Behaviour & Information Technology
Vol. 28, No. 2, March–April 2009, 121–137

Taylor & Francis
Taylor & Francis Group

The theoretical foundation and validity of a component-based usability questionnaire

W.-P. Brinkman[a]*, R. Haakma[b] and D.G. Bouwhuis[c]

[a]*Brunel University, Uxbridge, UK, and Delft University of Technology, Delft, The Netherlands;* [b]*Philips Research Laboratories Eindhoven, Eindhoven, The Netherlands;* [c]*Technische Universiteit Eindhoven, Eindhoven, The Netherlands*

(*Received October 2005; final version received February 2007*)

Although software engineers extensively use a component-based software engineering (CBSE) approach, existing usability questionnaires only support a holistic evaluation approach, which focuses on the usability of the system as a whole. Therefore, this paper discusses a component-specific questionnaire for measuring the perceived ease-of-use of individual interaction components. A theoretical framework is presented for this compositional evaluation approach, which builds on Taylor's layered protocol theory. The application and validity of the component-specific measure is evaluated by re-examining the results of four experiments. Here, participants were asked to use the questionnaire to evaluate a total of nine interaction components used in a mobile phone, a room thermostat, a web-enabled TV set and a calculator. The applicability of the questionnaire is discussed in the setting of a new usability study of an MP3 player. The findings suggest that at least part of the perceived usability of a product can be evaluated on a component-based level.

Keywords: usability evaluation; questionnaire; component-based usability evaluation; technology acceptance model (TAM)

1. Introduction

Usability evaluations, such as field studies, heuristic evaluations and usability tests, are often mentioned in textbooks on developing usable interactive systems (Nielsen 1993, Preece *et al*.1994, Love 2005). They suggest that engineers develop their system or prototype and examine its usability to understand potential points for improvement. Asking users to evaluate the design with a questionnaire is often mentioned in these books. In the field, however, usability professionals rate questionnaires and surveys among the bottom of methods they use or have used (Gulliksen *et al*.2004, Bark *et al*. 2005, Mao *et al*. 2005). To put this into perspective, 97% of 197 respondents in a survey (Bark *et al*. 2005) among HCI practitioners in Nordic countries had used a user test to evaluate a product, whereas 53% had used a questionnaire for this. Although this is still a substantially large group, practitioners rate the usefulness or importance relatively low. For example, 103 respondents who replied to a survey among attendees of the ACM CHI '2000 conference and members of the Usability Professional Association (UPA) rated surveys in the bottom five of important methods (Mao *et al*. 2005), and in the Nordic survey (Bark *et al*. 2005) respondents rated questionnaire surveys again among the least useful evaluation methods. Among other factors, Bark *et al*. (2005) suggest that practitioners might perceive a questionnaire as less useful because it is unable to give them sufficient detailed information. The source for this seems a mismatch between the focus of questionnaires and the interests of engineers. Existing usability questionnaires regard the system as a single entity and provide information only on this level. Engineers, on the other hand, often use a compositional system view; they regard the system as a synergy of components. They look at a word processor as a system built from several components, such as a menu, a text editor, a dictionary and a spell checker. This compositional engineering view is emphasised in engineering approaches such as component-based software engineering (CBSE). Instead of developing a system from scratch, this approach assembles a new system from already existing software parts. The approach reduces the complexity of large software projects and improves the maintenance and reliability of a system (Cox 1990), thereby hoping to reduce development cost and time (Aykin 1994). The approach is not new. McIlory (1979) already introduced the concept at the first conference on software engineering in 1968. He pointed at the inefficiency of software engineering when similar software parts had to be rewritten for different applications. He therefore

*Corresponding author. Email: w.p.brinkman@tudelft.nl

ISSN 0144-929X print/ISSN 1362-3001 online

DOI: 10.1080/01449290701306510
http://www.informaworld.com

envisioned a catalogue of software components from which developers could choose and re-use software.

CBSE has also been applied to interactive systems. For example, design patterns, a general repeatable solution to a recurring design problem, allow interaction designers to re-use a confined design solution across a range of applications (Borchers 2001). Patterns, however, are only structured textual and graphical descriptions of conceptual solutions. Engineers can instead also rely directly on ready-made software components. Myers (1998) cites the Andrew project (Palay *et al*. 1988) as among the first to demonstrate the practical application of developing a system with user interface components. More theoretical oriented work has also been done on explaining and predicting human–computer interaction within a compositional view. For example, Taylor (1988) has proposed a layered interaction framework. He explained how users and components of a system interact across multiple layers. Several interaction mechanisms have been studied within this framework, such as a general protocol grammar (Taylor *et al*. 1999), diviplexing and multiplexing (Taylor and Waugh 2000), communication synchronisation (Taylor 1989) and layered feedback (Haakma 1999). A compositional usability evaluation method within this framework has also been proposed (Brinkman *et al*. 2004a, 2005). The method analyses recorded interaction behaviour, which it relates to the amount of effort users invested in using a component. To create a log file, however, access to the source code to insert recording instructions is required, something which might not always be possible. Ready-made components might, for example, only be available in their shielded binary format, or engineers might not have the time or resources to develop the recording mechanism. Alternatively, usability questionnaires would not be restricted by this. They can be applied without modifying the system. Ideally, such a questionnaire should help engineers identify components that cause usability problems and indicate the severity of these problems in relation to the overall usability of the system. However, holistic questionnaires are unable to do this. They tell engineers whether the entire system scores high or low on specific usability dimensions, such as mental load, consistency, learnability, flexibility, or navigational support. Although these are all essential, they do not lead engineers to specific parts of a system that need improvement or replacement, such as the menu, the text editor or the dialogue box of the mail merge function. Engineers might overcome this with an evaluation strategy of asking users to complete a series of small tasks. Examining the overall usability data collected in each of these tasks gives them an indication of the usability of the components used in a task. This approach has a number of drawbacks. First, it uses an indirect measure, which is affected by the other components the user interacted with in that task. Next, only looking at scaled-down tasks might not bring out usability problems users experience in a complex everyday task. Therefore, a component-based questionnaire would be welcome, which an engineer can give to users after they have attempted a realistic task, especially if engineers can use the questionnaire on a wide range of systems, independently or in combination with other usability evaluation methods, such as a usability test or field study.

This paper examines such a questionnaire that would allow engineers to study the usability of a part of the system. More fundamentally, the paper looks at whether users can have an attitude towards the usability of an individual part of the system, instead of only towards the entire system. Attitude is seen here as 'a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor' (Eagly and Chaiken 1993, p. 1). The following section will show that user–system interaction can be broken down into a series of user–component interactions, which allows users to establish an attitude about the usability of each component. After this, the paper will move on to discussing how these attitudes can be measured by using a questionnaire. The data collected from four experiments that used this questionnaire will be re-examined to study the validity and reliability of the questionnaire. The application of the questionnaire will be examined in the setting of a new experiment, which evaluates the usability of an MP3 player and its components. The paper concludes by discussing the limitations of the study and looking at future research directions.

2. Attitude formation on a compositional level

When setting up a component-based usability questionnaire, one of the first questions engineers need to address is what are the components of the system, and secondly what do users perceive as components of the system? Engineers might think about windows, scrollbars and dialogue boxes, while users might be more task-oriented, thinking about things such as writing a text, storing it or printing it. For a questionnaire to present valid but also useful information for engineers, these views need to coincide with one another. Or at least, some common ground of what constitutes a component is needed. Fortunately, Taylor's layered protocol theory (LPT) provides help here. It offers a framework that brings the user and engineer perspectives together. The framework describes a system as a collection of communicating components, a view not uncommon for software architectures of interactive

systems. Take, for example, MVC (model–view–controller; Krasner and Pope 1988), PAC (presentation, abstraction, control; Coutaz 1987), ICON (input configurator; Dragivevic and Fekete 2001) and the CNUCE agent model (Paternò 2000); they all have components that communicate with each other, but also with the users, by exchanging messages. This is a key concept. Within systems, these messages are embodied, for example by function calls a component made to other components. Crossing the system's boundary are messages from the system to the user, such as symbols displayed on the screen, or from the user to the system, such as pressing a button. Figure 1 illustrates this idea for an oven. Its compositional architecture is built out of a temperature, a timer, a display and a heater component. They are responsible for setting the temperature and the cooking time, displaying feedback, and controlling the heating element, respectively. LPT regards the heater component as operating on a higher-level layer of interaction. Its communication with the user is mediated by the temperature and timer component. Similarly, mental control processes are often also regarded as operating in a hierarchy of layers (Powers 1973, Norman 1984, Nielsen 1986, Vallacher and Wegner 1987, Newell 1990, Carver and Scheier 1998). Processes operating on high levels are more abstract, focus on the users' main goals, and set sub-goals for lower level processes. Low-level processes are more physical in nature, such as body movement coordination. They focus on sub-goals and report upwards on their completion. For instance, the control heater process in the oven example relies on the control timer and the control temperature process to transform its abstract actions, such as cooking food at 250 degrees for 20 minutes, to physical actions with the oven. The strength of LPT lies in the fact that it combines these two notions of hierarchy into a single interaction framework. It sees mental processes and software components operating in mirroring hierarchies, which it aligns into a number of interaction layers. For example, the control temperature process and temperature component are operating in a single physical layer where the message exchange takes place by physical actions such as pressing buttons or displaying digits. The control heater process and heater component that operate on a higher-level layer have a virtual message exchange. They send and receive messages, such as 'set temperature to 250', to and from the lower-level layer that establishes the actual message exchange on their behalf.

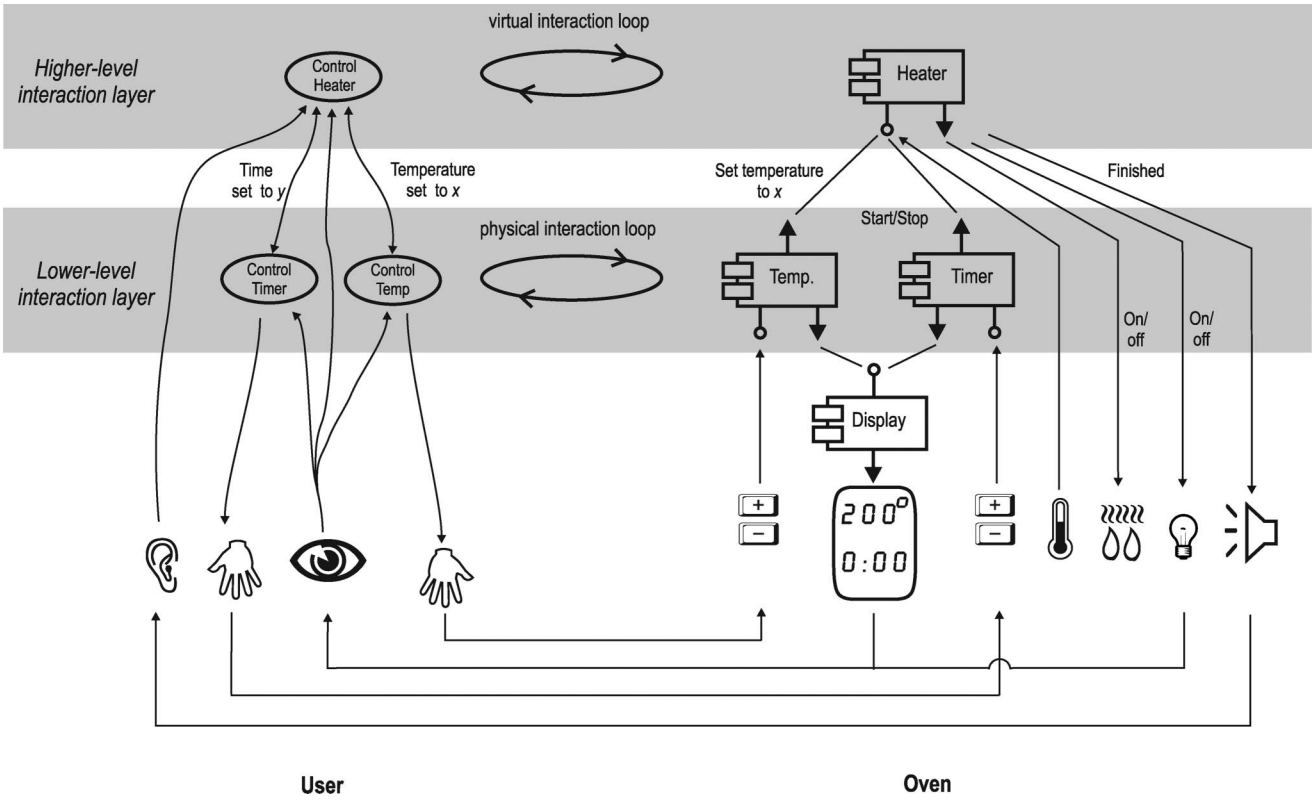

Figure 1. Layered interaction structure between a user and an oven.

The LPT framework illustrates the process in which user–system interaction takes place. It forms the foundation for the users' attitude formation towards individual components. First of all, it stresses that users can recognise distinct components in their interaction, which is of course a key requirement for asking them anything about components in a questionnaire. Users can perceive these components directly, such as the temperature component with its buttons and display, or indirectly, such as the heater component. But in both cases users can still perceive or deduce their behaviour and control them. This is essential when asking users about their attitude towards the *usability* of a component. Users need to associate their interaction experience with a component. They need to perceive a component as an independent interaction entity with a changeable state. Furthermore, the component should process users' input and inform users of its new state. With this two-way communication, users can control the behaviour of a component. They can perceive the component's response to their actions and compare this with their goal. In other words, the user and component are joined in a feedback loop, where users act upon the feedback from the component. For example, users would continue to press the upper '+' button to increase the temperature from 200 to 250 degrees (Figure 2) until the display actually shows 250 degrees. Their attitude towards the component's usability is therefore based on their experience when operating in this control loop. The ease with which users get a component in the desired state will determine how they will perceive the usability of that component. From an evaluation point of view, it seems therefore essential that a component has an independent, perceivable and controllable state, so users can tie their interaction experience with a specific part of the system. Software components that have these properties are referred to as interaction components (Brinkman *et al.* 2004a). In the case of the oven, the heater, the temperature, and the timer are interaction components, while the display is not. Its state is directly dependent on the state of the temperature and timer components. For users, it might be too difficult to separate the display from the other two components. Therefore, the usability of the display should be seen in the context of the interaction experience of the temperature or the timer component. These two components use the display to channel their feedback messages, making it part of their feedback loops, and hence their usability.

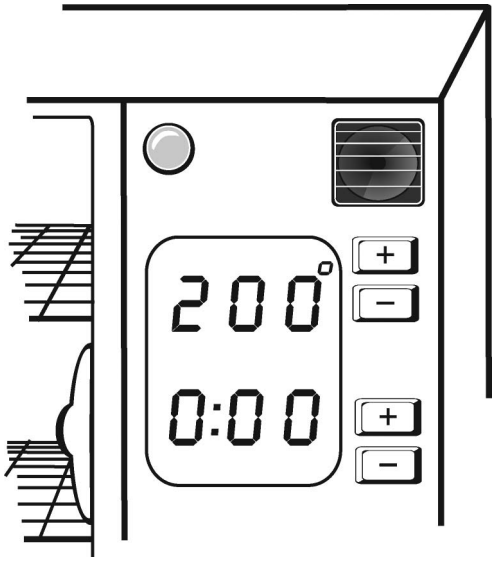

Figure 2. Part of the front of an oven.

3. Component-specific usability evaluation

The ISO's (1998) 9241-11 standard defines usability 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'. Within a compositional view, the product is decomposed into a set of joined interaction components. And as Figure 3 illustrates, users might have a different attitude towards the usability of each of these individual interaction components. Furthermore, how users perceive the usability of a component is based on their interaction experience with that component. The model is in essence similar, but more refined than the original ISO standard. Whereas the overall usability relates to the extent to which overall goals are achieved with a system, component-specific usability relates to the extent to which sub-goals are achieved with a specific component. Achieving these sub-goals will ultimately lead to achieving the overall goals. Therefore, the compositional model suggests that the users'

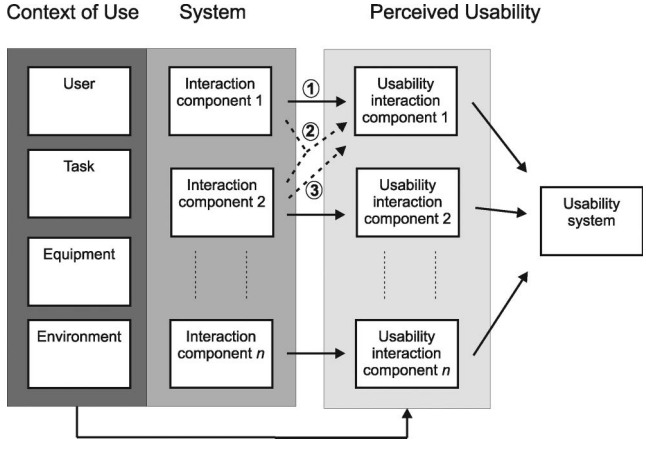

Figure 3. The compositional usability model.

attitude towards the overall usability is at least partly based on the attitude towards the usability of the individual components. The arrows in Figure 3 represent the relations between the various factors. The perceived usability of a component is first of all affected by the interaction component (1) in combination with the user, the task, the equipment, and the environment. However, the perceived usability of a component can also be affected by other interaction components (3), for example by the cognitive demand created by other components (Brinkman *et al*. 2004b). Furthermore, it can also be affected by a combination of interaction components (2), for example if components give users conflicting feedback (Brinkman *et al*. 2004c). Still, Figure 3 also shows that the users' attitude towards the entire system can be inferred once the users' attitude towards each individual component is established. In the example of the oven, if users perceive the usability of the temperature, the time and the heater interaction component as good, it is likely that they will perceive the usability of the entire oven also as good. Therefore, a validity criterion for any potential component-based usability questionnaire is that it should provide data to explain at least part of the overall usability. Not meeting this criterion means that engineers should not use a particular questionnaire because it provides no valid information about the usability of the system. This criterion will therefore be used later on in the validation section.

Figure 3 also shows the limitations of the conclusions that can be derived from attitudinal data on a compositional level. First of all, the users' attitude towards a component can be understood only within the context of the other components, the user group, the task, the equipment and the environment. In other words, placing a component in another context might change users' attitude towards it completely. This, though, is not a problem for engineers, who only want to know the usability of the components of a specific application. However, when confronted with a negative attitude towards a specific component, they need to consider the other components as well as the source of the problem. For example, if users find the temperature component difficult to use, engineers should look at the timer component as well. Its interaction protocol might be inconsistent and consequently also confuses users when interacting with the temperature component. Therefore, attitudinal data only gives an insight on how users perceive the situation, which of course could be a good starting point to understand a usability problem.

Although there are questionnaires that examine the usability of specific components, such as of a menu component (Norman 1991), a more generic questionnaire to examine all kinds of components would be more useful on a larger scale. It would be especially useful in situations with entirely new kinds of components. Two ways are possible to establish such a questionnaire: develop one from scratch or, more practical, adjust a well-established holistic questionnaire. The latter was applied in this project. Well-tested usability questionnaires (e.g. Chin *et al*. 1988, Davis 1989, Kirakowski and Corbett 1993, Lewis 1995, Brooke 1996, Lin *et al*. 1997, Kirakowski *et al*. 1998, Gediga *et al*. 1999, Hollemans 1999) were collected and examined for their potential as a component-specific questionnaire. Most questionnaires were unsuitable because they might be too lengthy, make reference to the appearance of the system, or are only relevant for a specific type of system, e.g. a web-based system (Kirakowski *et al*. 1998). Fortunately, the perceived usefulness and ease-of-use questionnaire (Davis 1989) provides a relatively small set of six questions (see Appendix 1) to capture well-formed beliefs about the ease of use of a system already after a brief initial exposure (Doll *et al*. 1998). These questions are statements which users rate on their likelihood. To make this a component-specific question, the name of the component is inserted in the statement on the place where normally the name of the system would be inserted. 'My interaction with the *temperature control* would be clear and understandable' would be an example of such a statement for the temperature component of the oven. In contrast to overall usability questions, component-based questions should help users to remember and reflect upon their interaction experience with a specific component (Coleman *et al*. 1985). It is, therefore, important that users can relate the name used in the statement, e.g. temperature control, with their actual interaction experience with this control. Users have to understand what is actually referred to in a question. Users can be helped by providing a clear description of a component alongside the questionnaire, or by providing pictures of the system, e.g. Figure 2, by pointing out or highlighting the component.

To measure the usability of a component, users would be asked to answer six questions. In the case of the oven example, the entire questionnaire would consist of 24 questions: six questions to measure ease-of-use of the oven and 18 questions to do the same for the three interaction components. To avoid systematic order effects, the 24 questions should be presented in random order. For example, the first statement a participant is asked to rate could be about the temperature control and the next about the entire oven. The order of the questions should be different for each participant, and applying a computer to generate the questions randomly seems useful here. After collecting the ratings, the ease-of-use measures are

calculated by taking the average of a participant's responses on the six ease-of-use questions.

4. Experiments

Potential problems in validation studies of usability evaluation methods have created serious concern in the human–computer interaction community about the effectiveness and validity of the results of these methods (Gray and Salzman 1998). Therefore, this section provides the results of a systematic attempt to test the main tenet of the compositional evaluation approach, which is that a system can be studied by analysing the usability of its interaction components. The analysis only used secondary data. No new experiment was set up for this part of the investigation. Instead the data collected in four experiments that had been conducted to test other research questions were re-examined. The four experiments were selected because they all used the component-based usability questionnaire and collected other usability measures as well. This small meta-study, therefore, gives an insight to the effectiveness and validity of the compositional evaluation approach for evaluating components in different devices. All four experiments were conducted in the context of a single project that studied the compositionality of usability (Brinkman 2003), and all participants in the experiments were students of Technische Universiteit Eindhoven. In each of the experiments, participants interacted with different versions of a PC emulator of a device, and afterwards they filled out a questionnaire on the computer. The questionnaire consisted of component-specific ease-of-use and satisfaction questions. Besides component-specific questions, the questionnaire also included questions about the ease-of-use and satisfaction of the entire application. The participant received one question at a time and, as suggested earlier, each participant received the questions in a random order to control possible systematic order effects.

Additional usability data were also collected in all four experiments, such as task time, log file data, and also transcriptions of the debriefing interviews. All PC emulators as well as the log file recording mechanisms were written in Delipi^{TM} 5. The usability of the devices was systematically manipulated by using usable or less usable components in the device. This means that, in contrast to other studies (e.g. Davis 1989), variance in the measured perceived usability was not only related to variance in the individuals' attitude, but also to variance in the 'actual' usability of the interaction components. Before looking at the data of these experiments, the following background sections give a short description of each experiment and its motivation.

4.1. Mobile phone

The first experiment was a usability experiment with a mobile phone (Brinkman *et al*. 2004a). This experiment was conducted to validate the findings from an earlier experiment (Brinkman 2003) that had resulted in a behaviour-based component-specific usability measure. Eighty participants, 53 male and 27 female, between the ages of 18 and 28 ($M = 21.43$, $SD = 2.27$) years old, were asked to use one of eight versions of a mobile phone to make a call, send a short text message, and to add a person to the address list. Three components were manipulated, a component responsible for selecting a function (function selector), a component responsible for creating characters (keypad), and a component responsible for sending a short text message (send text message). A more usable version and a less usable version were developed for each component based on the ideas of the cognitive complexity theory (Kieras and Polson 1985). This theory predicts that the number of rules users have to master to operate a system determines the complexity of that system, or in other words its usability. The components' dialogue structure was manipulated to require participants to learn more or fewer rules before they could interact effectively with them. For example, one version of the function selector was implemented with a relatively broad but shallow menu, and another version with a relative narrow but deep menu. Users operating a broad but shallow menu have been reported (Snowberry *et al*. 1983) to be faster and make fewer errors than when operating in a narrow but deep menu structure, as they have to master fewer rules. The keypad component was implemented with a repeated-key method or a modified-model-position method. Again the first version is reported (Detweiler *et al*. 1990) to be easier to use than the second version. Finally, the send text message component was implemented with an easier version, which guided users through the required step of sending a text message, while the less usable version left it to the user to master the right action sequences.

4.2. Room thermostat

The second experiment (Brinkman *et al*. 2004c) was set up to study the effect of inconsistency between components that operated in the same layer. A total of 48 participants, 16 female and 32 male, aged between 18 and 27 years ($M = 21.69$, $SD = 2.03$) were asked to operate one of four versions of a room thermostat that were developed by manipulating the control of the day and the night temperature. These two components were either implemented with the more usable *moving pointer* control or with the less

usable *moving scale* control (Sanders and McCormick 1993) to indicate the setting of the temperature. Furthermore, ergonomic design principles also warn against mixing these two types of control in a device when they are used for a related function (Sanders and McCormick 1993). The results of the experiment acknowledge this concern. The inconsistent combination of a moving pointer control for the day temperature and a moving scale control for the night temperature had a larger negative effect on the components' usability than could be explained by the usability of individual components independently.

4.3. Web-enabled TV set

The same 48 participants of the previous experiment also participated in the third experiment (Brinkman *et al*. 2004c). A similar inconsistency effect was found between two components, a browser and a web site. This time, however, the components operated on two different layers. Participants in the experiment were asked to use a web-enabled TV set to locate the web page with the departure time of a specific bus. As in the previous experiment, this experiment had a between-subjects design. Participants were assigned to one of two versions of the browser and to one of two versions of the web site. One browser, the linear-oriented version, interpreted the up and down buttons as *select the previous link* or *select the next link in succession*. The sequence went from left to right and continued on the left of screen to the next line. The left and right buttons were interpreted as *jumping to previous webpage* and *activate the selected link*. The other browser, the plane-oriented version, interpreted the up, down, left, and right buttons as moving the cursor in the associated direction. Activation of the link itself was done with a separate button. The two versions of the web site related to the layout of the web pages. In one version, the list layout, all links were placed in a vertical list, and in the other version, the matrix layout, all links were placed in a matrix layout. Ergonomic principles of spatial and movement compatibility (Sanders and McCormick 1993) would predict that the usability of the browser and the web site would be the lowest in the prototype that combined the linear-oriented browser with the matrix layout website. The design of this prototype is inconsistent with the expectation of the users. Users are therefore likely to use (unsuccessfully) the left and right button to move the cursor horizontally. This expectation was confirmed by the results of the experiment. The experiment consequently demonstrated that the usability of a component in one layer (e.g. browser) could be related to the design of a component in another layer (e.g. web site).

4.4. Calculator

The fourth experiment (Brinkman *et al*. 2004b) showed that the usability of a component could also be affected by the demand of mental effort created by the users' interaction with other interaction components. The 24 participants, 8 female and 16 male, aged between 19 and 25 years ($M = 21.33$, $SD = 2.16$), were asked to solve a number of equations by using two versions of a calculator: an editor with a small display that could only display a single value or an operator, or an editor with a large display that could display several lines of an equation. Cognitive ergonomic principles would favour the large display over the small display, especially for large and complex equations. Task information is preferably distributed more to the system than to the user side of the interaction (Zhang and Norman 1994) to reduce in this case the memory load for the user. In contrast to the other three experiments, this experiment had a within-subjects design. Participants were asked to use both versions of the calculator. Next, besides studying the usability of the manipulated editor component, the usability of a higher-level processor and memory component was also measured. Although this component was not manipulated, the results revealed that its interaction was also affected by the version of the editor.

5. Data analysis

The experiments seem successful for the purpose for which they were originally conducted; however, this re-examination of the collected data looked at the reliability and validity of the component-based usability questionnaire. Reliability refers to the extent to which the questionnaire yields consistent scores over repeated observations. Validity refers to the extent to which the questionnaire measures what it claims to measure – the user's attitude towards the usability of an individual component. Four types of validity are often mentioned for research methods in the area of social sciences (e.g. Neuman 1997), which are: face validity, whether the questionnaire looks valid; content validity, whether the full content of usability is represented in the measure; criterion validity, whether the results of the questionnaire agree with other known usability measures; and construct validity, whether the questionnaire measures the unobservable, theoretical construct – usability of a component. There are no statistical procedures to analyse face or content validity. They are assessed by studying the measuring procedure and the questions in the questionnaire. Both seem acceptable, since the component-based questionnaire is based on a well-established questionnaire and on a theoretical compositional framework. Reliability,

criterion validity and construct validity, on the other hand, can be assessed by applying statistical procedures on the data of the experiments. If the data show a low degree of reliability or validity, then the results of the component-based usability questionnaire is questionable. On the other hand, a high degree would support future use of the questionnaire.

However, before analysing the data, the data of the calculator experiment had to be restructured. As mentioned before, this experiment had a within-subjects design, whereas the other three experiments had a between-subjects design, where participants were only assigned to one version. To reduce the complexity of the analysis, the calculator experiment was therefore also treated as a between-subjects design, by splitting data from the participants into two groups. Data about the small display calculator was taken from participants who started the experiment with solving an equation on a small display calculator, and data about the large display calculator was taken from participants who started the experiment with solving an equation on a large display calculator.

5.1. Reliability and validity

5.1.1. Equivalence reliability

The first step of the analysis was to study the reliability of the component-based questionnaire. As multiple ease-of-use questions were used for each system and component, it was possible to measure the equivalence reliability. This type of reliability shows whether the measures yield consistent results across the different questions. Table 1 shows the Cronbach's α for the ease-of-use questions. All values are in line with the recommended reliability of level above 0.8 (Loewenthal 2001). This suggests that the six questions are related to the same underlying construct. Because of the high reliability, it seems acceptable to take the mean of the six questions as an aggregated measure for each component and for the entire system.

Table 1. Reliability of the component questionnaires.

Application/component	Cronbach's α
Mobile telephone	0.85
Function selector	0.87
Keypad	0.85
Send text message	0.89
Room thermostat	0.82
Daytime temperature	0.92
Nighttime temperature	0.92
Web-enabled TV set	0.91
Browser	0.90
Web pages	0.89
Calculator	0.96
Editor	0.97
Processor	0.92

5.1.2. Criterion validity

There are two basic types of criterion-related validity: predictive validity, the ability to predict something it should theoretically be able to predict; and concurrent validity, the ability to give similar results as other accepted standard measures collected at the same time. To start with predictive validity, the various versions of the components were designed to vary in usability. These predicted variations in usability were guided by established ergonomic principles such as cognitive complexity, consistency and mental load. Table 2 shows values of the Pearson correlation between the binary variables representing the versions of a specific component (low or high usability predicted) and its corresponding component-specific measure. The large number of significant correlations between the component-specific measures and the versions indicates an acceptable degree of predictive validity. Only the correlations between the measures associated with usability of the send text message and its version failed to reach a significant level.

How well the component-specific measures agreed with other measures (concurrent validity) is also illustrated in Table 2. Along with the ease-of-use questions, the questionnaire also included component-specific satisfaction questions and overall satisfaction questions. Participants rated each component and the application on two scales taken from Lewis' (1995) post-study system usability questionnaire. One question was a statement on how pleasant a specific component or application was and the other question a statement on how much a participant liked using it. Taking the average of ratings on both seven-point Likert scales results in a satisfaction measure for each component and for the overall application. Table 2 shows strong correlations between the ease-of-use and the satisfaction measures, suggesting that they relate to the same underlying construct, i.e. usability.

Along with the data for the perceived usability measures, objective interaction data were also recorded in the experiments. The overall perceived usability measures were correlated with the time participants took to complete a task, and the component-specific perceived usability measures were correlated with the number of messages received by a component, which is regarded as a component-specific usability measure (Brinkman *et al*. 2004a). In the case of the calculator these behavioural measures were first logarithmically transformed to limit the effect of outliers. The large number of significant correlations that was found indicates a considerable degree of agreement between results of the new measure and these already established measures. Besides collecting behavioural measures, participants of the mobile phone experiment

Table 2. Pearson correlations between, on one hand, perceived measures and, on the other hand, the version of components, satisfaction, behavioural measures and remarks made in the debriefing interview.

Interaction component	Version	Satisfaction	Behaviour	Debriefing
Mobile telephone		0.66**	−0.36**	−0.36**
Function selector	0.48**	0.77**	−0.36**	−0.51**
Keypad	0.33**	0.74**	−0.58**	−0.26*
Send text message	0.07	0.68**	−0.21	−0.18
Room thermostat		0.72**	−0.26	−0.38**
Daytime temperature	0.44**	0.85**	−0.19	−0.23
Nighttime temperature	0.52**	0.87**	−0.44**	−0.38**
Web-enabled TV set		0.82**	−0.69**	−0.23
Browser	0.33*	0.83**	−0.41**	−0.39**
Web pages	0.29*	0.77**	−0.57**	−0.02
Calculator		0.87**	−0.47*	−0.86**
Editor	0.86**	0.87**	−0.42*	−0.86**
Processor and memory	0.73**	0.86**	−0.31	−0.84
Mean	0.45	0.79	−0.41	−0.42

*Correlation is significant at the 0.05 level (2-tailed).
**Correlation is significant at the 0.01 level (2-tailed).

were also asked to fill out Norman's (1991) questionnaire on menu selection to evaluate the menu of the mobile phone. The results correlated significantly with the ease-of-use ($r = 0.67$; $p < 0.01$) measure of the function selector component. This suggests that similar data can be obtained from both the generic and specifically developed questionnaires for this component. Generalising this finding to other evaluations of components would mean that there is not always a need to develop specific, tailor-made component questionnaires, such as a menu questionnaire, when using a generic component-specific questionnaire.

Concurrent validity was also examined by comparing the perceived usability established via the questionnaire and what participants mentioned in the debriefing interview afterwards. Table 2 shows significant correlations between the overall measures and the number of usability problems mentioned by the participants. The table also shows significant correlations between the binary variables representing whether or not a participant mentioned a problem with a specific component and its corresponding component-specific measure. In the case of the calculator experiment, where all participants used two calculators, the debriefing measures were not based on the number of usability problems mentioned, but instead on the participants' preference for the version that they used to solve their first equation.

5.1.3. Divergent validity and explaining the overall usability

The next step of the analyses was to conduct a series of regression analyses with the overall measures as criterion variables and the component-specific measures as predictors. As mentioned in Section 3, an important validity criterion for any potential component-based usability questionnaire is that it should provide data to explain at least a part of the overall usability. In other words, what is the relationship between participants' attitude towards the usability of the individual components and the entire system? The regression model used in the analyses is an additional model which adds up the weighted (B_i) attitude (A_i) towards the individual components to explain the attitude towards the overall usability (A_{overall}), or more formally:

$$A_{\text{overall}} = \sum_{i=1}^{n} (B_i \times A_i) \quad (1)$$

Theories such as the theory of reasoned action (Ajzen and Fishbein 1980) or the multi-attribute utility theory (Keeney and Raiffa 1976) have suggested weighing factors to link individual items with an overall measure. In this case, weighting is appropriate as attitudes towards individual components might relate differently with users' attitude towards the overall system. Since both the overall attitude and the component-specific attitudes were collected in the experiments, regression analyses could fit the weighting factors. All regression analyses used the enter method, which resulted in all cases in significant models with adjusted R^2 values ranging from 79% to 93% (Table 3). To put this into context, according to Myers (as cited in Stevens 1996) behavioural scientists dealing in data reflecting human behaviour may feel fortunate with an R^2 as high as 70%. In these experiments the users' attitude towards the various versions of the individual interaction components could therefore explain between 79% and 93% of the variance in the users'

Table 3. Results of four regression analyses on the overall ease-of-use based on the ease-of-use of interaction components.

Application	R	R^2	Adj. R^2	SE	df_{reg}	df_{res}	F	p	MCC
Mobile telephone	0.90	0.80	0.80	0.475	3	76	104.23	<0.001	0.57
Room thermostat	0.90	0.80	0.79	0.367	2	45	91.75	<0.001	0.21
Web-enabled TV set	0.95	0.90	0.90	0.331	2	45	212.99	<0.001	0.94
Calculator	0.97	0.94	0.93	0.437	2	20	149.13	<0.001	0.78

attitude towards the overall usability of the various versions of the system. Furthermore, the mean errors of the models' predictions were relatively small. The standard error (*SE*) of the estimate overall measure shows on average an error of 0.4 on the seven-point Likert scale. It seems, therefore, that usability of individual components can explain at least a part of users' attitude towards the usability of the whole application. In the case of the four experiments, it was a considerable part that could be explained.

As there is only one ease-of-use measure for each component, construct validity entails only divergent validity, whether the rating of one component-specific measure is unrelated to the rating of another component-specific measure. The degree of divergent validity seems satisfactory in three of the four experiments. Although some similarity in the ratings can be accepted, correlations between the ratings should be below the adjusted R^2 value. As expected, this was clearly not the case for the web-enabled TV set, with a 0.94 mean correlation among its components (*MCC*) (Table 3). The participants did not make a distinction in their rating for the browser and the web pages, or even for the entire application since the *MCC* value is so close to 0.90 adjusted R^2 value. The Web-enabled TV set and calculator experiments were set up to demonstrate that the usability of one component could influence users' attitude towards another component, hence the high *MCC* values. With the mobile phone, participants especially had a problem with the send text message component. Its rating had a 0.58 correlation with the keypad rating and 0.67 correlation with the function selector rating. In contrast, participants had less difficulty distinguishing the other two components in the mobile phone. They only had a 0.46 correlation. Despite these concerns, the variance inflation factor (*VIF*), a multi-collinearity indicator, for each of these predictors is below 10 (Table 3). Myers (1990) argues that above this threshold predictor variables might be too confounded due to correlation among them. Table 4 shows that the component-specific measures were all significant predictors, or nearly significant in the case of the send text message or not at all in the case of the processor and memory components. These findings illustrate that the prediction of the perceived overall usability is not always simply linked with the usability of one single component, but in some cases with multiple components.

To conclude the section on reliability and validity, the component-specific measure seems to have an acceptable level of both predictive and concurrent validity and consequently criterion validity. Together with the findings on reliability and construct validity, it seems that, at least in some cases, this measure can be a valid indicator of how users perceive the usability of an interaction component. However, this was not true in all cases. For example, results related to the send text message component revealed a low degree of construct and criterion validity. Participants were not able to make a clear distinction between the usability of this component and the other components. One possible reason could be that participants did not attribute the name or description used in the questions to one single component. Testing this explicitly in a pilot run of a questionnaire in the future seems therefore advisable. Another related reason is that users' attitude towards a component can be influenced by other components. For example, a less usable keyboard could have affected participants' attitude towards the send text message component in the mobile phone.

5.2. *Norm data and an example usability test*

Up until now the analyses focused on the underlying evaluation model and the validity of the measures. However, to be of practical use in a usability test, engineers should be able to establish which component is usable and which is not. As developers of standardised usability tests, such as SUMI (Kirakowski and Corbett 1993) and WAMMI (Kirakowski *et al.* 1998), have realised, engineers need a norm to compare ratings with. The norm proposed here is based on the data obtained in the four experiments, which can be split in two: a sample set A of 18 cases where participants rated a difficult component implemented in a specific prototype, and a sample set B of 26 cases where participants rated a more usable component. The data shows a mean of 4.75 ($SD = 0.50$) for the average rating in set A, and a mean of 5.93 (*SD* 0.59) for the average rating in the set B. Using these means and standard deviations as a benchmark, it is

possible to determine a break-even point on the seven-point rating scale. As Figure 4 shows, this point is approximately 5.29. Only 14% of the difficult components would receive a mean rating higher than this point, and similarly only 14% of the more usable components would receive a mean rating lower than this point. Therefore, a rating above this point suggests a rating more comparable with the rating of sample set B, the more usable components, and less comparable with the rating of sample set A, the less usable components. Likewise, a rating below the break-even point suggests the opposite.

The remainder of this section will look at a small usability study with an MP3 player[1] to illustrate how engineers could apply the questionnaire and draw conclusions from the data. The MP3 player was developed to have two relatively easy-to-use and two relatively difficult-to-use components. The easy-to-use components were the play control and the file control components. With the play control, users could direct the playing of an MP3 file, for example to start playing a file, to pause it, or to jump to the next file in the list (Figure 5). With the file control (Figure 6), users could search and open a single or a group of MP3 files. Since both components used standard interaction elements, such as play, pause, next, fast forward buttons, and the Windows dialogue box for selecting and opening files, Windows users could be expected to be familiar with

Table 4. Estimated coefficients of regression models explaining the overall ease-of-use rating based on the component-specific rating.

Interaction component	*B*	*SE*	*β*	*t*	*p*	*VIF*
Mobile telephone						
Constant	0.02	0.299		0.06	0.949	
Function selector	0.47	0.063	0.512	7.48	<0.001	1.82
Keypad	0.41	0.063	0.405	6.46	<0.001	1.53
Send text message	0.13	0.067	0.146	1.94	0.056	2.18
Room thermostat						
Constant	1.46	0.345		4.22	<0.001	
Daytime temperature	0.45	0.045	0.670	9.90	<0.001	1.05
Nighttime temperature	0.32	0.046	0.470	6.95	<0.001	1.05
WebTV						
Constant	−0.52	0.353		−1.48	0.147	
Browser	0.67	0.146	0.626	4.58	<0.001	8.78
Web pages	0.41	0.164	0.338	2.48	0.017	8.78
Calculator						
Constant	0.22	0.437		0.51	0.618	
Editor	0.89	0.106	0.923	8.36	<0.001	3.88
Processor & memory	0.07	0.148	0.052	0.47	0.642	3.88

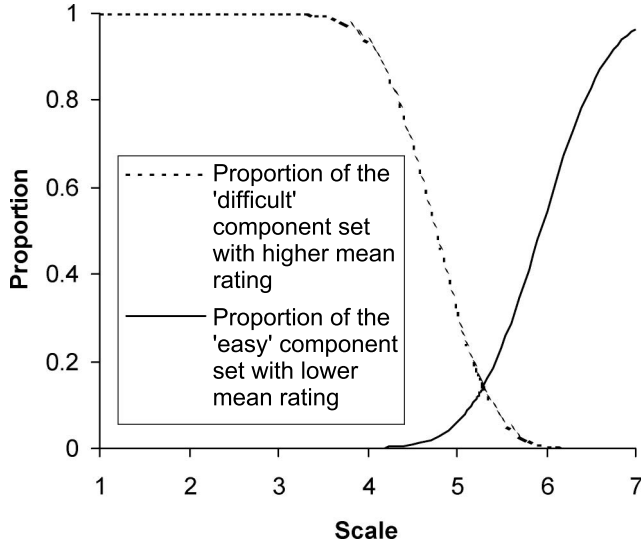

Figure 4. Proportion of components (from the difficult or easy set) that received higher or lower mean rating on the seven-point scale.

their use and therefore consider them easy to use. The opposite was the case for the other two components, the volume control and the information control. With the volume control, users could set the volume by using a combination of three buttons. For example, to decrease the volume, users first had to click on the button above the volume indicator, and secondly they had to click on the large button on the right side of the volume indicator (Figure 5). This would decrease the volume by 5% and reduce the size of the upper part of the indicator by 5% or, presumably more confusing, increase the size of the lower part of the volume indicator by 5%. To reduce the volume further they had to repeat the sequence of pressing the two buttons. To increase the volume, users had to follow the same procedure. However, this time instead of clicking on the button above the indicator, users had to click on the button below the volume indicator. This design was expected to be difficult to use because it violated several usability heuristics (Nielsen 1993), such as: simple and natural dialogue, consistency, and clear feedback. With the information control (Figure 7), the other difficult-to-use component, users could obtain information on an MP3 file regarding the song title, the artist, album's name, the year, the genre, and the track number. The design violated the ergonomic principle of consistency as each information element was obtained in a different way. For example, for the artist's name, the users had to select the song title in a dropdown box and press a search button, after which the name was displayed, whereas for the genre the users had to select only the song title and the genre would be displayed automatically, while for the year users had to select the song title and click with the right mouse button on the year text box. In contrast to these elements where users had to select the song in the information form (Figure 7), for the album's name users first had to select the song name in the directory list of the player (Figure 5), while the song title displayed at the left corner of the information form was that of the song currently played by the player.

To evaluate the MP3 player, an unmonitored, online usability test was conducted. Participants were asked to download the player and a number of MP3 files on their computer and afterwards to attempt to complete a series of tasks with them, such as playing a specific MP3 file, changing the volume to its maximum, and checking whether information about a song was correct. The participants were instructed to spend a maximum of 5 minutes on each task before moving on to the next task. After this they were asked to complete an evaluation questionnaire. The questionnaire had six sections: general background information and five sections to evaluate the individual components and the overall MP3 player. For practical reasons the questions related to a specific component or to the overall MP3 player were grouped together, combined with an open question which asked the participant to explain why they had given a specific rating. Eight versions of the questionnaire were developed to control for

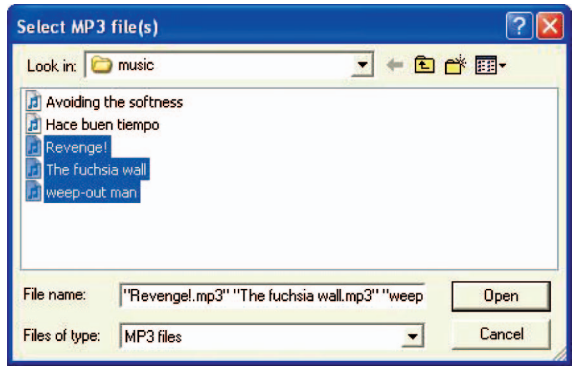

Figure 6. File control.

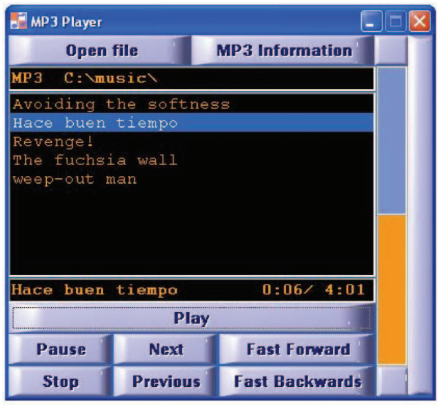

Figure 5. MP3 player.

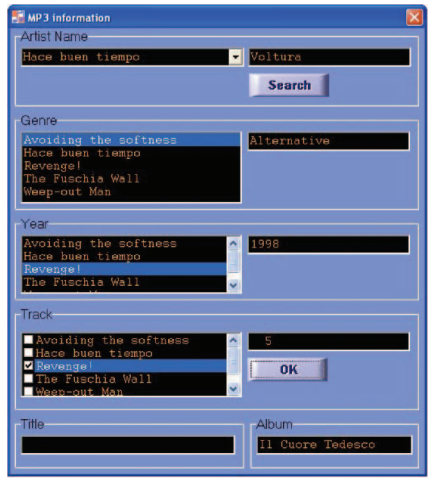

Figure 7. Information control.

potential order effects between the groups of questions. Four versions of the questionnaire started with the evaluation of the whole MP3 player, and four versions ended with this. The order of the evaluation of the components was also varied, ensuring that it was equally distributed. In the test, participants were randomly assigned to one of the questionnaire versions.

In total, 34 (14 females and 20 males) students and staff members (6 undergraduate students, 1 master student, 8 PhD students and 20 staff members[2]) of Brunel University participated. Their age ranged from 18 to 62 years old ($M = 31.44$, $SD = 9.88$). The first step of the analysis was to examine the reliability of the questions for each component and for the whole MP3 player. Participants seemed highly consistent in their rating of the six question items as Cronbach's α ran from 0.90 for the volume control to 0.98 for the play and the file control. Because of the high consistency, the analysis was continued by using the average rating for each component and the average rating for the MP3 player. The mean rating of the overall MP3 player was 4.37 ($SD = 1.59$), the mean ratings of the four components can be found in Table 6. The standard deviations, which range from 1.40 to 1.90, show that participants varied in their rating. Individuals might have perceived the usability differently, or they might have expressed it differently in their ratings. Therefore, questionnaire rating seems less suitable to be examined on an individual level. Instead engineers should examine the central tendency of the ratings across the participants as an indication of the average perceived usability of a component. Before doing this for the components of the MP3 player, the next step of the analysis focused on whether participants' rating agreed with the rationale they gave for the rating.

Two raters independently classified the comments participants had given for their ratings into four categories (Table 5). Afterward the agreement between the raters was examined by calculating the Cohen's κ, a measure for inter-raters reliability. Cohen's κ ranged from 0.93 to 0.95, which suggested a high level of agreement for the comments' rating of each component and the whole MP3 player. For the five cases on which raters initially disagreed, the raters discussed them and agreed on a classification. This resulted in single classification of each comment. Next, the average rating, which could range from 1 to 7, was recoded into three equal interval classes: *low* for a rating from 1 to 3, *medium* for a rating from 3 to 5 and *high* for a rating for 5 to 7. A frequency cross-tabulation (Table 5) of comments' rating set against average ratings shows the agreement between the participants' rating and their given rationale. Twelve per cent of the comment boxes were left open or participants gave answers that were not regarded as relating to the rating. This was mainly the case where a participant had given a high rating. Of the 149 relevant comments given by the participants only nine seem clearly inconsistent with their rating, e.g. mainly negative comments, and/or improvement advice combined with a high usability rating; or oppositely, mainly positive or no negative comments and no improvement advice combined with a low usability rating. When the five participants responsible for these nine comments were queried about the apparent inconsistency, they indicated that although they had made only negative comments or suggestions for improvement, this did not imply that they found the player or component difficult to operate, or as one participant replied 'Even thought [*sic*] I think that the play control is not flexible enough, I may think that I can get to use it in a pretty good way'. Another participant had a more lenient rating strategy as he stated 'I did not want to mark it down on the one problem which i had, and insted [*sic*] noted this down in the comments about the year option not working for me'. Language problems and misunderstanding of the

Table 5. Frequency cross-tabulation between classified comments and mean ratings.

	Mean rating			
Comments	Low [1, 3)	Medium [3, 5)	High [5, 7]	Total
Missing or not relevant	*3*	*2*	*16*	21
Mainly negative comments, and/or improvement advice	*39*	*20*	*7*	66
Both negative and positive comments, and/or improvement advice	*3*	*17*	*24*	44
Mainly positive or no negative comments and no improvement advice	*2*	*1*	*36*	39
Total	47	40	83	170

Table 6. Results of one-sample *t*-tests with test value 5.29 on the components' mean ease-of-use rating.

Interaction component	Mean	*SD*	*t*	*df*	*p*
Easy to use					
Play control	6.04	1.45	3.01	33	0.005
File control	5.90	1.68	2.12	33	0.042
Difficult to use					
Volume control	2.13	1.40	−13.13	33	<0.001
Information control	3.96	1.90	−4.10	33	<0.001

statement was also suggested as the cause of some confusion. Still, the remaining 140 comments seem reasonably consistent with the ratings. In other words, almost all participants seem to have been able to provide a rating that was consistent with their rationale, although this rationale did not always seem to agree with the usability heuristics. For example, for the information control, expected to be difficult to use, some participants gave comments such as 'It was very easy to me to use the information control and it was very clear and understandable', or 'It has been separated logically and made it easy for the used [*sic*] to understand each part functions'. Oppositely, for the file control, expected to be easy to use, a participant wrote down 'Hated this about the player – users shouldn't have to apply such rudimentary techniques to get the list in. A bit of a shocker. Worst feature of the player by far'. These variations in comments reflect the variations in the rating, which again supports an examination of the central tendency of the data set as was done in the final step of the analysis.

A series of one-sample *t*-tests was conducted to see whether the mean rating for a component deviates significantly from the 5.29 break-even point between easy- and difficult-to-use components of the four experiments discussed in the previous sections. Table 6 shows that the participants gave the easy-to-use components, play control and file control, average ratings that were significantly above the break-even point, while they rated the difficult-to-use components below it. With these results, engineers should therefore focus their improvement effort on the volume control and the information control. They should probably give the highest priority to the volume control as it received a significantly lower rating ($t(33) = -5.01$, $p < 0.001$, paired sample *t*-test) than the information control, and 68% of participants mentioned it in their rating rationale for the overall usability compared to 21% of the participants who mentioned the information control. If, however, the analysis would provide no significant results, in other words the test would be inconclusive, engineers are left with two options. First they might decide that participants perceived the usability as somewhere close to the break-even point. This decision seems appropriate when the sample size is sufficient enough. As an indication of appropriate sample sizes, to conduct a one-sample *t*-test with an 80% success rate of detecting a deviation on a significant-level of 0.05, at least 14 participants would be necessary to detect a deviation classified as large by Cohen (1988), and at least 33 participants for a medium classified deviation. The second option when faced with inconclusive results is to increase the power of the statistical test by including more participants in the study. This will help to cope with usability ratings that are not easily distinguishable by a one-sample *t*-test from the break-even point. A possible reason might be a relatively large standard deviation. In other words, participants are disagreeing about the usability of a component. Engineers should try examining other types of data as well, such as from interviews or observations, to understand the reason why participants vary in their opinion. Likewise, engineers should also look at user-related factors that influence usability, such as task experience, experience with related equipment, or the environment where the device was used.

6. Discussion and further research

This study represents a detailed and systematic effort to examine a compositional approach to usability evaluation. The findings obtained in the four experiments suggest that at least part of the usability of a product can be studied on a detailed, compositional level. The reliability and validity of the component-specific measure is quite acceptable, although not in all cases. The findings support an evaluation approach in which engineers ask users to rate their interaction experience with specific parts of the system and not only with the system as a whole; take, for example, the results of the MP3 player. Whereas the holistic questionnaire could only say that the overall usability was 4.37 on a seven-point scale, the component-specific questionnaire could tell that the usability of the play and file control was high, and the usability of the volume and information control was low compared to the norm data.[3] It illustrates how the component-based questionnaire can give improvement effort a clear direction.

The evaluation approach fits in well with the popular component-based software engineering (CBSE) approach. Engineers can identify specific elements in the system responsible for the reduction of the overall usability, and they can improve or replace these disruptive elements. The approach can also help usability evaluators to cope with a severe limitation often encountered in usability evaluations, the so-called evaluator effect. As two systematic studies (Hertzum and Jacobsen 2001, Molich *et al*. 2004) have demonstrated, different usability evaluators, or even entire teams, can end up defining completely different usability problems when analysing the same product. Compositional evaluation gives evaluators a sense of direction because results are directly related to specific parts of the system. It is not left to individual evaluators to make this link. Still, the problem of identifying the source of a problem is not entirely solved. Instead it now re-emerges in a reduced but compositional format. Where before evaluators had to

identify the usability problem affecting the users' attitude towards the overall usability, they now have to identify the usability problem affecting the users' attitude towards an individual component. The extra open question, introduced in the questionnaire of the MP3 player evaluation that asks participants to explain their rating, might provide valuable information here, especially because of consistency found between rating and comments.

Although the findings suggest that users' attitude towards a component can effectively be related to the usability of that component, the findings also suggest that it can be affected by other components. This also means that components rated as very usable in one application might not always be rated as very usable in another application with other components. The finding puts a limit to the idea of creating a library with highly usable components, to create highly usable applications. Whether a component is ultimately usable has to be evaluated in the context of the new application, user group, task, equipment and environment. It cannot simply be derived from evaluations in another context.

The compositional approach might also help to improve the technology acceptance model (TAM) (Taylor and Todd 1995). TAM is widely studied in the information systems research community and is an adaptation of the theory of reasoned action (Ajzen and Fishbein 1980). TAM uses Davis' (1989) ease-of-use construct, which has also been used in this study. Together with the perceived usefulness, these two constructs have been shown to be effective in explaining usage of information technology (Taylor and Todd 1995). Recent criticism (Lee *et al*. 2003), however, points at the limited practical use of TAM, as it does not explain how to improve the usefulness or ease-of-use of a product. A compositional approach as suggested here might provide part of the answer. Managers would have an indication of which system component to focus on if they want to change the usage of the technology.

The study presented here, however, is not without its limitations. First, ecological validity is always a point of concern of laboratory studies. Next, the ultimate criterion for effectiveness of a usability evaluation method is how well the method helps evaluators discover *real* usability problems. Or phrased in a more general question, has it something to say about what people do in 'real' culturally and economically significant situations? This is a question often posed when it comes to cognitive theories (e.g. Neisser 1976, Kaptelinin 1996, Hoc 2000). A usability problem is 'real' if it is a predictor of a problem that users will encounter in real work-context usage and that will have an impact on usability. In the experiments, the 'reality' was obtained by seeding *known* usability problems into the prototypes. This approach has been criticised (Hartson *et al*. 2001) as a standard approach to test *analytic* usability evaluation methods, because it heavily depends on the researchers' skill to shape a problem in the prototype. Furthermore, the prototypes were designed with the intent of testing and not of using, putting ecological validity again in doubt. Still, 'seeding' seems an appropriate approach here, as perceived usability measures were compared with other empirical measures. Next, the experiments were conducted with high-fidelity PC emulators of devices that used photographic material and gave highly realistic visual and auditory system feedback. Further research, however, could focus on applying the component-specific questionnaire in a field setting, where users are not instructed to perform a certain task, in a certain environment.

To conclude, the findings of this study support a compositional evaluation model, which underlies a usability evaluation approach that is in line with the CBSE approach. Furthermore, the evaluation approach can also be applied to applications not developed according to CBSE. In that case, evaluators have to identify components in the system which have a state that users can perceive and change. The evaluator would again benefit from the main advantage of the compositional evaluation approach of providing detailed information about the usability of specific interaction components, something an overall usability questionnaire fails to do. Although it is unlikely that among practitioners component-based usability questionnaires will surpass in popularity other usability evaluation methods such as user tests, it can be relatively easy to implement in parallel with these other methods. Therefore, integrating instead of replacing other evaluation methods might be a more viable approach, providing engineers with multiple views about the usability of a product.

Acknowledgements

We thank the anonymous reviewers, Kate Hone, Audrey Bink and Nayna Patel for their comments and advice that helped us to improve the paper.

Notes

1. The introduction text, task instruction, the MP3 player, music files, questionnaires, and the results of the test can be found at http://mmi.tudelft.nl/~willem-paul/mp3player/Intro.htm
2. One undergraduate student also worked for the university.
3. If the MP3 data is incorporated into the norm data set, the mean of sample set A becomes 4.58 ($SD = 0.77$), and 5.94 ($SD = 0.57$) for sample set B. The 2.13 mean rating of the volume control however is an extreme outlier

($< Q_1 - 3 \times IQR$), which would drive up the break-even point to 5.36. Ignoring this extreme outlier, results in a mean for sample set A of 4.71 ($SD = 0.52$) and break-even point of 5.29 again.

References

Ajzen, I. and Fishbein, M., 1980. *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Aykin, N., 1994. Software reuse: a case study on cost-benefits of adopting a common software development tool. *In*: R.G. Bias and D.J. Mayhew, eds. *Cost-justifying usability*. London: Academic Press, 177–202.

Bark, I., Følstad, A., and Gulliksen, J., 2005. Use and usefulness of HCI methods: results from an exploratory study among Nordic HCI practitioners. *In*: *Proceedings of HCI 2005*. 5 – 9 September 2005. Edinburgh, London: Springer-Verslag, 201–217.

Borchers, J., 2001. *A pattern approach to interaction design*. Chichester, UK: John Wiley.

Brinkman, W.-P., 2003. Is usability compositional? Technische Universiteit Eindhoven. PhD thesis.

Brinkman, W.-P., Haakma, R., and Bouwhuis, D.G., 2004a. Empirical usability testing in a component-based environment: improving test efficiency with component-specific usability measures. *In*: *Pre-Proceedings of EHCI-DSVIS*. 11 – 13 July 2004, Hamburg, Germany 340–356.

Brinkman, W.-P., Haakma, R., and Bouwhuis, D.G., 2004b. Memory load: a factor that links the usability of individual interaction components together. *In*: *Proceedings of HCI 2004*. 6 – 10 September 2004, Leeds, United Kingdom, British HCI Group, Vol. 2, 165–168.

Brinkman, W.-P., Haakma, R., and Bouwhuis, D.G., 2004c. Consistency: a factor that links the usability of individual interaction components together. *In*: *Proceedings of Twelfth European Conference on Cognitive Ergonomics*. 12 – 15 September 2004, York, United Kingdom, EACE, 57–64.

Brinkman, W.-P., Haakma, R., and Bouwhuis, D.G., 2005. Usability testing of interaction components: taking the message exchange as a measure of usability. *In*: R.J.K. Jacob, Q. Limbourg, and J. Vanderdonckt, eds. *Computer-aided design of user interfaces IV*. Dordrecht, The Netherlands: Kluwer Academic, 159–170.

Brooke, J., 1996. SUS: a ''quick and dirty'' usability scale. *In*: P.W. Jordan, B. Thomas, B.A. Weerdmeester, and A.L. McClelland, eds. *Usability evaluation in industry*. London: Taylor and Francis, 189–194.

Carver, C.S. and Scheier, M.F., 1998. *On the self-regulation of behavior*. New York: Cambridge University Press.

Chin, J.P., Diehl, V.A., and Norman, L.K., 1988. Development of an instrument measuring user satisfaction of the human-computer interface. *In*: *Proceedings of CHI*. Washington, New York, NY: ACM press, 213–218.

Cohen, J., 1988. *Statistical power analysis for the behavioural sciences*. New York: Academic Press.

Coleman, W.D., Williges, R.C., and Wixon, D.R., 1985. Collecting detailed user evaluations of software interfaces. *In*: *Proceedings of the Human Factors Society – 29th Annual Meeting*. Santa Monica, CA: Human Factors Society, 240–244.

Coutaz, J., 1987. PAC, an object oriented model for dialog design. *In*: *Proceedings of INTERACT'87*. Amsterdam: North-Holland, 431–436.

Cox, B.J., 1990. There is a silver bullet: a software industrial revolution based on reusable and interchangeable parts will alter the software universe. *Byte*, 15, 209–218.

Davis, F.D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13, 319–340.

Detweiler, M.C., Schumacher, M.C., and Gattuso, N.L., 1990. Alphabetic input on a telephone keypad. *In*: *Proceedings of the Human Factors Society – 34th Annual Meeting, vol. 1*. Santa Monica, CA: Human Factors Society, 212–216.

Doll, W.J., Hendrickson, A., and Deng, X., 1998. Using Davis's perceived usefulness and ease-of-use instruments for decision making: a confirmatory and multigroup invariance analysis. *Decision Sciences*, 29, 839–869.

Dragivevic, P. and Fekete, J.-D., 2001. Input device selection and interaction configuration with icon. *In*: *Proceedings of IHM-HCI*. 10 – 14 September 2001. Lille, France: Springer Verlag, 553–558.

Eagly, A.H. and Chaiken, S., 1993. *The psychology of attitudes*. Belmont, CA: Wadsworth.

Gediga, G., Hamborg, K.-C., and Düntsch, I., 1999. The IsoMetrics usability inventory: an operationalisation of ISO 9241/10 supporting summative and formative evaluation of software systems. *Behaviour and Information Technology*, 18, 151–164.

Gray, W.D. and Salzman, M.C., 1998. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-computer Interaction*, 13, 203–261.

Gulliksen, J., Boivie, I., Persson, J., Hektor, A., and Herulf, L., 2004. Making a difference – a survey of usability profession in Sweden. *In*: *Proceeding of NordiCHI*. 23–27 October 2004, Tampere, Finland. New York, NY: ACM Press, 207–215.

Haakma, R., 1999. Towards explaining the behaviour of novice users. *International Journal of Human-Computer Studies*, 50, 557–570.

Hartson, H.R., Andre, T.S., and Williges, R.C., 2001. Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction*, 13, 373–410.

Hertzum, M. and Jacobsen, N.E., 2001. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 13, 421–443.

Hoc, J.-M., 2000. Toward ecological validity of research on cognition. *In*: *Proceedings of IEA 2000/HFES 2000 congress, vol. 1*. Santa Monica, CA: Human factors and Ergonomics Society, 549–552.

Hollemans, G., 1999. User satisfaction measurement methodologies: extending the user satisfaction questionnaire. *In*: *Proceedings of International Conference on Human-Computer Interaction*. Mahwah, NJ: Lawrence Erlbaum, 1008–1012.

ISO, 1998. Ergonomic requirements for office work with visual display terminals (VDTs) Part 11. Guidance on usability (ISO no 9241-11). Geneva: International Organization for Standardization.

Kaptelinin, V., 1996. Activity theory: implication for human-computer interaction. *In*: B.A. Nardi, ed. *Context and consciousness*. London: MIT Press, 103–116.

Keeney, R.L. and Raiffa, H., 1976. Decisions with multiple objectives: preferences and value tradeoffs. New York: John Wiley.

Kieras, D. and Polson, P.G., 1985. An approach to the formal analysis of user complexity. *International Journal of Man-Machine Studies*, 22, 365–394.

Kirakowski, J. and Corbett, M., 1993. SUMI: the software usability measurement inventory. *British Journal of Educational Technology*, 24, 210–212.

Kirakowski, J., Claridge, N., and Whitehead, R., 1998. Human centered measures of success in web site design. *In*: *Proceedings of the Fourth Conference on Human Factors and the Web*. Available online at www.research.att.com/conf/hfweb/proceedings/kirakowski/ (accessed 16 May 2005).

Krasner, G.E. and Pope, S.T., 1988. A cookbook for using the model-view-controller user interface paradigm in Smalltalk-80. *Journal of Object-Oriented Programming*, 1, 7–49.

Lee, Y., Kozar, K., and Larsen, K.R.T., 2003. The technology acceptance model: past, present, and future. *Communications of the Association for Information Systems*, 12, 752–780.

Lewis, J.R., 1995. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7, 57–78.

Lin, H.X., Choong, Y.-Y., and Salvendy, G., 1997. A proposed index of usability: a method for comparing the relative usability of different software systems. *Behaviour and Information Technology*, 16, 267–278.

Loewenthal, K.M., 2001. *An introduction to psychological tests and scales*. Hove: Psychology Press.

Love, S., 2005. *Understanding mobile human-computer interaction*. Amsterdam: Elsevier.

Mao, J.-Y., Vredenburg, K., Smith, P.W., and Carey, T., 2005. The state of user-centered design practice. *Communications of the ACM*, 48, 105–109.

McIlory, M.D., 1979. Mass produced software components. *In*: *Software Engineering: Concepts and Techniques: Proceedings of the NATO Conferences*. 1968 and 1969, Garmisch, Germany, and Rome, New York, NY: Mason/Charter, 88–98.

Molich, R., Ede, M.R., Kaasgaard, K., and Baryukin, B., 2004. Comparative usability evaluation. *Behaviour & Information Technology*, 23, 65–74.

Myers, B.A., 1998. A brief history of human-computer interaction technology. *Interactions*, 5, 44–54.

Myers, R., 1990. *Classical and modern regression with applications*. Boston, MA: PWS-KENT.

Neisser, U., 1976. *Cognition and reality*. San Francisco, CA: W.H. Freeman.

Neuman, W.L, 1997. *Social research methods: Qualitative and quantitative approaches*. Boston, MA: Allyn and Bacon.

Newell, A., 1990. *Unified theories of cognition*. London: Harvard University Press.

Nielsen, J., 1986. A virtual protocol model for computer-human interaction. *International Journal of Man-Machine Studies*, 24, 301–312.

Nielsen, J., 1993. *Usability engineering*. Boston: AP Professional.

Norman, D.A., 1984. Stages and levels in human-machine interaction. *International Journal of Man-Machine Studies*, 21, 365–375.

Norman, K.L., 1991. *The psychology of menu selection: Designing cognitive control of the human/computer interface*. Norwood, NJ: Ablex.

Palay, A.J., Hansen, W.J., Kazar, M.L., Sherman, M., Wadlow, M.G., Neuenorffer, T.P., Stern, Z., Bader, M., and Peters, T., 1988. The Andrew toolkit – an overview. *In*: *Proceedings of the Winter USENIX Conference*. Berkeley, CA: USENIX Association, 9–21.

Paternò, F., 2000. Model-based design and evaluation of interactive applications. London: Springer.

Powers, W.T., 1973. *Behavior: The control of perception*. Chicago, IL: Aldine.

Preece., J., Rogers, Y., Sharp, H, Benyon, D., Holland, S., and Carey, T., 1994. *Human-computer interaction*. Wokingham, England: Addison-Wesley.

Sanders, M.S. and McCormick, E.J., 1993. *Human factors in engineering and design*. New York, NY: McGraw-Hill.

Stevens, J., 1996. *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Lawrence Erlbaum.

Snowberry, K., Parkinson, S.R., and Sisson, N., 1983. Computer display menu. *Ergonomics*, 26, 699–712.

Taylor, M.M., 1988. Layered protocol for computer-human dialogue. I. Principles. *International Journal of Man-Machine Studies*, 28, 175–218.

Taylor, M.M., 1989. Response timing in layered protocols: a cybernetic view of natural dialogue. *In*: M.M. Taylor, F. Néel, and D.G. Bouwhuis, eds. *The structure of multimodel dialogue*. Amsterdam: Elsevier Science, 159–172.

Taylor, M.M., Farrell, P.S.E., and Hollands, J.G., 1999. Perceptual control and layered protocols in interface design. II. The general protocol grammar. *International Journal of Human-Computer Studies*, 50, 521–555.

Taylor, S. and Todd, P.A., 1995. Understanding information technology usage: a test of competing models. *Information Systems Research*, 6, 144–176.

Taylor, M.M. and Waugh, D.A., 2000. Multiplexing, diviplexing, and the control of mutimodal dialogue. *In*: M.M. Taylor, F. Néel, and D.G. Bouwhuis, eds. *The structure of multimodal dialogue II*. Amsterdam: John Benjamins, 439–456.

Vallacher, R.R. and Wegner, D.M., 1987. What do people think they're are doing? Action identification and human behavior. *Psychological Review*, 94, 3–15.

Zhang, J. and Norman, D.A., 1994. Representation in distributed cognitive tasks. *Cognitive Science*, 18, 87–122.

Appendix 1: Questionnaire template

The six perceived ease-of-use questions taken from the perceived usefulness and ease-of-use (PUEU) questionnaire (Davis 1989). © 1986 by the Regents of the University of Minnesota; used with permission.

Questions

(1) Learning to operate *[name]* would be easy for me.
(2) I would find it easy to get *[name]* to do what I want it to do.
(3) My interaction with *[name]* would be clear and understandable.
(4) I would find *[name]* to be flexible to interact with.
(5) It would be easy for me to become skilful at using *[name]*.
(6) I would find *[name]* easy to use.

Scale

Unlikely	O	O	O	O	O	O	O	Likely
	extremely	quite	slightly	neither	slightly	quite	extremely	
	1	2	3	4	5	6	7	