schraagen@tno.nl
Phone: +31 888 665 945

[1] Present address: Monash Injury Research Institute, Building 70, 21 Alliance Lane - Monash University, Clayton Campus, Victoria 3800, Australia, phone (mobile): +61432251295

Final accepted manuscript. To be published in Applied Ergonomics (2016).

Highlights

- Infusion pump interface designed with situated Cognitive Engineering was validated
- Usability validation took place by comparison with a reference interface
- A novel process tracing technique was used for analysis
- Task performance increased with novel interface
- Novel method was feasible for validating safety of medical devices

Abstract

We validated the usability of a new infusion pump interface designed with a situated Cognitive Engineering approach by comparing it to a reference interface using a novel testing method employing repeated measurements and process measures, in addition to traditional outcome measures. The sample consisted of 25 nurses who performed eight critical tasks three times. Performance measures consisted of number and type of errors, deviations from a pre-established normative path solution, task completion times, number of keystrokes, mental effort and preferences in use. Results showed that interaction with the new interface resulted in 18% fewer errors, 90% fewer normative path deviations, 42% lower task completion times, 40% fewer keystrokes, 39% lower mental effort and 76% more subjective preferences in use. These outcomes suggest that within the scope of this case study, combining the situated Cognitive Engineering approach with a novel testing method addresses various shortcomings of earlier testing methods.

Keywords: Medical device usability testing; infusion pump; human-machine interaction

1 Introduction

While infusion pumps contribute to patient care, they are not without risks. From 2005 to 2009, around 56,000 adverse drug events associated with the use of infusion pumps were reported (Center for Devices and Radiological Health, 2011). Many of those use-related hazards were related to user-interface design deficiencies (Center for Devices and Radiological Health, 2010), the critical impact of which on the patient's safety is a well-known problem (Obradovich and

Woods, 1996; Vicente et al., 2003). Approaching designs using Human Factors engineering has proven to be an effective means to enhance positive performance outcomes, such as fewer errors, less time to performance tasks and lower mental effort (Lin et al., 1998; Syroid et al., 2012).

Nevertheless, the current practice in studying medical device technology has methodological shortcomings, as evidenced by an extensive literature study in this field (Schraagen and Verhoeven, 2013). First, previous studies lack a profound analysis of the user-device interaction by mainly focusing on final task outcomes (errors) and completion time as primary performance measures (Schraagen and Verhoeven, 2013). It has been suggested that only 69.5% of the practitioners pressed keystrokes contributing towards the goal state that is aimed to be achieved (Nunnally et al., 2004). Hence, merely measuring erroneous task outcomes undervalues the impact of complex menu structures on the process of task completion, and thus the occurrence of near accidents. Secondly, past studies draw their conclusions upon single user-device interactions, and are therefore unable to investigate the impact of learning effects on the infusion pump's usability (Garmer, 2002; Schraagen and Verhoeven, 2013). Lastly, previous studies lack a combination of subjective and objective measures in order to gain a more complete picture of the user-device interaction (Hornbæk, 2006).

1.1 Goal of the present study

The aforementioned shortcomings in studying medical device technology potentially limit the informative value with respect to the effectiveness of Human Factors engineering in medical device design. The aim of this study is to evaluate the usefulness of a novel testing method, in a case study involving the comparison of two infusion pump interfaces. The study compared an existing infusion pump interface with a new infusion pump interface that has been designed with a situated Cognitive Engineering Human Factors approach. The novel testing method utilized in this study addresses the current limitations in the study of medical device design. Specifically, this study addresses these shortcomings by combining qualitative and quantitative analyses with objective and subjective measures in a usability validation study with repeated measures. Hence, the main aim of this study is to investigate whether this novel testing method would address shortcomings of previous methods.

1.2 Novel method for studying medical device technology

As reviewed by Schraagen and Verhoeven (2013), contemporary methods for studying medical device technology mostly report traditional outcome measures rather than 'process tracing techniques' placing emphasis on cognitive processes. To address this shortcoming, we introduce a novel, replicable method for a standardized representation of the user's task completion process. Our proposed method is feasible for qualitative or quantitative research, as well as mixed-method approaches. We applied the Goals, Operators, Methods, and Selection rules (GOMS) model (John and Kieras, 1996) as a framework for data coding to achieve a formal representation of task execution processes. In addition, introduction of novel interaction design requires initial learning (MacKenzie and Zhang, 1999) and may even be hampered by inappropriate transfer (Besnard and Cacitti, 2005). In order to capture performance differences beyond the first encounter, we explored the impact of task repetition on performance.

2 Methods

Designing interfaces of medical devices is a complex consideration of multiple aspects. Several frameworks, tools, methods, and case studies are available regarding the application of Human Factors to the design of medical devices (Furniss et al., 2014). Despite this, a recent study showed that development teams still face challenges in incorporating Human Factors when designing interactive medical devices (Vincent et al., 2014). An integrative framework, that addresses both users and the technology, and in which results from both theoretical and field research can account for choices in the design process could not be identified for medical device design. Therefore, we adopted a situation Cognitive Engineering systems perspective that has been successfully applied in other complex task environments such as space laboratories, ship control centers etc. (Neerincx and Lindenberg, 2008). This is a coherent three-phase-process (see Table 1 for a phase description) with accompanying methods to systematically arrive at validated user interface requirements. The core of the methodology is the theory-driven specification of claims (phase 1 and 2) and their empirical validation (phase 3).

Table 1. Design process of new interface based on situated Cognitive Engineering

Phase	Methods	Output
1. Derive: Integrated analysis of the operational, human factors, and technological drivers or constraints	Contextual interviews (n=5) to determine type of errors that occur with the use of infusion pumps and identify set needs and wants regarding 'ideal pump interface' Systematic literature review to identify existing design requirements of infusion pump interfaces (Schraagen and Verhoeven, 2013))	Requirements baseline: 9 use cases and 82 requirements, paper prototype of test interface
2. Specify: Specification of requirements baseline with its design rationale (claims and use cases)	Interviews (n=7) to prioritize and correct inadequacies before requirements were translated into a concept user interface (Expert meeting (n=8) to rank prioritized requirements	Prioritized requirements (9 use cases and 41 requirements), revised paper prototype of test interface
3. Test and refine: Three evaluation approaches: reviews, human-in-the-loop evaluations and simulations.	Paper prototyping (n=7) to validate whether requirements baseline were correctly translated into user interface Expert meeting (n=8) to decide upon definite set of requirements that were incorporated into dynamic working prototype of test interface Experiment (n=25) to validate user requirements by means of comparing new interface with reference interface (the experiment reported in this paper)	Validated requirements (9 use cases and 41 requirements) and validated, working prototype of test interface

Next to phases and methods, the methodology also provides a specific format to specify and validate user requirements. Figure 1 depicts an example of this format. The key elements of the format are use cases, requirements and claims. Use cases describe the general behavior requirements for the device that is being designed. Nine use cases were formulated, eight of them describing an interaction between the infusion pump and a user: (1) start and stop infusion, (2) inserting and removing syringe, (3) pausing infusion, (4) alarms, (5) switching pump on and off, (6) bolus, (7) drug group, (8) occlusion. The ninth use case concerned a rest category containing requirements on a more general level, relevant for each user-pump interaction, such as font size, contrast, spacing and distinctiveness of buttons, etcetera, which we labeled "usability/ ergonomics". Each use case referred to multiple requirements (indicating what the user should be able to do with the infusion pump) and to one or more claims, containing the evidence from literature or empirical research for the need of the particular requirement. Claims are included to justify design decisions, highlighting the upsides, downsides and trade-offs involved. If the claims are an adequate justification of the requirements, then a system adhering to the requirements will help reach the design objective. Claims have to be specific and testable, and defined in terms of outcome measures such as effectiveness (accurate and complete), efficiency (time), satisfaction, etc. In the end, the new interface (working prototype) was based on 41 validated user requirements. Figure 1 depicts an example of a use case and one accompanying

requirement. As can be seen, the "claim"- requirement integrates evidence from literature and empirical research that we conducted to inform the interface design, facilitating the complex issues we had to consider when designing the new interface.

UC 6	Start/ stop infusion
Description:	Starting and stopping infusion (not temporarily, but starting and finishing infusion according to infusion plan)
Goal:	User is able to start and stop infusion safely according to a minimal amount of actions (not temporarily) (De Jong, 2004)
Actor:	1. Operation theatre: anesthesiologist, anesthesia assistant 2. ICU: IC nurse, intensivist 3. Ward: Physician, nurse
Pre-condition:	Infusion pump has been started and required syringe according to infusion plan has been inserted. Now, infusion has to be started. User does this safely with a minimal amount of actions.
Post-Conditions	Infusion has been started In case medication has to be stopped according to infusion plan, user stops infusion safely with a minimal amount of effort.
Requirements:	User does not use infusion pump unintended User does not forget to start infusion User can clearly hear alarm User does not erase pump settings by mistake User relates pump settings to patient data User knows what to do in case of an alarm User sees infusion volume and speed simultaneously User immediately sees whether or not pump is infusing **User can distinguish decimals easily** User can select previous used settings User can titrate without stopping infusion User can set volume limit

Requirement 6	User can easily distinguish decimals
Type:	Functional
Description:	In order to infuse medication according to the right volume and speed, it is essential to enable users to distinguish decimals before and behind the comma.
Claim (evidence):	Theory: Decimals after the decimal point have to be depicted distinctively, e.g., larger or with a different color (Doyle, 2010; Thimbleby, 2010) and not too close to each other (Paul, 2010). Empirical: Contextual interviews, expert meeting, and paper prototyping
+	Number of errors reduce and mental effort decreases
-	User is not used working with decimal points displayed and therefore is confused, mental effort increases.
Use cases:	UC6: Start/ stop infusion

Figure 1. Example use case and accompanying requirement

The working prototype we used as the test interface involved a dynamic simulation that users could interact with and that stored user key presses (see Figure 2a). As reference, the interface of the Braun Perfusor® Space syringe pump was used (see Figure 2b) which is a commonly used infusion pump in Dutch hospitals. The reference interface provides a completely different implementation of the same display design principles that were used for the development of the new interface. Table 2 lists five basic display design principles and shows how the reference interface scores more poorly in terms of the design principles compared to the new design. These are hypotheses to be tested in the qualitative analysis part of our results.

(a) New interface

(b) Reference interface

Figure 2. Tested infusion pump interfaces

Table 2. Comparison of interface design between new and reference interface (categorization based on basic display design principles Wickens et al., 2004)

	New interface	*Reference interface*
Minimization of information access cost	Directly accessible, flat menu structure	Navigation through deep and broad menu structure
Usage of discriminable elements	Differing shape and size of buttons	Uniform shape and size of buttons
	Infusion speed and volume to be infused are displayed large, other information (like remaining time) is displayed in the righter display	Infusion speed and volume are displayed in different sizes, no other information such as remaining time displayed.
	Decimals are displayed significantly smaller than numbers before decimal point	Decimals are displayed in similar size as numbers before decimal point
	Lock-modus to prevent patients or cleaning staff to unintentionally push buttons and change settings	No lock-modus
Visibility & Legibility	Large rectangular touchscreen with a large font covering majority of interface	Narrow rectangular shape with small screen covering left part of interface, small font (no touch screen)
	Syringe is visible (placed under interface with etiquette displaying barcode and name of medication)	Syringe invisible (placed behind interface)
Distinctiveness	Buttons have stand-alone functions, no multi-functionality, e.g., two separate on and off-buttons, separate buttons for manual and automatic bolus	Buttons are multi-functional, e.g., combined on- and off-button, no separate buttons for manual and automatic bolus

	Only relevant buttons are displayed	All buttons are continuously presented to practitioner
Consistency	Button functions remain consistent throughout use	Button functions differ depending on the context of operations

* Alarm function also differed regarding visibility and legibility between new and reference interface. Since we did not test the alarm function in the experiment, we do not elaborate on the alarm function here.

2.1 Experimental design

A 2 x 3 (type of interface x session) within-subjects design was employed. Each participant had to accomplish three task sets with both interfaces, each task set consisted of eight tasks. Thus, participants had to accomplish a total of six task sets. Each task session comprised two task sets; one regarding the new interface and one regarding the reference interface. All three task sessions were conducted directly after each other. Tasks were similar regarding the critical operations they tested and only differed in the precise medication doses that had to be administered. In order to control for order effects and systematic biases, the order of interaction with interfaces, task variations and tasks themselves was counterbalanced using a standardized protocol (see Table 3). Interaction with interfaces was counterbalanced by alternating the type of interface (new or reference) the participant started the experiment with. Each of the three succeeding measures included two task sets (for the new and control interface respectively). The sets of task variations were rotated; each participant completed each of the three task set variations with both interfaces, regardless with which interface they started. This, in turn, enabled a comparison between participants starting with either the reference or new interface.

Table 3. Research design for counterbalancing the start of interface and task set variations

Start interface	*Measure*	*Task set variations*	*Order by task number*	*Interface*
Reference	1	1	12345678	Reference
		2	14762538	New
	2	3	13652478	Reference
		1	12345678	New
	3	2	14762538	Reference
		3	13652478	New
New	1	1	12345678	New
		2	14762538	Reference
	2	3	13652478	New
		1	12345678	Reference
	3	2	14762538	New
		3	13652478	Reference

In order to control for order effects *within* task set variations, the sequence of individual tasks was manually randomized for each set of task variations. An exception were the first (starting the infusion pump) and the last task (stopping the infusion pump). As they naturally occur in the beginning and ending, we did not include them in the randomizing order.

2.2 Sample

The sample consisted of 25 nurses (20 female, 5 male) from both the General Care Unit (GCU) (N=13) and the Intensive Care Unit (ICU) (N=12). Experience with infusion pumps ranged from zero to 31 years (M = 15.2, SE = 1.92). Frequency of use of infusion pumps varied from zero up to more than four times a day. Participation was voluntary and recruiting took place by means of the non-probability snowball sampling technique. Thus, a few participants were initially asked if they would like to volunteer. Interested participants then contacted us. After completing the study, they asked other nurses in their department that might be interested too. Only participants having zero experience with the *Braun Perfusor® Space syringe pump* were included in the sample, thus equating prior experience with the two interfaces. The Braun Perfusor is a widely used pump in Dutch hospitals, implying that only a limited number of hospitals could be identified where users had no experience with this particular pump.

2.3 Tasks and use scenarios

Eight tasks were selected, based on use cases resulting from the situated Cognitive Engineering-methodology (see Table 1). Inserting and removing syringe could not be simulated in the task as, our new interface was not incorporated in a physical infusion pump, but presented to respondents on a touch screen tablet (see paragraph 2.4). For every task, three variations were created, that concerned the same user activities and interface functions, but differed in their specific patient scenario and content (e.g., rate and type of medication). Different patient-scenarios were created for the two user groups (GCU and ICU), adapted to the respective work environment. Mostly, this concerned different type of medications and infusion rates, which were higher for participants of the ICU. Finally, tasks were combined into three task variation sets (per user group), allowing for within-subject repeated testing of interface functions. Table 4 shows the tasks related to optimal number of key strokes.

Table 4. Tested critical tasks and optimal numbers of keystrokes

Task	*Content/tested function/ use case*	*Optimal number of keystrokes for each task**	
		New	*Reference*
1	Switching on the infusion (use case: switch pump on and off)	2	1
2	Adjusting values and starting infusion (use case: start and stop infusion)	5	10
3	Administration of a (manual) bolus while infusion is active (use case: bolus)	1	5
4	Adjusting infusion rate while infusion is active 1 (use case: start and stop infusion)	4	4
5	Adjusting infusion rate while infusion is active 2 (use case: start and stop infusion)	4	4
6	Retrieving diagnostic information: looking up the drug dosage that has been administered to patient (use case: drug group)	0	0
7	Administration of (automatic) bolus while infusion is active (use case: bolus)	1	5
8	Stopping and switching off infusion (use case: pausing infusion, switching pump on and off)	2	2

**Note*. The optimal number of keystrokes per task depends on the values entered. Keystrokes presented in the table are required to enter the following settings: rate 2 ml/2 hours; adjusting from 4 ml to 2 ml; administering bolus of 2ml. Task 1,7 and 8 do not require numeric adjustments.

2.4 Apparatus and experimental set-up

All sessions were recorded on video as support for the subsequent analysis. Interfaces were presented on a tablet (*Fujitsu StylisticQ550*, screen size 10.1 inches, 1280x800 pixels) in their original size and quality. Using the tablet's touchscreen, the participants could operate the interface and perform the given tasks. Pre-programmed tasks were loaded on an external laptop and were sent to the tablet via WLAN. Log files of the pressed keystrokes were saved on the tablet and later assessed for analysis.

2.5 Procedure

The study was conducted in an isolated room, at the hospital where the respective respondent was employed. Two researchers were present at each experimental trial: one was responsible for instructing the participant, the other for managing task representation on the tablet. The participant received general information about the experiment, informed consent and a non-

disclosure agreement. After signing the informed consent, the participant completed a pre-questionnaire concerning experience with infusion pumps and demographics. Then, a training video explaining the pump's basic functions was presented to the participant. Subsequently, the experiment started and the participant performed the first set of task variations. Each task was presented on a separate sheet of paper which was handed to the participant by the researcher. During the task objective performance measures were recorded. After every task, a self-report scale for mental demand was administered, in order to measure subjective mental workload. After completion of the first device, the training video of the second infusion pump interface was presented and the participant completed the second set of task variations with the second pump. With the exception of the training videos this procedure was repeated until each task set variation was completed with each interface (six measures in total). During this procedure the researchers did not engage in verbal conversation with the participant other than to provide task-related instructions. After completion of the experimental trial the session finished with a post-interview concerning the participant's preferences in use of both interfaces. Each experimental trial took about 90 minutes and all participants received a financial reimbursement of 50 Euro.

2.6 Measures

Objective performance measures concerned the number of successfully completed tasks, normative path deviations, number of keystrokes and time to completion. A completed task was scored as erroneous when the participants' executed operations did not achieve their intended outcomes, in agreement with Reason (1990). For quantifying the normative path deviations, a novel method was developed, that would also support identification of usability issues in a mixed-method setting. The method is described in detail in the following section. In addition to objective performance, subjectively experienced mental demand was measured using the RSME scale (*Rating Scale Mental Effort*), a one-dimensional anchored subjective workload scale (Zijlstra and Doorn, 1985). Ratings of invested effort are indicated by a cross on a continuous line. The line runs from 0 to 150 mm, and every 10 mm is indicated. Along the line, at several anchor points, statements related to invested effort are given (see Table 5). The scale is scored by measurement of the distance from the origin to the mark in mm. On the RSME the amount of *invested effort* into the task has to be indicated, and not the more abstract aspects of mental workload (e.g., mental demand, as in the NASA TLX). These properties make the RSME a good

candidate for self-report workload measurement. Previous research has shown that unidimensional scales are better in providing a global rating of workload, while being easier and quicker to administer (Hendy et al., 1993; Pickup et al., 2005). In addition, subjective preferences of individual functions of both interfaces were assessed by a structured post-interview.

Table 5. Subjective cognitive effort scale (RSME) – anchor points and statements

Anchor point	*Statement*
Between 0 and 10	Not strenuous at all
Between 10 and 20	Barely strenuous
Between 20 and 30	A bit strenuous
Between 30 and 40	Somewhat strenuous
Between 50 and 60	Considerably strenuous
Between 70 and 80	Fairly strenuous
Between 80 and 90	Very strenuous
Between 100 and 110	Very much strenuous
Between 110 and 120	Enormously strenuous

**Note*. Statements translated from Dutch by the authors

2.7 Coding scheme for normative path deviations

Derivation of normative path deviations was accomplished by conducting a sequence of analytical steps (see Figure 3). For this purpose, a standardized coding system for user actions was developed. Establishing the appropriate level of detail for the data presented a challenge and required a compromise between not losing too much relevant information but also not getting lost in irrelevant details.

Using goal-directed keystrokes as a performance indicator for medical devices was previously introduced by Nunnally et al., 2004, but this approach would not reveal any qualitative information about specific types of path deviations and how those differ between interfaces. As our approach aims to be feasible for mixed-methods analyses, a coding system that provided more qualitative information about specific path deviation was required.

The GOMS (*Goals, Operators, Methods, Selection rules)* model (see John and Kieras, 1996) served as the initial theoretical framework for the first step of data reduction. The GOMS model is frequently used for task analysis and user interface design- and evaluation in human-machine interactive systems (John and Kieras, 1996) and can be used to formally represent a task

by reducing observations to essential actions. Here, methods for accomplishing a task (goal) are a sequence of low-level user actions (‘operators’, e.g., ‘adjusting rate’).

Using this analysis framework, the user-device interaction is reduced to essential actions (‘operators’). During the behavioral analysis, distinct letters are assigned to each defined operator, such that the sequence of letters describes the *observed path*. The same set of letters are used to describe the *optimal path*, which is either the results of an expert analysis, or may also be derived from requirements documents or user manuals. Then the *Levenshtein algorithm* (Levenshtein, 1966) is applied on the two resulting letters strings for (1) the normative path solution of a task and (2) the observed path, which results in a metric for *deviation from the optimal path*. A strong deviation is taken as an indicator of inferior, or even faulty task completion by the user, and is used to for both quantitative as well as qualitative analyses.

2.7.1 Operators and methods (normative paths)

In order to achieve some formal representation of our tasks, we defined lower-level user actions (operators, e.g. ‘adjusting rate’) that, when ordered in a specific sequence (method), accomplish a task goal normatively. Operators were defined just unspecific for both interfaces, although more operators were defined for the reference interface, as this interface required more operators to accomplish task goals. This was amongst others due to its menu structure requiring navigation, as opposed to the new interface. As such, the methods to accomplish tasks with the new interface can mainly be seen as a subset of the methods needed to accomplish tasks with the reference interface.

Operators were defined on the basis of actions the users needed to perform in order to accomplish the task. Hence, when observing, we summarized obtained keystrokes to the distinct operators defined before. For example, when defining the operator ‘adjusting rate’, all pressed keystrokes belonging to those operators were added up and grouped under that operator. As such, adjusting the rate from zero to 6.1 resulted in seven keystrokes for the operator “adjusting rate” for the new interface. As the reference interface required the operation of the right-arrow in order to move to the decimal digit, eight keystrokes would be the minimum number of required keystrokes for the same operator. One virtual operator was added which referred to reading information from screen (thus, no keystrokes were required). In the next step, a distinct letter was assigned to each defined operator. Subsequently, the sequence of operators (i.e. method) for each

task and interface was established, resulting in a distinct letter string reflecting the respective method (in usability terms, normative path solution).

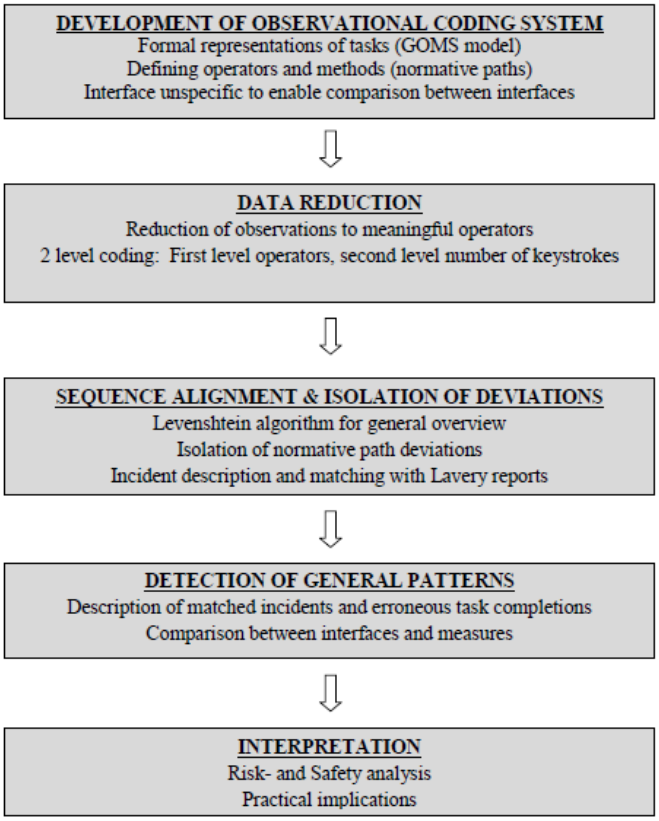

Figure 3. Analytical steps for derivation of normative path deviations

2.7.2 Keystroke-Level-Model

In addition to the sequence of operators we also reported the minimum number of required keystrokes at every step. This approach is similar to the *KLM* (*Keystroke-Level-Model*) variant of the GOMS techniques (John and Kieras, 1996) and thereby offers a quantitative, more simplified presentation of the obtained data. The result is a two-level coding system, consisting of a (1) operator level and (2) keystroke level. While the keystroke level represents efficiency of use, the operator level just captures deviations and is therefore a measure (see next section) for proneness to err, as well as an indicator for areas that might need further improvement.

2.7.3 Levenshtein distance as a deviation measure

For comparing normative paths with the actual interaction, the observed sequences were compared to the letter sequences of the normative paths of the same task (“sequence alignment”)

by applying the Levenshtein algorithm, similar as in Guan et al., 2006. The Levenshtein algorithm measures the distance between two sequences by counting the minimal number of edits necessary to transform one sequence into the other (Kruskal, 1983; Levenshtein, 1966). Thus, a Levenshtein distance of zero means that the participant's path equals the normative path, and that the higher the Levenshtein distance the more deviations from the normative paths occurred.

The Levenshtein distance comprises three types of edits: insertions, deletions and substitutions. For example, for the task of administering a bolus of three milliliters, the normative sequence (method) for the reference interface would have been:

Method: H I J K

Number of keystrokes: 1 1 3 1

(H = Selection of bolus mode, I = selection of bolus volume adjustment mode, J = Adjusting bolus volume, K = Administration of bolus). A frequently observed sequence was the following:

Method: H I J L

Number of keystrokes: 1 1 3 1

(L = Stopping the pump, i.e. the task was completed erroneously as no bolus was supplied).

In this example, the total Levenshtein distance would be 1, as it is required to execute one edit: substituting L with K.

Noteworthy, the Levenshtein distance does not indicate whether a task was accomplished correctly or erroneously, for even sequences with many deviations can result in a correct outcome. Deviations from the normative path were also used as guidance for the qualitative analysis, as this allowed for efficient identification of potentially critical incidences (those with strong deviations). Specific types of deviations and errors were isolated and described using structured report forms (Lavery et al., 1997).

2.8 Data analysis

Type and number of errors and normative path deviations, number of keystrokes and completion times were assessed via log files and post hoc video analyses. Mental effort was analyzed by means of the outcomes of the RSME.

The data set comprised a complex repeated measures structure. Therefore, inference was based on generalized linear mixed-effects models. Separate regressions were estimated for all performance parameters, with interface (β_{IF}) and session (β_S) as fixed effects. For objective efficiency measures (path deviations, completion time and keystrokes), only results from correctly completed tasks were included. Final errors were treated with logistic regression, whereas Poisson regression was used for normative path deviations and number of keystrokes. Using these models accounts for skewed residual distributions, non-constant variance structures and bounded response variables (Fox, 2008). Mental effort and completion time were modelled as Gaussian regressions. Because completion time residuals exhibited skew and heteroscedasticity, log-transformation was applied which effectively solved these problems. Participants and tasks entered the regressions as intercept random effects. All possible slope random effects were added, to establish the maximum random effects structure (Barr et al., 2013). Asymptotic estimation procedures for mixed-effects models are known to be inaccurate (Bolker et al., 2009), especially for small sample sizes. Therefore, all regressions were estimated using the accurate MCMC procedure, as provided by the MCMCglmm command (Hadfield, 2010) of the statistical computing environment R (R Development Core Team, 2011). Due to that, the reported significance levels and credibility intervals referred to the posterior distribution of the respective parameter.

3 Results

3.1 Quantitative analysis

3.1.1 Errors

With the reference interface, 426 of 600 tasks were completed successfully (71%), compared to 457 with the new interface (76%). Using logistic regression, an overall training effect was observed for the transition from first (68.5% correct) to second session (75.8%), and to a lesser extent from the second to third session (76.5%). Despite being rather small, the training effect was confirmed by mixed effects logistic regression ($\beta_S = 18.6,\ 95\%\ \mathrm{CI}\ [4.7;\ 32.7]\ , p = .001$), whereas the differences between the two interfaces were not statistically significant ($\beta_{IF} = 37.1,\ 95\%\ \mathrm{CI}\ [-15.6;\ 101.3]\ ,\ p = .120$). See Figure 4 for an illustration.

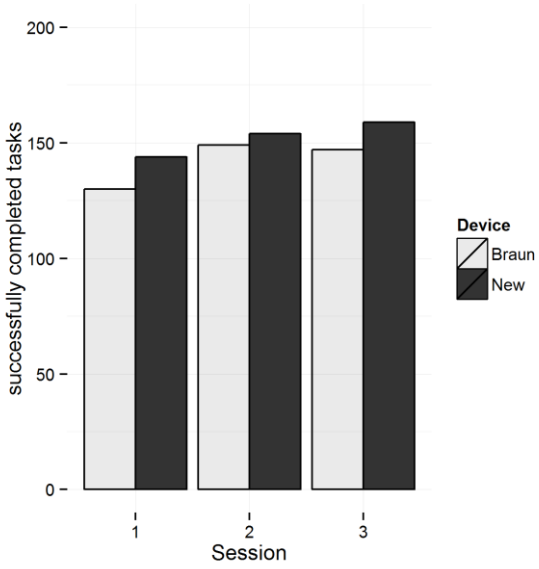

Figure 4. Successfully completed tasks for the two interfaces across sessions

3.1.2 Deviations from normative path

Overall, normative path deviations as measured by Levenshtein distance were $M = 2.05$ ($SD = 2.83$) for the reference interface and $M = 0.28$ ($SD = .67$) for the new interface. Poisson mixed-effects regression confirms that normative path deviations were reduced by almost 90% with the new interface ($\beta_{IF} = -2.17,\ 95\%\ \text{CI}\ [-2.94; -1.39]\ ,\exp(\beta_{IF}) = .11,\ p < .001$).

In addition, a training effect was observed, with the initial Levenshtein distance of $M = 1.50$ ($SD = 3.01$) reducing to $M = 1.09$ ($SD = 1.84$) for the second and $M = .86$ ($SD = 1.44$) to the third session. Regression results confirm that with every new repetition the distance is reduced by almost one quarter ($\beta_S = -.26,\ 95\%\ \text{CI}\ [-.37; -.15]\ ,\exp(\beta_S) = .77,\ p < .001$).

3.1.3 Completion times

Average task completion was reduced by 42% with the new interface ($M = 14.5s, SD = 14.8$) compared to the reference interface ($M = 24.8s, SD = 24.8$). Mixed-effects regression (on log-transformed completion time) confirms the improvement ($\beta_{IF} = -.69, 95\%\ \text{CI}\ [-1.28; -.13]\ , p = .022$). In addition, overall completion time decreased from the first session ($M = 28.4s, SD = 27.5$) to the second ($M = 17.1s, SD = 16.6$) and to the third ($M = 13.8s, SD = 14.1$). This is confirmed by regression results ($\beta_S = -.38, p < .001$).

3.1.4 Keystrokes

Next to completion time, efficiency of operation was measured by number of single keystrokes to complete a task. Similar to completion times, number of keystrokes were reduced by 40% for the new interface ($M = 7.0, SD = 8.9$) as compared to the reference interface ($M = 11.6, SD = 12.7$). However, this effect did not reach statistical significance in the Poisson regression ($\beta_{IF} = -.75, 95\%$ CI $[-1.69; .18]$, $\exp(\beta_{IF}) = .47, p = .11$). Again, a training effect was observed, with initial $M = 10.3$ keystrokes ($SD = 11.9$) being reduced to $M = 9.0$ (SD $= 11.5$) in the second session and $M = 8.5$ ($SD = 9.9$) in the third session. Despite not appearing very strong, this effect was statistically highly significant ($\beta_S = -.11, 95\%$ CI $[-.16; -.06]$, p $<$ .001).

3.1.5 Mental effort

Objective efficiency measures were complemented by one subjective measure: reported mental effort. Mental effort scores with the new interface ($M = 9.9, SD = 16.2$) were 39% lower than with the reference interface ($M = 16.2,$ SD $= 18.9$). This effect was confirmed by the regression ($\beta_{IF} = -6.46, 95\%$ CI $[-10.58; -2.30]$, $p = .003$). Participants reported lower mental effort with every repetition, with initial $M = 18.0$ ($SD = 21.1$) reducing to $M = 12.3$ ($SD = 16.2$) in the second and $M = 8.9$ ($SD = 14.8$) in the third session. This effect was statistically significant ($\beta_S = -4.75, 95\%$ CI $[-6.18; -3.35]$, $p < .001$).

3.1.6 Subjective preferences

The new interface was preferred by the majority of participants: 83.3% (N = 10) of the ICU and 69.2% (N=9) of the GCU employees preferred using the new interface, accounting for a total of 76%. Most frequently mentioned reasons for preferences concerned a clear overview on possible modes and visibility of the syringe (N=16), easiness and directness in use (N=16), button distinctiveness (N=9), and easy information access by lack of deep menu structure (N=6). These were precisely the design aspects that differed between the new and reference interface (see Table 2).

3.1.7 Additional observations

For the above regression models, the most simple fixed effects structure was chosen, with only one main effect for interface and one linear coefficient for training effects across the sessions. In fact, we have explored three variants for all five outcome variables: non-linear training effects, different training rates between the two interfaces, and differences between professional user groups. We compared the added fixed effects against the default model using the deviance information criterion (DIC). The DIC is a measure of model fit that penalizes model complexity and is commonly used for model comparison (Bolker et al., 2009). When two models are compared, a smaller DIC value indicates better model fit, with a decrease by five being roughly equivalent to a significance level of .05 in asymptotic omnibus tests (Spiegelhalter et al., 2002).

First, we modelled the training effect as a linear coefficient, albeit training curves typically are non-linear, with performance asymptotically leaning towards a maximum (for example, zero errors). Replacing the linear regressor β_S with a factor allows the slope to vary freely between subsequent sessions, albeit at the cost of an additional parameter. For all outcome variables, the linear coefficient fitted the data better and was retained. Second, faster initial learning does not necessarily lead to better performance. Potentially, the new interface could be more intuitive at first use, but less efficient when users gain more experience. This can be modelled as an interaction effect between device and sessions. However, no relevant or statistically significant interaction effects were observed for any of the outcome variables. Third, professional groups (intensive care, general care) could differ in overall performance, which we modelled as an additional factor, again without any relevant effect.

3.2 Qualitative analysis

In order to gain information about the specific type and consequence of normative path deviations and errors, these were further analysed utilizing a qualitative approach. By using structured report forms, single normative path deviations were isolated. Thus, individual keystrokes that deviated from the normative path solution as revealed by the Levenshtein distance were classified into patterns of deviations, e.g., ‘confusion between starting the infusion and confirming settings’. Thus, individual keystrokes were summarized and coded into distinct patterns of deviations. We mean these patterns of deviations when referring to ‘normative path deviations’ in the following qualitative analysis. We coded a total of N = 230 normative path

deviations for the new interface, and a total of N = 672 normative path deviations for the reference interface. In Table 6 the number of coded normative path deviations and final errors are listed per tested function, across user groups. The relation between deviations and errors may be read as followed: the number of errors (e.g., 6 in task 2 with the reference interface, measure one) indicates how often the respective 38 deviations resulted in an error. For example, in the first measurement, 38 normative path deviations occurred and six of these resulted in a final error. Hence, 32 deviations did not result in a final error.

Table 6. Overview number of errors and normative path deviations for the three succeeding task runs (measure 1-3).

Interface		*Measure 1*	*Measure 2*	*Measure 3*
Switching on infusion (task 1)				
New	Deviations	25	2	2
	Errors	3	2	2
Reference	Deviations	32	27	19
	Errors	5	5	4
Adjusting settings and starting infusion (task 2)				
New	Deviations	10	3	1
	Errors	7	1	0
Reference	Deviations	38	32	21
	Errors	6	2	5
Adjusting settings while infusion is in progress (task 4 and 5)				
New	Deviations	22	16	20
	Errors	0	0	0
Reference	Deviations	27	26	22
	Errors	8	9	6
Finding diagnostic information (task 6)				
New	Deviations	17	16	15
	Errors	17	16	15
Reference	Deviations	31	23	20
	Errors	21	14	13
Stopping and switching off infusion (task 8)				
New	Deviations	0	0	0

	Errors	0	0	0
Reference	Deviations	25	24	21
	Errors	1	1	1

*_Note_. The switch-on function of the reference interface could not be tested, which is why the standby function of the reference interface was used for comparison.

Normative path deviations and errors occurring with the new interface were sorted by their severity, defined as frequency and risk of harm. Only recurrent errors and normative path deviations not decreasing by more than 70 percent between measures are presented. Thus, in this table we only present normative path deviations that kept occurring within each of the three repeated measurements. We decided to summarize the qualitative data for this article in this way in order to focus on the most prevalent found deviations and reduce the amount of qualitative data obtained in the original study. Table 7 provides a description of the type of deviation, its design-related cause, clinical consequences and Human Factors design aspects (Wickens et al., 2004) involved.

Table 7. Description of frequently found deviations with the new interface

Function/ use case	*Description*	*Cause*	*Consequence*	*Design aspect (*Wickens et al., 2004), see also Table 2
Finding diagnostic information	Retrieving wrong diagnostic information (volume to be infused instead of delivered volume)	Terminology, not clear what TIV (Te Infunderen Volume= Volume To Be Infused) means.	Wrong clinical information is administered	Visibility and legibility
Bolus Administration	Repeated administration of automatic bolus	Lack of control and diagnostic feedback	Drug overdoses	Visibility and legibility: extensive user feedback is presented in right-hand display
	Adjusting main settings before administration of bolus	Not clear enough that bolus function stands for itself	Main settings are incorrect after bolus administration	
	Confusion manual/automatic bolus	Indistinctiveness of bolus functions	Wrong bolus volume may be administered when automatic bolus was set at a different value	Distinctiveness: separate buttons for manual and automatic bolus

Adjusting main settings while infusion is active/ start and stop infusion	Infusion is re-started although already active	Visibility of system state not sufficient	No consequences; repeatedly executing the Start-function does not change state of infusion.	Visibility and legibility

Normative path deviations and errors occurring with the reference interface especially concerned button indistinctiveness (design principle 'distinctiveness', see Table 2) and multi-functionality (design principle 'consistency', see Table 2). The function of starting the infusion was frequently confused with the confirmation of adjusted values executed by means of the OK button, reflecting that the Start- and OK function are not distinctive enough. Further, due to one multifunctional button, participants frequently stopped the infusion in an attempt to re-start it when adjusting settings while an infusion was active. Indistinctiveness of button functions also affected bolus administrations. Frequently, participants tried to administer boluses via the OK- or Start-infusion function, thereby resulting in no bolus administrations, delays in supply and interrupted infusions. By violating visibility and legibility-guidelines (Wickens et al., 2004) concerning screens/menus and graphics, both (1) the little screen containing a high amount of visual information, displayed in monochrome colors and (2) the physical buttons on the interface's right part which are consistently visible contributed to aforementioned performance outcomes. The retrieval of wrong diagnostic information occurred frequently with both interfaces.

4 Discussion

The aim of this study was to investigate if a novel method of testing infusion pump interfaces, in conjunction with a situated Cognitive Engineering method for designing such interfaces, would address various shortcomings of earlier testing methods of interfaces. We have done this by addressing methodological shortcomings of existing usability research and studies on the interaction between practitioners and infusion device technology: lack of process-tracing techniques, lack of repeated measurements to reveal learning effects, and lack of combination of subjective and objective measures (Hornbæk, 2006; Schraagen and Verhoeven, 2013).

Results showed that the new interface outperformed the reference interface: especially with regards to completion times and normative path deviations, as well as perceived mental effort, numbers were significantly reduced. These findings are in line with previous research

conducted by Garmer, 2002, Lin et al., 1998 and Syroid et al., 2012. Still, the rate of errors (24%) and path deviations with the new interface remained high. This reflects that application of the situated Cognitive Engineering method utilized in this study does not eliminate use-related hazards completely. However, it provides very specific directions for improvement (see Table 8) that need to be addressed prior to clinical implementation.

By utilizing a qualitative approach, this study revealed differences between interfaces concerning the occurrence of specific normative path deviations and errors. Particularly design choices regarding visibility/ legibility and distinctiveness (Wickens et al., 2004) appeared to make a significant difference between reference and test interface. Based on our situated Cognitive Engineering design process, we decided to let buttons have stand-alone functions rather than being multi-functional, displaying only relevant buttons on the touchscreen rather than continuously confronting users with all buttons, and using a large rectangular touchscreen rather than a narrow screen (see Table 2). Possibly, these were the design aspects that reduced the number of errors and normative path deviations in the new interface.

Whereas errors and normative path deviations in our study primarily concerned visibility/legibility and distinctiveness (in line with Garmer et al., 2002), previous research showed that mainly consistency was a usability problem (Graham et al., 2004; Zhang et al., 2003). This was not confirmed in our experiment, probably because we ensured the button functions to remain consistent throughout use in our new interface in contrast to variable button functions in our reference interface.

As a review of frequently employed methods for studying medical device technology indicated a lack of 'process tracing techniques' (Schraagen and Verhoeven, 2013), we introduced a novel, replicable method for a standardized representation of the user's task completion process. The results of our study suggest the feasibility of our analytical approach, which is especially practical for mixed-methods approaches and comparative usability validation studies. The GOMS model provided an unspecific representation of the interaction with both interfaces, thereby enabling a direct comparison between both devices. Coding the interaction with interfaces on the basis of operators made feasible to identify the location of use-related hazards, which could be more easily isolated in the subsequent qualitative analysis. Thus, in contrast to previous research, the present study distinguished between normative path deviations in the process of task completion and erroneous tasks. Comparing these two measures, the difference in normative path deviations seems much more pronounced, and the number of deviations declined

faster with repeated exposure. In addition, normative path deviations proved effective in identifying interface issues that may cause use-related hazards. The frequency of these normative path deviations is an important indicator for an interactive system's safety; although not necessarily resulting in erroneous task outcomes, they greatly reflect the stability of an interactive system. This is especially crucial in clinical real world conditions characterized by interruptions and time constraints (Nunnally et al., 2004). Under such stressful conditions it is likely that normative path deviations easily occur. Although no erroneous input might be executed initially, a correction of a keystroke is challenging under time pressure. A stable interaction offers more resilience against time constraints and interruptions, typical characteristics for fast-paced environments such as the operating room or intensive care unit (Nunnally et al., 2004).

Contrary to previous research, we additionally considered learning effects. Our results demonstrate that performance with both infusion devices improves with repeated interaction. The fact that performance with the new interface remains higher at all points of measurement, indicates that it will reach better maximum performance in the long term. Therefore, this study provides stronger evidence for superior performance of the new interface.

5 Limitations and recommendations

This study has some limitations that need to be mentioned. First, we would like to address a limitation that is inherent to case study research. Since this study reports only one case study involving one experimental and one control interface and a relatively small sample, conclusions with respect to the effectiveness of the situated Cognitive Engineering approach cannot be validly drawn. For a stronger claim in favor of situated Cognitive Engineering, future research should compare design outcomes of this method with other established usability methods, such as heuristic evaluation or cognitive walkthrough. If technologies designed with the situated Cognitive Engineering method show enhanced performance repeatedly when compared to the other methods, a stronger claim in favor of this method can be made.

A second limitation is that some use cases could not be tested: inserting and removing a syringe, occlusion and alarms. Although these functionalities were all three integrated in the interface design, they were not fully optimized to be incorporated in the testing. Another limitation involved the artificial testing environment. Users were confronted with critical tasks they had to accomplish in an isolated, unthreatening situation. This study cannot account for how

users would perform in real healthcare settings characterized by distractions, operation of multiple devices and time pressure. For addressing those limitations in future studies, we recommend using a more realistic, less isolated testing environment, the inclusion of alarms and the operation of multiple infusion interfaces.

A further limitation with regard to the identification of user requirements is that this activity relied to a large extent on interviews with end-users and experts. Their understanding of the medical device under investigation and its requirements may possibly be incomplete and biased. However, we regard the input of actual end-users of the medical device under study as the most relevant source of information. Although they may have a biased understanding, they are the sharp-end practitioners (in our sample, on average 15 years of experience) that have to interact with that particular device on a daily basis. As such, they are the ones who are most aware of the device's requirements, and how the interface design can be optimally embedded in their clinical environment. Moreover, the situated Cognitive Engineering approach describes interviewing as the most appropriate method to detect requirements.

As far as recommendations are concerned, for future studies on similar topics, we recommend the use of process tracing techniques such as the application of the Levenshtein algorithm. The resulting Levenshtein distance was feasible for combining mixed-methods approaches in order to study the interaction between practitioner and interface. Thereby, performance measures included both outcome and process measures. As the latter is usually achieved with qualitative analyses, this study presents a feasible impetus on how to include process measures in usability studies in a quantitative way. In this context we recommend that the link between normative path deviations and use-related hazards should be further investigated in future studies. Moreover, we advocate repeated measurements rather than single interactions between users and interface, as well as combining objective and subjective measures.

6 Conclusions

Using our novel testing method, we showed that (1) the inclusion of repeated measures is a relevant add-on for revealing learning effects and (2) both focusing on task outcomes and on process measures reveals a more realistic picture of the user-device interaction and the high-risk system's stability.

This study also showed that following a standardized Human Factors design approach does not automatically result in an exhaustive detection of usability problems (see also Schmettow et al., 2013) and a complete elimination of use errors.

Acknowledgements

Anita Cremers is gratefully acknowledged as participating investigator for specifying user requirements. Cor Kalkman served as scientific advisor and critically reviewed the proposal. Anita Aarts designed the new interface. Bert Bierman implemented both interfaces. Frauke van Beek and Jan Sommer were instrumental in collecting data. Rutger van Merkerk managed the project. This work was supported by the Pieken in de Delta-program of the Ministry of Economic Affairs, Agriculture and Innovation, and by the city of Utrecht and the province of Utrecht (grant number PID 101060).

References

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. J. Mem. Lang. 68, 255–278. doi:10.1016/j.jml.2012.11.001

Besnard, D., Cacitti, L., 2005. Interface changes causing accidents. An empirical study of negative transfer. Int. J. Hum. Comput. Stud. 62, 105–125. doi:10.1016/j.ijhcs.2004.08.002

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.-S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24, 127–35. doi:10.1016/j.tree.2008.10.008

Center for Devices and Radiological Health, U.S. Food and Drug Administration, 2010. Infusion Pump Improvement Initiative.

Center for Devices and Radiological Health, U.S. Food and Drug Administration, 2011. Applying human factors and usability engineering to optimize medical device design. Draft guidance.

Fox, J., 2008. Generalized Linear Models, in: Applied Regression Analysis and Generalized Linear Models. SAGE Publications, pp. 379 – 424.

Furniss, D., Masci, P., Curzon, P., Mayer, A., Blandford, A., 2014. 7 Themes for guiding situated ergonomic assessments of medical devices: A case study of an inpatient glucometer. Appl. Ergon. 45, 1668–1677. doi:10.1016/j.apergo.2014.05.012

Garmer, K., 2002. Application of usability testing to the development of medical equipment. Usability testing of a frequently used infusion pump and a new user interface for an infusion pump developed with a Human Factors approach. Int. J. Ind. Ergon. 29, 145–159. doi:10.1016/S0169-8141(01)00060-9

Garmer, K., Liljegren, E., Osvalder, A.-L., Dahlman, S., 2002. Application of usability testing to the development of medical equipment. Usability testing of a frequently used infusion pump and a new user interface for an infusion pump developed with a Human Factors approach. Int. J. Ind. Ergon. 29, 145–159. doi:10.1016/S0169-8141(01)00060-9

Graham, M.J., Kubose, T.K., Jordan, D., Zhang, J., Johnson, T.R., Patel, V.L., 2004. Heuristic evaluation of infusion pumps: implications for patient safety in Intensive Care Units. Int. J. Med. Inform. 73, 771–9. doi:10.1016/j.ijmedinf.2004.08.002

Guan, Z., Lee, S., Cuddihy, E., Ramey, J., 2006. The validity of the stimulated retrospective think-aloud method as measured by eye tracking, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '06. ACM Press, New York, New York, USA, p. 1253-1262. doi:10.1145/1124772.1124961

Hadfield, J., 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J. Stat. Softw. 33, 1–22. doi:10.1002/ana.23792

Hendy, K., Hamilton, K., Landry, L., 1993. Measuring subjective workload: When is one scale better than many? Hum. Factors 35, 579–601. doi:10.1177/001872089303500401

Hornbæk, K., 2006. Current practice in measuring usability: Challenges to usability studies and research. Int. J. Hum. Comput. Stud. 64, 79–102. doi:10.1016/j.ijhcs.2005.06.002

John, B.E., Kieras, D.E., 1996. Using GOMS for user interface design and evaluation: which technique? ACM Trans. Comput. Interact. 3, 287–319. doi:10.1145/235833.236050

Kruskal, J.B., 1983. An overview of sequence comparison: Time warps, string edits, and macromolecules. SIAM Rev. 25(2), 201–237

Lavery, D., Cockton, G., Atkinson, M.P., 1997. Comparison of evaluation methods using structured usability problem reports. Behav. Inf. Technol. 16, 246-266. doi:10.1080/014492997119824

Levenshtein, V.I., 1966. Binary Codes capable of correcting deletions, insertions, and reversals. Sov. Phys. Dokl. 10, 707–710.

Lin, L., Isla, R., Doniz, K., Harkness, H., Vicente, K.J., Doyle, D.J., 1998. Applying human factors to the design of medical equipment: patient-controlled analgesia. J. Clin. Monit. Comput. 14, 253–63.

MacKenzie, I.S., Zhang, S.X., 1999. The design and evaluation of a high-performance soft keyboard, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: The CHI Is the Limit. ACM, pp. 25–31.

Neerincx, M.A., Lindenberg, J., 2008. Situated cognitive engineering for complex task environments, in: Schraagen, J.M., Militello, L.G., Ormerod, T., Lipshitz, R. (Eds.), Naturalistic decision making and macrocognition. Ahsgate Publishing Limited, Aldershot, England, pp.373-389.

Nunnally, M., Nemeth, C.P., Brunetti, V., Cook, R.I., 2004. Lost in menuspace: User interactions with complex medical devices. IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans 34, 736–742. doi:10.1109/TSMCA.2004.836780

Obradovich, J.H., Woods, D.D., 1996. Users as designers: How people cope with poor HCI design in computer-based medical devices. Hum. Factors 38, 574–92.

Pickup, L., Wilson, J.R., Norris, B.J., Mitchell, L., Morrisroe, G., 2005. The Integrated Workload Scale (IWS): A new self-report tool to assess railway signaller workload. Appl. Ergon. 36, 681–693. doi:10.1016/j.apergo.2005.05.004

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Reason, J., 1990. Human Error. Cambridge University Press, New York.

Schmettow, M., Vos, W., Schraagen, J.M., 2013. With how many users should you test a medical infusion pump? Sampling strategies for usability tests on high-risk systems. J. Biomed. Inform. 46, 626–641. doi:10.1016/j.jbi.2013.04.007

Schraagen, J.M., Verhoeven, F., 2013. Methods for studying medical device technology and practitioner cognition: the case of user-interface issues with infusion pumps. J. Biomed. Inform. 46, 181–95. doi:10.1016/j.jbi.2012.10.005

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B Stat. Methodol. 64, 583–616. doi:10.1111/1467-9868.00353

Syroid, N., Liu, D., Albert, R., Agutter, J., Egan, T.D., Pace, N.L., Johnson, K.B., Dowdle, M.R., Pulsipher, D., Westenskow, D.R., 2012. Graphical user interface simplifies infusion pump programming and enhances the ability to detect pump-related faults. Anesthesia & Analgesia 115, 1087–1097. doi:10.1213/ANE.0b013e31826b46bc

Vicente, K.J., Kada-Bekhaled, K., Hillel, G., Cassano, A., Orser, B. A, 2003. Programming errors contribute to death from patient-controlled analgesia: Case report and estimate of probability. Can. J. Anaesth. 50, 328–32. doi:10.1007/BF03021027

Vincent, C.J., Li, Y., Blandford, A., 2014. Integration of human factors and ergonomics during medical device design and development: It's all about communication. Appl. Ergon. 45, 413–419. doi:10.1016/j.apergo.2013.05.009

Wickens, C.D., Lee, J.D., Liu, Y. Gordon-Becker S.E., E., 2004. An Introduction to Human Factors Engineering, second ed. Pearson Education, New Jersey.

Zhang, J., Johnson, T.R., Patel, V.L., Paige, D.L., Kubose, T., 2003. Using usability heuristics to evaluate patient safety of medical devices. J. Biomed. Inform. 36, 23–30. doi:10.1016/S1532-0464(03)00060-1

Zijlstra, F.R.H., Doorn, L., 1985. The construction of a scale to measure subjective effort. Delft, Netherlands Delft Univ. Technol. Dep. Philos. Soc. Sci.