Design Parameters of Rating Scales for Web Sites

PAUL VAN SCHAIK
University of Teesside
and
JONATHAN LING
Keele University

The effects of design parameters of rating scales on the perceived quality of interaction with web sites were investigated, using four scales (Disorientation, Perceived ease of use, Perceived usefulness and Flow). Overall, the scales exhibited good psychometric properties. In Experiment 1, psychometric results generally converged between two response formats (visual analogue scale and Likert scale). However, in Experiment 2, presentation of one questionnaire item per page was better than all items presented on a single page and direct interaction (using radio buttons) was better than indirect interaction (using a drop-down box). Practical implications and a framework for measurement are presented.

Categories and Subject Descriptors: H.1.2 [**Models and Principles**]: User/Machine Systems—*human factors; human information processing, software psychology*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*ergonomics; evaluation / methodology; graphical user interfaces (GUI); Interaction styles (e.g., commands, menus, forms, direct manipulation), screen design (e.g., text, graphics, color), style guides*; H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*web-based interaction*; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia

General Terms: Human Factors; Measurement

Additional Key Words and Phrases: Human-computer interaction, web site, psychometrics, online questionnaires, interaction mechanism, questionnaire layout, response format, visual analogue scale, Likert scale, screen design.

ACM Reference Format:

van Schaik, P. and Ling J. 2007. Design parameters of rating scales for web sites. ACM Trans. Comput.-Hum. Interact. 14, 1, Article 4 (May 2007), 35 pages. DOI = 10.1145/ 1229855.1229859 http://doi.acm.org/ 10.1145/1229855.1229859

Authors' addresses: P. van Schaik, School of Social Sciences and Law (Psychology Subject Group), University of Teesside, Teesside, UK; email: p.van-schaik@tees.ac.uk; J. Ling, School of Psychology, Keele University, Keele, UK; email: j.r.ling@psy.keele.ac.uk.

1. INTRODUCTION

Surveys and questionnaires are increasingly being conducted on the World Wide Web (Web) (see Buchanan [2000]) and have great potential for the online measurement of users' perceptions. There is an extensive literature on designing and administering surveys printed on paper to gather factual information (e.g., Sapsford [1999]) and psychometric questionnaires (to measure people's perceptions (e.g., Kline [2000]). There is also research on surveys for gathering factual information online [Couper et al. 2004; Tourangeau 2004]. However, there is a lack of research on the design of web-based psychometric questionnaires. In this article, we report two experiments investigating three parameters of questionnaire design:

(1) *response format* (a discrete Likert scale versus a visual analogue scale);
(2) *questionnaire layout* (single question per page versus multiple questions per page);
(3) *interaction mechanism* (radio buttons versus drop-down menus for response selection).

We examined the effect of these parameters on psychometric properties of questionnaires. This type of research applies to situations where people's perceptions are measured using rating scales rather than facts (as in surveys). In particular, our work focuses on the use of rating scales for measuring the quality of human-computer interaction in web sites. Specifically, we studied the psychometric properties of existing psychometric instruments for measuring the following four aspects of the quality of human-computer interaction:

(1) perceived ease of use;
(2) perceived usefulness;
(3) disorientation;
(4) flow.

The questionnaire items that make up these instruments are described in more detail in Section 2.

In order to measure the effectiveness of the questionnaires, we asked participants in our experiments to navigate and find information in two versions of the same web site. The content of the two versions was identical, but they were designed so that the quality of interaction would be poor in one version compared to the other. By creating this deliberately marked contrast between the two versions, it was possible to predict differences in perceived ease of use, perceived usefulness, disorientation and flow and to determine to what extent the predictions were confirmed, that is to the extent to which the instruments were sensitive to differences in web designs. In the next section, we define and describe properties to evaluate psychometric instruments as well as the instruments we used in our research. The three subsequent sections describe the design of the versions of the web site, and two experiments conducted using these. The final sections discuss the results and future research and present final conclusions.

1.1 Psychometrics and Human-Computer Interaction

This article addresses psychometric measurement in Human-Computer Interaction in the context of web pages. In the psychometric measurement of perceptions using questionnaires, spontaneous responses are required that are based on a first reading of each item and without long deliberation or reference to earlier responses. Meeting this requirement could be facilitated by presenting items in isolation without distracting from other items. In contrast, surveys (not studied in this article) require responses that are factually accurate, and therefore respondents may go back to their answers to previous questions and make corrections.

One application of online questionnaires is the measurement of users' experience of web sites. Many variables related to this experience, such as disorientation [Ajuha and Webster 2001], can significantly impact on or be affected by users' performance with and acceptance of these sites. However, a prerequisite for the use of questionnaires that measure constructs, such as disorientation, is that they are psychometrically sound. Psychometrics is a branch of psychology that focuses on the operationalization of variables for the purposes of measurement [Vogt 1999]. For Human-Computer Interaction, four key aspects of the quality of questionnaires are factor structure, reliability, validity and sensitivity [Lewis 2002].

Factor structure is commonly assessed by using factor analysis. The goal of factor analysis is to identify the underlying structure of a set of questionnaire items by reducing them to a smaller number of underlying factors. Establishing factor structure is important as a first step towards identifying a set of items that will measure each of the underlying aspects of human-computer interaction that are to be assessed. For example, Davis and Wiedenbeck [2001] found that nine questionnaire items for the measurement of flow could be reduced to two factors, one measuring users' involvement in interaction with an application and the other measuring users' perception of control over the application. The *reliability* of the items making up the factors is an important property of psychometric instruments because by using reliable scales, measurement error is reduced. One major type of reliability is internal consistency. This is the degree to which the items that make up a factor are related and is usually assessed by employing Cronbach's coefficient alpha. If alpha is sufficiently high (> 0.70), then the items are added up or averaged to produce a scale, thereby reducing the larger set of item scores to a single scale value. Reliability is a prerequisite for *validity*, which is the degree to which an instrument measures what it purports to measure. Validity is an important property of instruments because this clarifies what they measure in relation to other instruments. Two main types are *discriminant validity* and *criterion-related validity* [Bagozzi et al. 1992], both usually analyzed with Pearson's correlation coefficient r. Discriminant validity appraises the level of differentiation between measures of distinct constructs. Criterion validity assesses the relationship between one indicator of a construct (e.g., relaxation) and another construct (e.g., heart rate variability) that are expected to co-vary. *Sensitivity* (or responsiveness) is the ability of a scale to discriminate among various systems, user

populations or tasks. In order to be useful, an instrument needs to be sensitive, that it needs to have the power to detect differences that are expected to exist. Usually, analysis of variance is used to test sensitivity, for example with system and user type as independent variables.

There is a dearth of research simultaneously investigating the psychometric quality of a range of key measures of interaction when presented online. Previous research with online questionnaires has focused on surveys (e.g., Couper et al. [2004] and Norman et al. [2001]), which usually measure factual information rather than on psychometric instruments.

The content of online questionnaires is of paramount importance. In the context of web-based systems, recent research has also investigated the psychometric properties of some key constructs in Human-Computer Interaction and hypermedia research. Davis [1989] introduced and measured two key concepts, perceived ease of use and perceived usefulness, and investigated their role within the framework of technology acceptance. *Perceived ease of use* was defined as the extent to which an individual believes that using a computer system will be free of effort [Davis et al. 1989, p. 985]. Davis and Wiedenbeck [2001] found that their perceived ease of use scale possessed reliability and criterion validity. Ahuja and Webster [2001] confirmed the reliability of their perceived ease of use scale, but found a lack of sensitivity to navigation support in website designs. *Perceived usefulness* was defined as the degree to which a person believes that a computer system will aid job performance [Davis et al. 1989, p. 985]. Davis and Wiedenbeck's results showed that their perceived usefulness scale was reliable. The current study used Davis and Wiedenbeck's perceived ease of use (consisting of three items) and perceived usefulness scales (consisting of four items; see Appendix A).

Disorientation has been defined as "the feeling experienced by users who do not know where they are within hypertext documents [such as web sites] or how to move to desired locations" [Ahuja and Webster, 2001, p. 16]. Using factor analysis, Ahuja and Webster found that disorientation and perceived ease of use were two distinct factors and that their disorientation scale possessed reliability, discriminant validity and sensitivity to navigation support. The current study employed Ajuha and Webster's disorientation scale (comprised of seven items; see Appendix A).

Flow is a psychological state in which a person feels cognitively efficient, motivated and happy [Moneta and Csikszentmihalyi 1996, p. 277]. When people are in the state of flow, they become absorbed in their activities and irrelevant thoughts and perceptions are screened out [Chen et al. 1999]. Davis and Wiedenbeck [2001] established that their discrete scale for measuring flow possessed a two-factor structure. The flow scale was reliable and sensitive to both training and interaction style. This scale was used in the present study (consisting of four items for involvement and five items for control; see Appendix A).

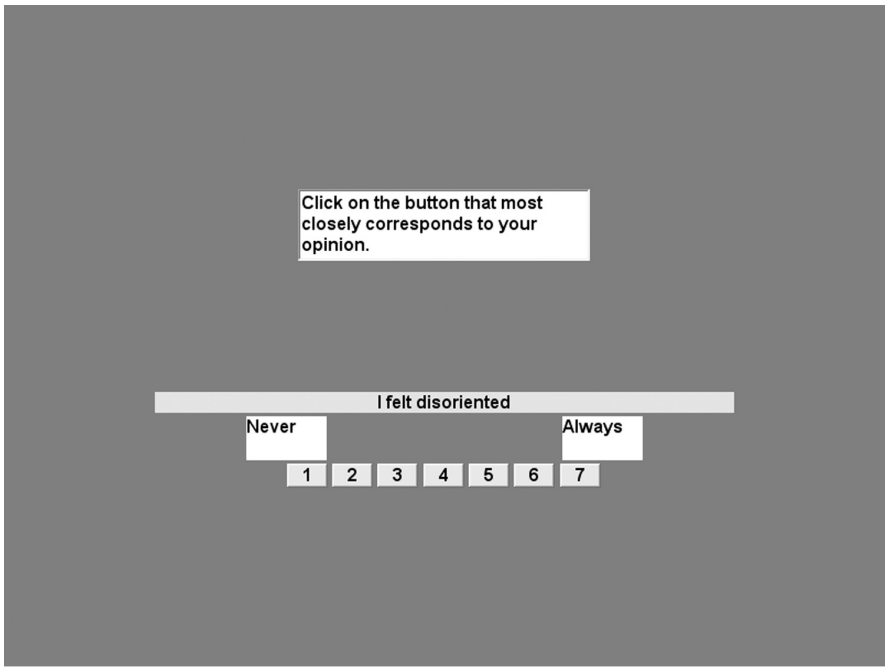
In summary, a number of psychometric criteria exist on which scales for measuring the quality of interaction in terms of key concepts can be assessed. The psychometric properties of questionnaires may depend on the presentation—in terms of *response format* and *questionnaire layout*—as well as *interaction mechanisms* that are employed in the administration of scale items.

1.1.1 *Response Format.* A large corpus of research has demonstrated differences in responses according to the form of survey mode employed (see, e.g., Dillman and Christian [2005]); however, this work has not addressed psychometric measurement. Two types of *response format* have been used in psychological and health research using psychometric instruments: discrete (typically a Likert scale; see Figure 1(a)) and analogue (usually a visual analogue scale; see Figure 1(b)).

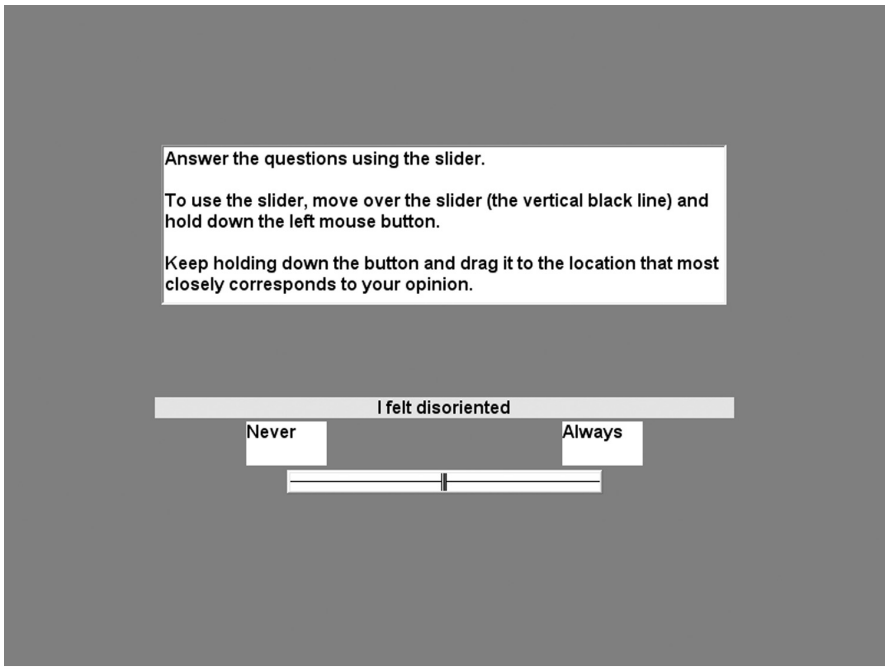
In Human-Computer Interaction, discrete *response formats* are typically used rather than analogue ones (Gillan and Cooke, 1995); however, a scientific justification for this choice in this field is lacking. Discrete *response formats* such as Likert frequently present seven graduated categories to choose from (although 5- and 9-point scales are also used), anchored with descriptive phrases usually representing only the lowest and highest response categories. Respondents select the category most representative of, for instance, their perceived quality of interaction with a web site. In contrast, continuous *response formats* such as visual analogue scale are frequently presented as a 10-cm horizontal line, anchored with two verbal descriptors at the extremes (e.g., strongly agree and strongly disagree); respondents indicate their perceived status by placing a mark along the horizontal line at the most appropriate point. Advantages of Likert and visual analogue scale formats reported by van Schaik and Ling [2003a] are presented Table I. Van Schaik and Ling concluded from their review of mainly medical research on *response formats* that Likert and visual analogue scales may differ across a number of properties, including level of rating, reproducibility and responsiveness. The majority of direct comparisons of *response formats* has focused on subjective assessments of pain or fatigue. However, all of these properties are important for the evaluation of users' interaction with computers, but there is currently a lack of research measuring the quality of this interaction.

In relation to the potential difficulty of using a visual analogue scale, Pfennings et al. [1995] found a higher variability of scores when using a 10-cm visual analogue scale compared to 10-point Likert. In contrast, other work has found significant positive correlations between ratings of acute pain between 10-point Likert and 10-cm visual analogue scale [Murphy et al. 1988] and ratings of fatigue using 5-point Likert and 10-cm visual analogue scale [Brunier and Graydon 1996]. In terms of sensitivity, Price et al. [1994] showed that, when rating intensity and experience of pain, a 15-cm visual analogue scale and an 11-point Likert (ranging from 0 to 10) were equally sensitive to differences in temperature and Hayes et al. [1996] found that both a visual analogue scale and a 5-point Likert for measuring pain were sensitive to differences in the treatment of corneal rust ring. However, Likert scales with fewer than seven response alternatives or different scale lengths may produce nonequivalent results. For example, using a smaller number of response alternatives (4-point Likert) resulted in a lack of responsiveness for Likert compared to visual analogue scale [Bellamy et al. 1999a, 1999b; Joyce et al. 1975].

In summary, different *response formats* each have their own advantages and disadvantages, although previous research has found some evidence for equivalence of *response formats* when Likert with a larger number of response



(a)



(b)

Fig. 1. *Response formats* used in Experiment 1. (a) Likert *response format*; (b) Visual analogue scale *response format*.

Table I. Reported Advantages and Disadvantages of Likert and Visual Analogue Scale Response Formats (van Schaik & Ling, 2003a)

	Likert	Visual analogue scale
Advantages	Relatively easy to learn because all possible responses are presented.	Effect of individual interpretation of Likert graduations avoided.
	Relevant changes in scores more easily interpreted by researchers.	Better match between subjective state and response through very large response range.[a]
Disadvantages	Poorer match between subjective state and response because of restricted range of responses.	Difficulty in (learning to) use because of lack of indication of intermediate points (only end-points are displayed).
	Variability due to individual interpretation of Likert graduations.	Extra work required to convert analogue responses into numeric scores after data collection.

[a]This is often assumed, but is not consistent with the research cited in Nunnally and Bernstein [1994] on the effect of increase the number of scale steps on the reliability of scales, where scales become more reliable with an increasing number of scales steps, but the with rapidly diminishing returns; in particular, after 11 steps, reliability increases very little.

categories used. However, there is a lack of research that studies instruments measuring the quality of human-computer interaction online.

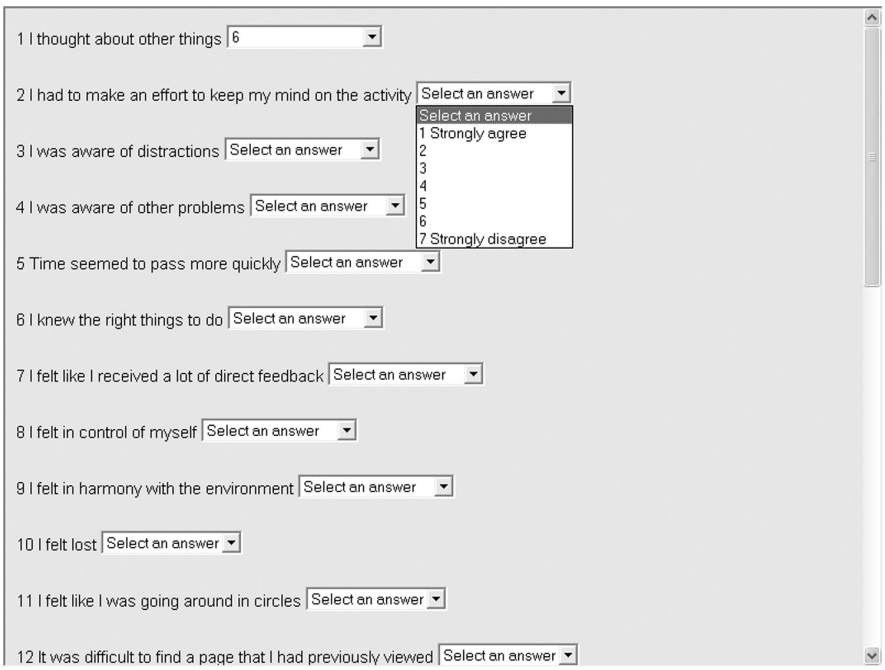
1.1.2 *Questionnaire Layout.* Norman et al. [2001] studied the effect of another parameter in online questionnaire design, *questionnaire layout*, on speed and subjective measures in one data entry task and two editing tasks. Four types of partitioning a survey were used: whole form (all items on one page, which - as a result - required scrolling), semantic sections (items of related sections, which required scrolling, depending on the group size of related items), screen pages (each page was "filled up" with items, but no scrolling was required) and single items (one item per page and no scrolling was required). These four types were combined with two levels of navigation support (present, by having an index to access other parts of the survey, and absent, without an index) to produce eight online survey designs. (Examples of whole-form designs used in the present study are shown in Figures 2(a) and 2(c) and examples of single-items designs are shown in Figures 2(b) and 2(d).) An advantage of single-items designs is that there is no distraction by other items (as in the other three designs) because these are each presented on separate pages. In Norman et al.'s [2001] data entry task, online survey design did not affect completion time. However, in a text-editing task, screen pages with an index and single items with an index were faster than whole form and semantic sections. In the numeric editing task, single items with an index were faster than the other designs and screen pages with an index and single items without an index were the slowest. No systematic differences occurred between online survey designs in terms of subjective measures.

Norman et al.'s [2001] results, showing advantages in terms of speed for certain online survey designs, demonstrate that *questionnaire layout* is an important factor in completing online surveys (see also Couper et al., 2004). However, the tasks where differences in speed were found—survey editing of text and

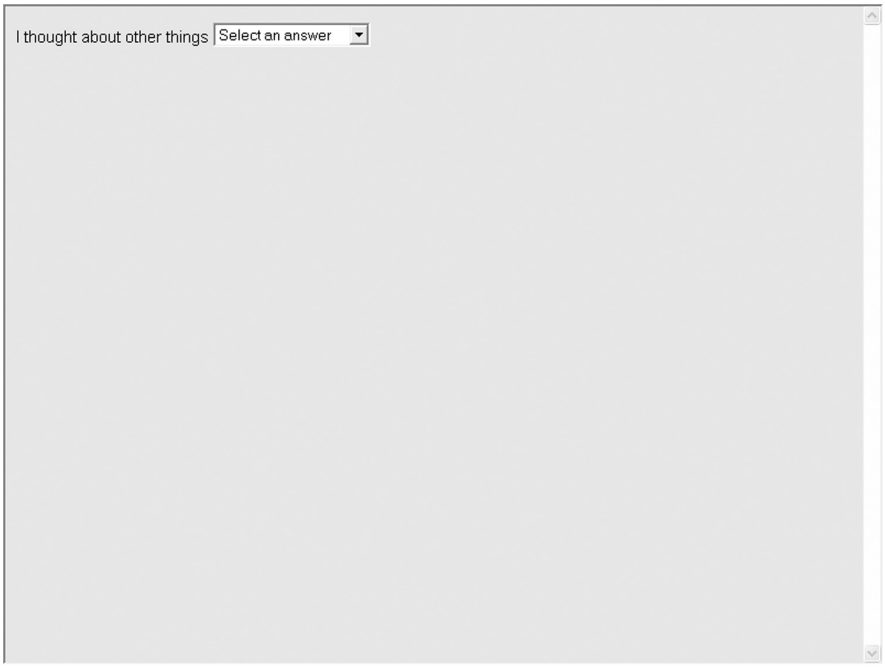
(a)

(b)

Fig. 2. *Interaction mechanism*s and *questionnaire layout* used in Experiment 2. (a) Direct interaction, whole form (b) Direct interaction, single items (c) Indirect interaction, whole form (d) Indirect interaction, single items (*continues*).



(c)



(d)

Fig. 2. (*Continued*).

numerical information—do not apply to psychometric instruments, which are the focus of the current study. In particular, in surveys, factual data are frequently collected and factual accuracy is an important consideration, which can be improved by making changes to responses, in contrast to psychometric data collection where changes are not desired. Furthermore, navigation support (through an index in Norman et al.'s online surveys) does not usually apply when completing psychometric scales, because spontaneous responses are required that are based on a first reading of each item and without long deliberation or reference to earlier responses. In summary, *questionnaire layout* has been found to effect responses on online questionnaires, but this has not been investigated with psychometric questionnaires.

1.1.3 *Interaction Mechanism.* A further parameter in the design of online questionnaires is *interaction mechanism* [van Schaik and Ling 2003a]. With the use of radio buttons (see Figures 2(a) and 2(b)), all responses are readily visible and (after moving the cursor to the desired response) selection involves a single action (usually clicking a mouse button). When a drop-down list is employed (Figures 2(c) and 2(d)), responses are not visible until a respondent selects the list and response selection then potentially involves scrolling through the list, followed by selecting one response. Compared to a drop-down list, an advantage of radio buttons is that no action is required to make responses visible. An advantage of a drop-down list is a reduction in the potential for distraction caused by the presentation of the range of response categories of other items on the same page; however, this advantage only occurs when there is more than one item on a page (see Couper et al. 2004).

In their study, Couper et al. [2004] investigated the effect of *response format* on user responses to web sites. Participants were presented with a series of questions, which they answered using radio buttons, a drop box with no options initially visible (participants had to click on the box to view potential responses), and a drop box which displayed five options, but required scrolling to display remaining ones. Couper et al. found that *response format* had a significant effect on responses, but the effect of visibility of response options was stronger, with visible response options endorsed more frequently. In summary, research has shown that *interaction mechanism* affects responses in online questionnaires, but psychometric instruments have not been investigated.

1.1.4 *The Current Study.* The preceding review of research has indicated first, a lack of evidence for the psychometric properties of psychometrics instruments when simultaneously measuring several key concepts in the quality of experience of web sites, and second, a lack of empirical evidence for the best *response format*, *interaction mechanism* and layout for questionnaires measuring the quality of users' interaction with computers. It is essential to study psychometric properties and both presentation format and *interaction mechanism* simultaneously rather than separately in order to produce generalizable results. Fundamentally, if the two factors are tested separately, it will not be possible to establish if they are independently having an effect and the generalizability of findings would be unknown at best. If both factors are tested

simultaneously and if the results of statistical analysis show a moderator effect [Jaccard 1998], this would demonstrate that the effect of presentation format depends on the type of *interaction mechanism* and vice-versa and thus demonstrate a limitation in generalizability. A comprehensive set of measures rather than separate measures need to be presented in an optimal way in order to simultaneously measure the quality of interaction comprehensively. Certain presentation formats may produce better results than others; researchers and practitioners need research evidence that allows them to use the format with the best psychometric properties for a range of key concepts in the quality of human-computer interaction. The current study therefore aimed to compare two *response formats* (7-point Likert and 10-cm visual analogue scale), two *interaction mechanism*s (direct and indirect) and two *questionnaire layout* (single items and whole form) in terms of psychometric properties of existing psychometrics instruments for measuring four key concepts in the quality of interaction with web sites (disorientation, perceived ease of use, perceived usefulness and flow). In this way, the study investigated the HCI design of questionnaires that measure the quality of interaction with web sites.

1.2 Design of the Web site

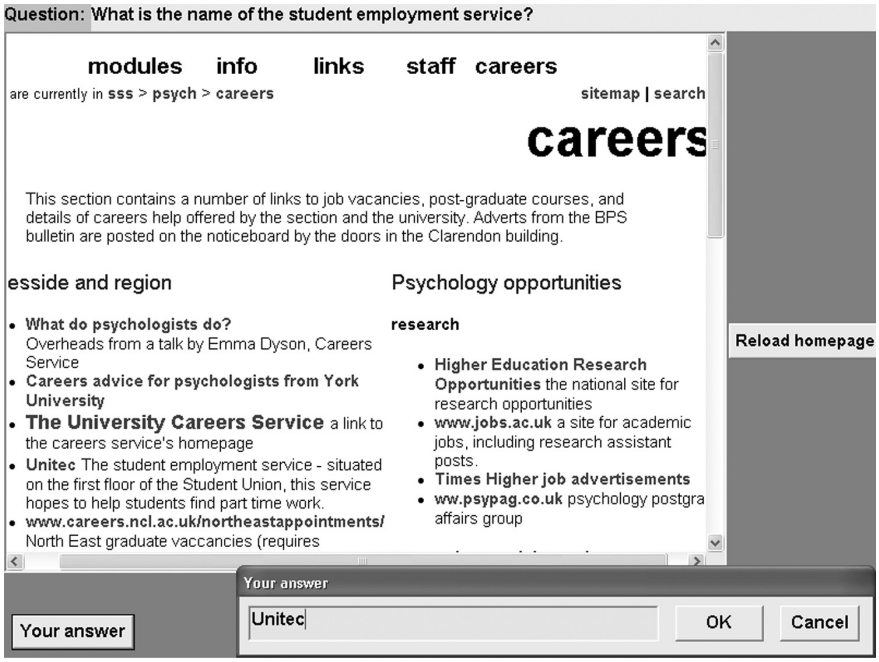
One method to establish the sensitivity of measures of the quality of interaction and used in the current study is to compare measurements between different human-computer interfaces [Lewis 2002]. Methods from experimental psychology are used to make systematic variations to user interfaces of particular web sites, using different web designs, and observe their effect on outcome measures. This can be achieved by complying with or violating (user interface or cognitive) *design principle*s for web sites, related to, for example, the overall structure of screen design [Ward and Marsden 2003], support for users' schemas of web designs [van Schaik and Ling 2003b], and representation of information categories [Dalal et al. 2000]. In both experiments reported in this article, the following *design principle*s were applied in one version of the web site or violated in the other version (see Figure 3):

—preserve the context of information units [Dalal et al., 2000]: headings of "content area subregions" [Blackmon et al. 2005] were included or removed, for example "Teesside regions" and "Psychology opportunities";

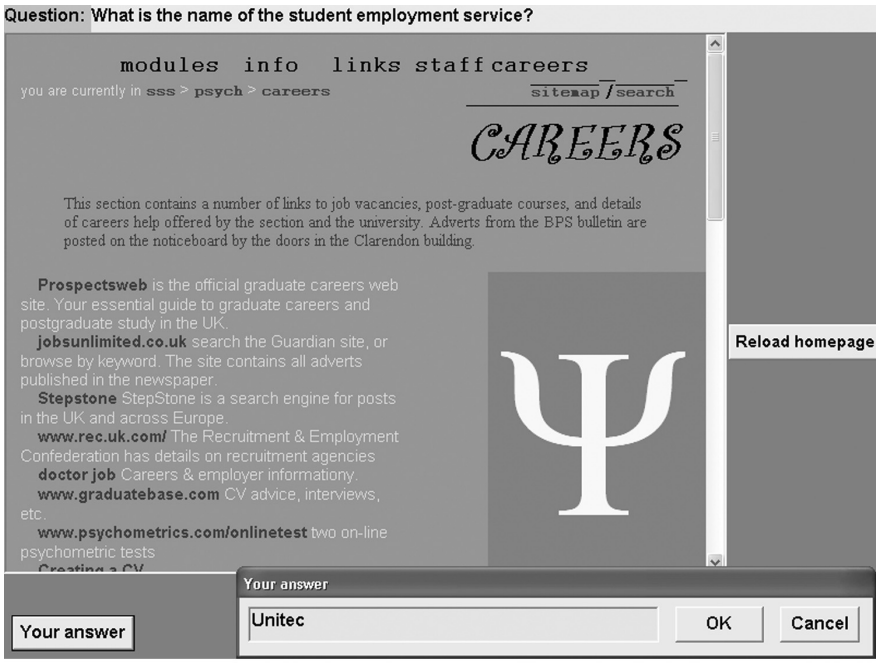
—use higher-order information units (Dalal et al. 2000): (hyper)links were grouped into content area subregions or one long list was presented, for example headings for the previously mentioned subregions and subheadings "research" and "postgraduate training and careers";

—avoid gratuitous animation [Ward and Marsden 2003]: an unnecessary logo in the form of a conspicuous Ψ symbol (representing the discipline of 'Psychology') was present or absent;

—be consistent [Ward and Marsden, 2003]: the same shape of mouse cursor was used consistently over different text areas in a web page or different shapes were used over different text areas; the use of color and type faces was consistent or inconsistent;



(a)



(b)

Fig. 3. Typical Web Pages used in Experiments 1 and 2. (a) *Design principle*s applied (b) *Design principle*s violated.

—use conventions for appearance [Ward and Marsden 2003]: an I-beam-shaped cursor was used for text and a hand-shaped cursor for links (conventions followed) or an I-beam-shaped cursor was used for links and either a hand-shaped or an arrow-shaped cursor for text (conventions violated);
—use color contrast to enhance readability [Ware 2004]: good contrast (principle followed) or poor contrast (principle violated) between text and background of web pages.

The research used an existing intranet site, developed by the University of Teesside's Psychology subject group for its students. The site consisted of a home page, six main content pages (see Figure 3), a sitemap, a search page and links to several hundred further pages of content, including lecture notes, generic study information, careers advice and staff details as well as links to other university intranet sites and external Internet sites. Two versions of the web site were produced, which were identical in content, links and navigation structure. The two versions differed in that *design principle*s were applied in one version and violated in another. This manipulation of the HCI design of the web site was necessary to study the sensitivity of the questionnaires used in this study and is separate from the manipulation of the HCI design of the questionnaires (in terms of their *response format*, *interaction mechanism* and *layout*).

Participants undertook a series of 10 tasks (information retrieval questions) by finding information in the intranet site (see Appendix B) and were required to answer all the questions. Two experiments were conducted, using the two versions of the site, in order to establish the effect of *response format*, *interaction mechanism* and *questionnaire layout* on the properties of psychometric instruments for measuring the quality of interaction.

2. EXPERIMENT 1

Experiment 1 aimed to establish the extent to which two different formats of online questionnaires are equivalent in terms of their psychometric properties.

2.1 Method

2.1.1 *Experimental Design.* The experiment used a 2x(2) experimental design with two factors. The first, independent measures, factor was application of *design principle*s with two levels: *design principle*s applied or violated. The second, repeated measures, factor was *response format* also with two levels: Likert (7-point) and visual analogue scale, with single-items presentation (see Figure 1). Participants were allocated to screen designs with *design principle*s that were either applied or violated and order of *response format* was counterbalanced. Therefore, participants were allocated into one of four separate conditions: *design principle*s applied and Likert first, *design principle*s applied and visual analogue scale first, *design principle*s violated and Likert first, and *design principle*s violated and visual analogue scale first. Participants gave responses to three questionnaires: the disorientation and perceived ease of use questionnaires developed by Ahuja and Webster [2001], and the flow questionnaire used by Davis and Wiedenbeck [2001].

2.1.2 *Participants.* There were 103 participants, consisting of 84 females and 19 males, with a mean age of 25 years (SD = 8.38). All were undergraduate psychology students who took part in the experiment as a course requirement. Of these, 50 took part in the conditions with *design principle*s applied (with 24 completing the Likert statements first and 26 completing the visual analogue scale statements first) and 53 took part in the conditions with *design principle*s violated (with 28 completing Likert statements first and 25 completing visual analogue scale statements first). All participants had experience using the Web and the vast majority had been using the Web for more than one year (92%). Frequency of Web use varied from more than once a day to less than once a month, with the majority (60%) using the Web at least once a day.

2.1.3 *Materials and Apparatus.* The experiment ran on personal computers (Intel Pentium, 333 MHz, 64 Mb RAM, Microsoft NT4 operating system, 14-inch monitors). The screen dimensions were 800×600 pixels. Contrast and brightness were set to optimal levels.

Participants also completed a series of three on-screen questionnaires to measure the three key concepts in the quality of the web site designs being tested. The first measured disorientation and comprised of seven statements, such as "I felt like I was going round in circles". The second assessed perceived ease of use and consisted of three statements, such as "learning to use the site was easy". The third measured participants' intensity of flow and consisted of nine statements, such as "I thought about other things".

Measurements were made using two *response formats* for each questionnaire: Likert and visual analogue scale. The Likert *response format* was presented on a 7-point scale, which ranged from "Strongly agree" to "Strongly disagree" at either end of the scale (see Figure 1(a)). Responses to statements presented in this format were given by clicking on a button beneath one of these points. Responses to the visual analogue scale statements were given by dragging a slider along the scale (see Figure 1(b)). The slider always started at the midpoint of the scale. This *response format* did not include subdivisions, although "Strongly agree" and "Strongly disagree" were presented at either end of the scale. After making a response, participants had to explicitly take an action to either amend their response or proceed (to the next question or to the next part of the experiment).

2.1.4 *Procedure.* The experiment consisted of two parts: an information retrieval task, followed by a questionnaire. The experiment was run in a computer lab with groups of 15 to 20 participants who worked independently. In the information retrieval task, typical tasks were included that users perform with web sites (see e.g., van Schaik and Ling [2003b]. Participants first completed a practice run consisting of three trials to accustom them to the site before moving on to the main information retrieval task, which had 10 further trials. Examples of tasks included "What is PsycINFO?", "What is the name of the student employment service?" and "Who deals with requests for coursework extensions?". In each trial, a question appeared at the top of the screen. Once participants had read the question, they had to click on a button labeled "Show

web site". The home page of the intranet site then appeared on the screen and they had to find the answer to the question using the site. Participants were told to take the most direct route possible to locate the answer. Having found the answer, they clicked on a button labeled "Your answer", which opened a dialog box at the bottom of the screen. Participants typed their answers into the box and, after clicking on "OK", moved on to the next question. After the information retrieval task, participants completed the questionnaires using both Likert and visual analogue scale as *response formats*. Following this, participants were first asked which *response format* they preferred and then presented with a series of four screens containing text boxes into which they were asked to type advantages and disadvantages of both *response formats* that they had identified. The experiment took approximately 45 minutes to complete.

3. RESULTS

The data were analyzed to evaluate the effects of *design principle*s on task performance and navigation behavior, the psychometric properties of the questionnaires, and participants' preference of *response format*.

3.1 Task Performance and Navigation Behavior

Prior to analysis of the psychometric instruments, objective measures were analyzed to establish that the two versions did differ in their level of usability. A t test showed that the effect of *design principle*s on average number of additional visits to the homepage was significant, $t\,(101) = -17.56$, $p < 0.001$, $\eta^2 = 0.75$ (*design principle*s applied: mean = 0.002; SD = 0.014; *design principle*s violated: mean = 1.46; SD = 0.59). The effect of *design principle*s on percentage of correct answers was also significant, $t\,(101) = 2.51$, $p < 0.05$, $\eta^2 = 0.06$ (applied: mean = 88%; SD = 16%; violated: mean = 81%; SD = 11%). The effect of *design principle*s on number of links visited before a correct answer was close to significance, $t\,(101) = -1.95$, $p = 0.054$, $\eta^2 = 0.04$ (applied: mean = 3.88; SD = 1.28; violated: mean = 4.59; SD = 2.26). The effect of *design principle*s on time-on-task before a correct answer was not significant, $t\,(101) = -1.57$, $p > 0.05$. These results show that the experimental manipulation of *design principles* was effective.

3.2 Psychometric Properties

Psychometric instruments for the evaluation of human-computer interaction need to have a well-defined factor structure and demonstrate reliability, validity and sensitivity [Lewis 2002]. As is clear from Lewis's work, sensitivity—in this case to different web site designs presenting the same content—is not a separate issue, but an integral part of the psychometric evaluation of sites. We investigated these characteristics for the three scales used in this experiment.

3.2.1 *Factor Structure.* Correlations among the following sets of items were all > 0.3 and statistically significant for both *response formats*: the seven disorientation items, the three perceived ease of use items and the first set of four flow items (the second set of four flow items in Davis and Wiedenbeck [2001]) for both *response formats*. However, the correlations within the second

Table II. Factor Analyses (Experiment 1)

	Likert			Visual analogue scale		
Item	Factor 1: Disorientation	Factor 2: Ease of use	Factor 3: Flow	Factor 1: Disorientation	Factor 2: Ease of use	Factor 3: Flow
FLO1	−0.082	−0.172	**0.723**	−0.049	−0.063	**0.714**
FLO2	0.145	−0.134	**0.693**	0.372	0.037	**0.410**
FLO3	−0.040	0.074	**0.755**	−0.024	−0.030	**0.570**
FLO4	0.072	0.069	**0.697**	0.033	0.109	**0.781**
EOU1	0.096	**0.801**	−0.069	0.015	**0.814**	0.055
EOU2	−0.016	**0.822**	−0.035	0.075	**0.971**	0.008
EOU3	−0.162	**0.848**	0.000	−0.189	**0.742**	−0.089
DIS1	**0.670**	−0.198	−0.088	**0.848**	−0.020	0.064
DIS2	**0.869**	−0.054	−0.032	**0.855**	0.011	0.072
DIS3	**0.684**	0.159	0.179	**0.705**	−0.007	−0.008
DIS4	**0.836**	0.131	0.073	**0.852**	0.084	−0.108
DIS5	**0.823**	−0.045	0.013	**0.820**	−0.087	0.011
DIS6	**0.814**	−0.070	−0.041	**0.852**	−0.070	0.071
DIS7	**0.843**	0.042	0.001	**0.876**	−0.041	−0.068

Notes. N = 103. Extraction method: principle axis factoring; rotation method: direct oblimin. FLO: flow; EOU: perceived ease of use; DIS: disorientation. Items are numbered per scale, for example EOU1: perceived ease of use, Item 1. Figures are factor loadings.

set of five flow items (Davis and Wiedenbeck's first set of five items) were generally low; therefore, this set was unsuitable for factor analysis and not analyzed further.

A three-factor solution, using principle axis factoring and direct oblimin (oblique) rotation, showed simple structure (i.e., there were a number of items loading highly on each factor and each item only loaded highly on one factor) and confirmed the following factors (1) disorientation, (2) perceived ease of use and (3) intensity of flow (defined by the first set of four flow items) for both *response formats* (see Table II). The two *response formats* possessed essentially the same factor structure and the percentage of explained variance was similar (unrotated three-factor solutions explained 64% of variance with visual analogue scale and 65% with Likert format). Subsequently, we investigated the psychometric properties of the scales *Disorientation*, *Perceived ease of use* and *Intensity of flow*, as defined by the items that loaded highly on each of the three factors.

3.2.2 *Reliability*. Employing Cronbach's alpha, the scales *Disorientation* (alpha = Likert: 0.92, visual analogue scale: 0.94), *Perceived ease of use* (alpha = Likert: 0.87, visual analogue scale: 0.88) and *Intensity of flow* (alpha = Likert: 0.82, visual analogue scale: 0.74) all proved to be reliable for both formats. Subsequently, overall scores were calculated by averaging item scores per scale and overall scores for Likert format were converted to the range [0; 100]. (For visual analogue scale, the range was already [0; 100].)

3.2.3 *Validity*. Regarding discriminant validity, for the Likert format all three correlations between pairs of *Intensity of flow*, *Perceived ease of use* and *Disorientation* were moderate and significant (see Table III). The same pattern of correlations was found for the visual analogue scale format, with significant

Table III. Correlations between Scales (Experiment 1).

	Intensity of flow (Likert)	Disorientation (Likert)	Perceived ease of use (Likert)
Intensity of flow (visual analogue)	**0.884	**0.311	*−0.232
Disorientation (visual analogue)	**0.410	**0.867	*−0.246
Perceived ease of use (visual analogue)	−0.162	**−0.372	**0.555

Note. Upper right: correlations between items with Likert format; lower left: correlations between items with visual analogue format; diagonal: correlations of Likert with visual analogue for the same items.
*Correlation significant at the 0.05 level (two-tailed).
**Correlation significant at the 0.01 level (two-tailed).

correlations of *Disorientation* with *Intensity of flow* and *Perceived ease of use.* There was no significant correlation between *Intensity of flow* and *Perceived ease of use*. Correlations between the two *response formats* were significant for all three scales, although they were higher for *Intensity of flow* and *Disorientation*, ($\geq$ 75% overlap in variance) than for *Perceived ease of use* (31% overlap). Tests for differences in correlations between subjective measures between the two *response formats*, using Fisher's z for nonindependent groups, showed that none of the correlations differed significantly.

Criterion validity of psychometric scales can be established by calculating correlations between scales, but also by calculating correlations between scales and other measures of the quality of interaction, such as measures of task performance and navigation behavior. Task performance is frequently measured in terms of speed (usually measured as time-on-task) and accuracy (often measured as percentage of correct answers). Typical measures of navigation behavior include number of pages visited, revisitation rate (which can, incidentally, be considered as a behavioral measure of disorientation), pages visited once, and number and percentage of visits to search pages [Cockburn and McKenzie 2001]). The correlation of *Disorientation* with both of the following measures of task performance was significant: percentage of correct answers, for Likert Pearson's $r = -0.35$, $p < 0.001$, and for visual analogue scale $r = -0.49$, $p < 0.001$, and average time-on-task before a correct answer, for Likert $r = 0.29$, $p < 0.005$, and for visual analogue scale $r = 0.28$, $p < 0.005$. The correlation of *Disorientation* with both of the following behavioral measures was significant: average number of links before a correct answer, $r = 0.34$, $p < 0.001$, for both *response formats*, and number of visits to the homepage, for Likert $r = 0.46$, $p < 0.001$, and for visual analogue scale $r = 0.49$, $p < 0.001$. Tests for differences in correlations between subjective measures and a common objective measure between the two *response formats* showed that the correlation of *Disorientation* with percentage of correct answers was significantly greater with visual analogue scale than with Likert, $t\ (100) = 3.17$, $p < 0.005$.

3.2.4 *Sensitivity.* Using flow, perceived ease of use and disorientation as dependent variables, 2×2 analyses of variance (ANOVAs) were conducted with the independent variables of *design principle*s and *response format*, analyzing responses for first *response format* used. In this way any potential carry-over effect from using one type of format to another on the results was avoided.

A two-way ANOVA revealed that the effect of *design principle*s on *Disorientation*, was significant, $F\ (1, 99) = 26.18$, $p < 0.001$, $\eta^2 = 0.21$, with mean = 29.46

(SD = 14.42) for *design principle*s applied and mean = 48.08 (SD = 21.88) for *design principle*s violated. However, neither the effect of *response format*, F (1, 99) = 1.32, $p > 0.05$, nor the interaction effect, $F < 1$, were significant.

A further ANOVA showed that the effect of *design principle*s on *Perceived ease of use*, was not significant, F (1, 102) = 2.55, $p > 0.05$, $\eta^2 = 0.03$, nor were the effect of *response format* or the interaction effect, both $F < 1$. Another two-way ANOVA demonstrated that the effect (application) of *design principle*s on *Intensity of flow* was not significant, F (1, 102) = 1.82, $p > 0.05$, $\eta^2 = 0.03$. The effect of *response format* and the interaction effect were not significant, both $F < 1$.

3.3 Preference of Response Format

A chi square test showed that a significant majority (82%) preferred the Likert *response format*, χ^2 (1) = 41.02, $p < 0.001$. Based on participants' responses to the open-ended questions presented at the end of the experiment, a number of advantages and disadvantages were identified for both *response formats* (see Table IV). Participants believed that first, visual analogue scale allowed a greater range and Likert a more restricted range of responding, and second, visual analogue scale made it more difficult to give consistent answers between statements. However, in reality, the variability in scores was similar between the two *response formats* and indeed there was a tendency for *higher* variance for Likert, with variance ratios of visual analogue scale to Likert of approximately 0.83 (0.72, 0.81 and 0.96 for *Intensity of flow*, *Perceived ease of use* and *Disorientation*). Participants believed that Likert might lead to a bias towards neutral answers. Nevertheless, after converting overall scores on the three scales to a 7-point range, differences in frequencies for the middle neutral response category were within 8% between the two *response formats* (with *more* neutral answers for visual analogue scale than Likert for *Intensity of flow*, *Perceived ease of use* and *Disorientation*). Participants believed that Likert might lead to the avoidance of extreme responses. However, after conversion to a 7-point range, differences in frequencies for the extreme lowest response category was within 3% between the two *response formats*. Similarly, differences in frequencies for the extreme highest response category were within 2% between the two *response formats*. In Experiment 1, timing data measuring the time spent on completing the items were not collected and therefore not available for analysis.

3.4 Summary of Results

Experiment 1 established the psychometric properties of scales for measuring three key concepts (disorientation, perceived ease of use and flow) in the quality of users' interaction with web sites simultaneously. Even though a strong preference for Likert was observed, overall psychometric results in terms of factor structure, reliability, validity and sensitivity for both *response formats* converged. Respondents' negative perceptions of response tendencies with Likert and visual analogue scale were not confirmed by their actual responses to the scales. The results can be explained in terms of diminishing returns of

Table IV. Advantages and Disadvantages of Response Formats (Experiment 1)

Advantages			
Likert		Visual analogue scale	
Clarity of response	[a]55	Degree of choice	44
Ease/speed of use	22	Ease/speed of use	13
Disadvantages			
Likert		Visual analogue scale	
Difficulty of mapping judgment to 7-point numerical scale	49	Lack of clarity and consistency	68
Response set[b]	7	Usability	22

[a]Frequency.
[b]Responding to questions in a particular way independently of question content.

more scale steps from 7 (with Likert) to 100 (with visual analogue scale) and do not lend support to the cited advantages disadvantages of both scales (see Table I), such as the lack of (visibility of) scale steps with visual analogue scale. Because of the equivalence of *response formats*, the commonly used Likert format was included in Experiment 2 and two other design parameters of online questionnaires were investigated with this format.

4. EXPERIMENT 2

Experiment 2 set out to investigate the extent to which two different *interaction mechanism*s and two different *questionnaire layout* are equivalent in terms of their psychometric properties.

4.1 Method

4.1.1 *Experimental Design.* The experiment used a 2x2x(2) experimental design with three factors. The first, independent measures, factor was application of *design principle*s with two levels: *design principle*s applied or violated. The second, independent measures, factor was *interaction mechanism* for psychometric items also with two levels: selection from a set of immediately visible options using radio buttons ("direct interaction") and selection from a set of options that became visible when interacting with the control—a drop-down list—in which they were embedded ("indirect interaction") (see Figure 2). The third, repeated measures, factor was *page layout* of psychometric items also with two levels: whole form and single items (see Figure 2). Participants were allocated to screen designs with *design principle*s either applied or violated, direct or indirect *interaction mechanism* and order of *questionnaire layout* was counterbalanced. Participants gave responses to four questionnaires: the disorientation and perceived ease of use questionnaires developed by Ahuja and Webster [2001], and the flow and perceived usefulness questionnaires used by Davis and Wiedenbeck [2001].

4.1.2 *Participants, Materials and Procedure.* There were 127 participants, consisting of 100 females and 27 males (mean age: 23 years; SD = 7.45). They were all undergraduate psychology students and took part in the experiment as a course requirement. Of these, 68 took part in the conditions with *design*

*principle*s applied (with 36 completing statements in direct interaction, and 32 completing statements in indirect interaction) and 59 took part in the conditions with *design principle*s violated (with 27 completing statements in direct interaction, and 32 completing statements in indirect interaction). All but one of the participants had experience using the Web and the vast majority had used it for more than one year (96%). Frequency of Web use varied from more than once a day to less than once a month, with the majority (76%) using the Web at least once a day.

The same materials and procedure were used as in Experiment 1, with the addition of a 4-item perceived usefulness questionnaire. Questionnaire responses were presented with direct (see Figures 2(a) and 2(b)) and indirect interaction (see Figures 2(c) and 2(d)) and using two *questionnaire layouts*: whole form (see Figures 2(a) and 2(c)) and single items (see Figures 2(b) and 2(d)). At the end of the experiment, participants were first asked which *questionnaire layout* they preferred and then presented with a series of four screens containing text boxes into which they were asked to type advantages and disadvantages of both *questionnaire layouts* that they had identified.

5. RESULTS

The data were analyzed to evaluate the effects of *design principle*s on task performance and navigation behavior, the psychometric properties of the questionnaires, speed of and changes in completing questionnaires, and participants' preference of *response format*.

5.1 Task Performance and Navigation Behavior

Before the psychometric instruments were assessed, objective measures were analyzed in order to establish that the two versions did differ in their level of usability. A t test showed that the effect of *design principle*s on average number of additional visits to the homepage was significant, $t\ (125) = -4.69$, $p < 0.001$, $\eta^2 = 0.15$ (*design principle*s applied: mean = 1.27; SD = 0.26; *design principle*s violated: mean = 1.62; SD = 0.57). The effect of *design principle*s on percentage of correct answers was also significant, $t\ (125) = 2.89$, $p < 0.005$, $\eta^2 = 0.06$ (applied: mean = 84; SD = 15; violated: mean = 76; SD = 16). Furthermore, the effect of *design principle*s on number of links visited before a correct answer was significant, $t\ (125) = 2.84$, $p < 0.01$, $\eta^2 = 0.06$ (applied: mean = 3.97; SD = 1.70; violated: mean = 4.94; SD = 2.14). The effect of *design principle*s on number of links visited before an incorrect answer was significant as well, $t\ (125) = 2.85$, $p < 0.01$, $\eta^2 = 0.06$ (applied: mean = 6.75; SD = 5.76; applied: mean = 10.46; SD = 8.77) as was the effect of *design principle*s on time-on-task before a correct answer, $t\ (125) = 4.41$, $p < 0.01$, $\eta^2 = 0.14$ (applied: mean = 63s; SD = 40s; applied: mean = 43s; SD = 23s). These findings show that the experimental manipulation of *design principles* was effective.

5.2 Psychometric Properties

Factor structure, reliability, validity and sensitivity were investigated for the four scales used in this experiment.

5.2.1 *Factor Structure.* Correlations among the following sets of items were always > 0.3 and statistically significant for both *interaction mechanism*s when presented with single items and mostly > 0.3 when all presented with whole form: the seven disorientation items, the three perceived ease of use items, the four perceived usefulness items and the first set of four flow items (the second set of four flow items in Davis and Wiedenbeck [2001]) for both *interaction mechanism*s. However, the correlations within the second set of five flow items (Davis and Wiedenbeck's first set of five items) were generally low; this set was therefore unsuitable for factor analysis and not analyzed further.

A four-factor solution, using principle axis factoring and direct oblimin (oblique) rotation was identified, which confirmed the following factors: (1) disorientation, (2) perceived usefulness, (3) perceived ease of use and (4) intensity of flow (defined by the first set of four flow items) for both *interaction mechanism*s (see Table V). Percentages of explained variance were 72% for indirect interaction/single items, 68% for direct interaction/single items, 64% for indirect interaction/whole form and 64% for direct interaction/whole form. Solutions for single items showed mostly simple structure, whereas in the solutions for whole form disorientation items 3 and 4 had severe cross-loadings for both *interaction mechanism*s. Subsequently, the psychometric properties were investigated of the scales *Disorientation*, *Perceived ease of use, Perceived usefulness* and *Intensity of flow*, as defined by their corresponding items. However, for whole-form presentations the items with severe cross-loadings were excluded.

5.2.2 *Reliability.* Reliability values for the scales *Disorientation*, *Perceived ease of use, Perceived usefulness* and *Intensity of flow* are presented in Table VI. Over all four scales, averaged reliability coefficients varied from 0.93 (indirect/single items) to 0.91 (direct/single items) to 0.86 (indirect/whole form) to 0.83 (direct/whole form). Average reliability was 0.84 for single items and 0.92 when for whole form, and average reliability was 0.87 for direct interaction and 0.89 for indirect. Subsequently, overall scores were calculated per scale by averaging item scores.

5.2.3 *Validity.* Pearson's correlation was used to establish validity. Concerning discriminant validity (see Table VII), for direct interaction and both *questionnaire layouts Intensity of flow* was not significantly correlated with the other scales, but the correlations between each of the other scales were moderate and significant. For indirect *interaction mechanism* and both *questionnaire layouts*, all correlations between scales were moderate and significant, except the nonsignificant correlation between *Intensity of flow* and *Perceived usefulness*. Correlations between the two *questionnaire layouts* were significant for all four scales, but highest for *Intensity of flow* (> 70% overlap in variance for both *interaction mechanism*s). Tests for differences in correlations between subjective measures between the two *interaction mechanism*s, using Fisher's z for independent groups, showed that none of the correlations differed statistically significantly. Tests for differences in correlations between subjective measures between the two *questionnaire layout*, using Fisher's z for

Table V. Factor Analyses (Experiment 2)

	Single Items							
	Direct *Interaction Mechanism*				Indirect *Interaction Mechanism*			
Item	Factor 1: Disorientation	Factor 2: Usefulness	Factor 3: Flow	Factor 4: Ease of use	Factor 1: Disorientation	Factor 2: Usefulness	Factor 3: Flow	Factor 4: Ease of use
FLO1	0.185	0.090	**0.788**	−0.224	0.054	0.003	**0.739**	0.008
FLO2	0.093	0.052	**0.847**	−0.160	0.091	−0.122	**0.918**	−0.130
FLO3	−0.140	−0.075	**0.779**	0.105	−0.017	0.014	**0.858**	0.047
FLO4	−0.072	−0.021	**0.837**	0.219	−0.088	0.122	**0.829**	0.079
EOU1	0.234	−0.005	−0.027	**0.633**	0.051	0.035	0.072	**0.881**
EOU2	0.260	0.183	−0.006	**0.608**	0.017	0.056	0.052	**0.898**
EOU3	0.167	0.033	0.049	**0.706**	0.260	0.165	−0.064	**0.643**
DIS1	**−0.393**	−0.182	−0.104	−0.350	**−0.619**	−0.098	−0.086	−0.194
DIS2	**−0.441**	−0.172	−0.161	−0.197	**−0.773**	−0.009	−0.013	−0.008
DIS3	**−0.779**	0.188	−0.057	−0.027	**−0.498**	0.166	−0.066	−0.454
DIS4	**−0.874**	0.071	0.019	−0.060	**−0.667**	0.009	0.062	−0.268
DIS5	**−0.758**	0.035	0.146	−0.189	**−0.560**	0.136	−0.079	−0.341
DIS6	**−0.628**	−0.156	0.038	−0.174	**−0.942**	−0.021	−0.049	0.037
DIS7	**−0.842**	−0.221	−0.049	0.039	**−0.827**	−0.143	−0.035	0.101
USF1	−0.106	**0.899**	0.004	0.163	−0.018	**0.950**	0.090	0.002
USF2	−0.112	**0.988**	−0.030	0.108	0.051	**0.954**	−0.007	−0.024
USF3	0.046	**0.987**	−0.048	−0.044	0.161	**0.930**	−0.081	−0.092
USF4	0.102	**0.826**	0.078	−0.158	−0.071	**0.760**	0.023	0.170
	Whole form							
FLO1	−0.085	0.057	**0.632**	0.041	0.034	−0.124	**0.660**	0.097
FLO2	−0.096	−0.013	**0.522**	−0.019	−0.047	0.007	**0.844**	−0.074
FLO3	0.193	−0.088	**0.610**	−0.122	−0.109	−0.039	**0.664**	0.372
FLO4	0.209	−0.032	**0.634**	0.129	0.224	0.167	**0.687**	−0.179
EOU1	0.084	0.068	0.028	**0.541**	−0.079	0.318	0.092	**0.753**
EOU2	−0.046	0.088	−0.105	**0.891**	0.002	0.070	−0.012	**0.731**
EOU3	−0.129	0.042	0.217	**0.916**	0.139	0.109	0.006	**0.591**
DIS1	**−0.733**	−0.049	−0.087	0.016	**−0.586**	−0.142	−0.049	0.078
DIS2	**−0.438**	−0.209	−0.284	−0.082	**−0.770**	−0.031	0.008	−0.068
DIS3	**−0.153**	0.165	0.077	[a] **−0.481**	**−0.176**	0.041	−0.043	[a] **−0.665**
DIS4	**−0.273**	−0.030	0.153	[a] **−0.566**	**−0.319**	0.069	−0.026	[a] **−0.612**
DIS5	**−0.762**	−0.049	0.040	0.030	**−0.775**	0.066	0.198	−0.205
DIS6	**−0.797**	−0.037	−0.044	−0.142	**−0.560**	0.052	−0.098	−0.279
DIS7	**−0.452**	−0.029	0.062	−0.171	**−0.823**	0.033	−0.157	0.013
USF1	0.147	**0.773**	−0.082	0.026	−0.021	**0.863**	−0.046	0.155
USF2	−0.047	**0.980**	0.030	−0.015	−0.033	**0.936**	−0.040	−0.007
USF3	−0.012	**0.962**	0.050	0.042	0.016	**0.886**	0.077	−0.044
USF4	0.045	**0.898**	−0.072	−0.029	0.086	**0.706**	−0.035	0.090

Notes. N = 127. Extraction method: principle axis factoring; rotation method: direct oblimin. FLO: flow; EOU: perceived ease of use; DIS: disorientation; USF: perceived usefulness. Items are numbered per scale, for example EOU1: perceived ease of use, Item 1. Figures are factor loadings.
[a] Severe cross-loading.

Table VI. Reliability of Scale (Experiment 2)

	Direct (radio buttons)		Indirect (drop-down list)	
Scale	Whole Form	Single Items	Whole Form	Single Items
Intensity of flow	0.70	0.88	0.81	0.90
Disorientation	0.84	0.92	0.86	0.94
Perceived ease of use	0.84	0.88	0.83	0.94
Perceived usefulness	0.95	0.96	0.92	0.95

Note. Reliability coefficient: Cronbach's alpha.

Table VII. Correlations between Scales (Experiment 2)

		Intensity of Flow	Disorientation	Perceived Ease of Use	Perceived Usefulness
Direct *interaction mechanism*	*Intensity of flow*	**0.854	−0.216	0.052	0.023
	Disorientation	−0.174	**0.824	**−0.491	**−0.380
	Perceived ease of use	0.069	**−0.719	**0.777	*0.252
	Perceived usefulness	0.169	**−0.331	*0.315	**0.850
Indirect *interaction mechanism*	*Intensity of flow*	**0.884	**−0.289	*0.261	0.093
	Disorientation	**−0.370	**0.852	**−0.492	*−0.250
	Perceived ease of use	*0.299	**−0.777	**0.826	**0.414
	Perceived usefulness	0.115	*−0.300	**0.330	**0.752

Note. Upper right: correlations between items with whole-form presentation; lower left: correlations between items with single-items presentation; diagonal: correlations of whole-form- with single-items-presentation for the same items.

*Correlation significant at the 0.05 level (two-tailed).

**Correlation significant at the 0.01 level (two-tailed).

nonindependent groups, revealed that, for both *interaction mechanism*s, the size of the correlation between *Disorientation* and *Perceived ease of use* was greater with single items than with whole form, $z = 2.68$, $p < 0.01$ for direct interaction and $z = 3.82$, $p < 0.001$ for indirect interaction.

Concerning criterion validity, the correlation of *Disorientation* with both percentage of correct answers and average time-on-task before a correct answer was significant. For percentage of correct answers, Pearson's $r = -0.27$, $p < 0.05$ for direct/whole form, $r = -0.34$, $p < 0.01$ for direct/single items, $r = -0.37$, $p < 0.005$ for indirect/whole form and $r = -0.39$, $p < 0.005$ for indirect/single items. For average time-on-task before a correct answer, $r = 0.29$, $p < 0.05$ for direct/whole form and $r = 0.27$, $p < 0.05$ for direct/single items. *Disorientation* was also significantly correlated with number of visits to the homepage, Pearson's $r = 0.26$, $p < 0.05$ for direct/whole form, $r = 0.35$, $p < 0.005$ for direct/single items, $r = 0.32$, $p < 0.05$ for indirect/whole form and $r = 0.31$, $p < 0.05$ for indirect/single items. *Perceived ease of use* was significantly correlated with percentage of correct answers, Pearson's $r = 0.40$, $p < 0.005$ for direct/whole form, $r = 0.37$, $p < 0.005$ for direct/single items, $r = 0.29$, $p < 0.05$ for indirect/whole form and $r = 0.29$, $p < 0.05$ for indirect/single items. *Intensity of flow* was significantly correlated with percentage of correct answers, $r = 0.25$, $p < 0.05$ for direct/whole form and $r = 0.29$, $p < 0.05$, for direct/single items. Tests for differences in correlations between subjective measures and a common objective measure between the two *questionnaire layouts*, and tests for differences in correlations between subjective measures and a common objective measure between the two *interaction mechanism*s, showed that none of the correlations differed significantly.

5.2.4 *Sensitivity*. Using flow, perceived ease of use, perceived usefulness and disorientation as dependent variables, 2×2×2 ANOVAs were conducted with independent variables *design principle*s, *interaction mechanism* and *questionnaire layout*, analyzing responses for first *response format* used.

For *Disorientation*, the main effect of *design principle*s was significant, $F(1,119) = 15.97$, $p < 0.0001$, $\eta^2 = 0.11$, as well as the three-way interaction of

Table VIII. The Effects of Questionnaire Layout and Design Principles on Disorientation (Experiment 2)

		Questionnaire Layout		
*Design Principle*s	*Interaction Mechanism*	Single Items M (SD)	Whole Form M (SD)	Overall M (SD)
Applied	Direct	2.72 (1.17)	3.48 (1.20)	3.10 (1.23)
	Indirect	2.64 (1.34)	2.69 (0.92)	2.67 (1.12)
	Overall	2.68 (1.23)	3.10 (1.13)	2.90 (1.19)
Violated	Direct	3.83 (1.40)	3.58 (1.00)	3.71 (1.21)
	Indirect	3.29 (1.28)	4.22 (1.13)	3.78 (1.27)
	Overall	3.55 (1.34)	3.95 (1.11)	3.75 (1.23)
Overall	Direct	3.21 (1.37)	3.52 (1.11)	3.36 (1.25)
	Indirect	2.96 (1.33)	3.46 (1.28)	3.23 (1.32)
	Overall	3.09 (1.35)	3.49 (1.19)	3.29 (1.28)

*design principle*s by *interaction mechanism* by *page layout*, F (1, 119) = 4.91, $p < 0.05$, $\eta^2 = 0.03$ (see Table VIII). None of the other main or interaction effects was significant. A simple effect test showed that, with single-items presentation, the effect of *design principle*s was significant, F (1, 58) = 7.05, $p = 0.01$, $\eta^2 = 0.11$, though neither the effect of page layout nor the interaction effect were significant, both $F < 1$. However, a further test demonstrated that, with whole-form presentation, the effect of *design principle*s was significant, F (1, 61) = 9.26, $p < 0.005$, $\eta^2 = 0.12$, as well as the interaction effect of *design principle*s with *interaction mechanism*, F (1, 64) = 6.70, $p = 0.01$, $\eta^2 = 0.09$, but not the main effect of *interaction mechanism*, $F < 1$. Further simple effect tests showed that the effect of *design principle*s was significant when using whole-form presentation with indirect interaction, t (32) = 4.33, $p < 0.01$, d = 0.77, but not with direct interaction, $t < 1$.

For *Perceived ease of use*, the main effect of *design principle*s was significant, F (1, 119) = 14.81, $p < 0.001$, $\eta^2 = 0.11$, with mean = 4.91 (SD = 1.16) for *design principle*s applied and mean = 4.01 (SD = 1.41) for *design principle*s violated. None of the other main- or interaction effects was significant. For *Perceived usefulness*, the main effect of *design principle*s was significant, F (1, 119) = 4.58, $p < 0.05$, $\eta^2 = 0.04$, with mean = 4.25 (SD = 1.41) for *design principle*s applied and mean = 3.67 (SD = 1.57) for *design principle*s violated. None of the other main- or interaction-effects was significant. For *Intensity of flow* (involvement), none of the main- or interaction effects were significant.

5.3 Speed of and Changes in Completing Questionnaire Items

A 2×2×(2) analysis of variance showed that the effects of *questionnaire layout*, F (1, 123) = 31.80, $p < 0.001$, $\eta^2 = 0.07$, and *interaction mechanism*, F (1, 123) = 8.37, $p < 0.005$, $\eta^2 = 0.04$, on time to complete questionnaire items (see Table IX) were significant. Single items was faster than whole form and direct *interaction mechanism* was faster than indirect. The effect of *design principle*s was not significant, F (1, 123) = 2.13, $p > 0.05$. None of the interaction effects were significant.

With whole-form presentation, mean number of changes was 1.05 (SD = 2.63) with direct interaction, 0.28 (SD = 1.09) with indirect interaction and

Table IX. Effects of Interaction Mechanism and Questionnaire Layout on Time to Complete Questionnaire Items (Experiment 2)

	Questionnaire Layout		
	Whole Form	Single Items	Overall
Interaction Mechanism	M (SD)	M (SD)	M (SD)
Direct	122.50 (45.47)	96.89 (59.14)	109.70 (42.17)
Indirect	140.20 (41.71)	115.08 (31.71)	127.64 (32.89)
Overall	131.42 (44.34)	106.06 (48.03)	118.74 (38.69)

0.66 (SD = 2.04) averaged over *interaction mechanism*s. A 2×2×(2) analysis of variance showed that the effects of *questionnaire layout*, F (1, 123) = 12.83, $p < 0.001$, $\eta^2 = 0.05$, and *interaction mechanism*, F (1, 123) = 8.22, $p < 0.05$, $\eta^2 = 0.02$, on total number of changes in completing questionnaire items were significant, but not the effect of *design principle*s, $F < 1$. None of the interaction effects were significant.

5.4 Preference of Questionnaire Layout

A chi square test showed that a statistically significant majority (63%) of those who had used indirect interaction preferred whole form, χ^2 (1) = 4.00, $p < 0.05$; however, among those who had used direct interaction, there was no statistically significant majority (49%) for whole form interaction, χ^2 (1) < 1.

Based on participants' responses to the open-ended questions presented at the end of the experiment, a number of advantages and disadvantages were identified for both *interaction mechanism*s (see Table X). Regarding changes made to responses when completing the questionnaire, 47% of variance in time to complete questionnaire items was explained by the number of changes made for the direct *interaction mechanism*, $r = 0.69$, $p < 0.001$, and 15% of variance explained for indirect, $r = 0.39$, $p < 0.005$. The difference between these orientations was significant, $z = 2.40$, $p < 0.05$. Part of the remaining variance with both *interaction mechanism*s may have been due to the disadvantage of distraction by other items that participants mentioned and, as a consequence, forgetting to answer one or more questions, which would have to be rectified at the end (on submission of an incomplete set of responses an error message was displayed and submission was only accepted after all items had been completed). Despite participants' feelings that presentation of single items would be more time-consuming, and whole form would be quicker and easier to read, completion was actually quicker with single items. This finding may be explained by the following advantages of single items that respondents stated: focus and lack of influence by responses to other items (the most frequent answer, given by 69% of participants), ease of reading, increased clarity and greater interest, and by the following perceived disadvantages of whole form: distraction by other items, modification of responses to fit with earlier ones and usability issues. Participants reported usability problems as a disadvantage for presentations with single items even though this type of presentation was faster and required no explicit action to submit a response (this happened automatically on answering each question, whereas with whole form a submit action was required after completing all items); on the other hand, single items did not allow

Table X. Advantages and Disadvantages of Questionnaire Layouts (Experiment 2)

Advantages			
Whole Form		Single Items	
Knowing how many questions left/size of task	[a]41	Focus/Uninfluenced by other responses	87
Ability to go compare/change responses	32	No advantages/Don't know	20
Speed	21	Miscellaneous	6
Ease of reading	17	Ease of reading	5
No advantages/Don't know	13	Increased clarity	5
Miscellaneous	4	Greater interest	3
Disadvantages			
Whole Form		Single Items	
Too much information/distraction	54	Uncertainty about length of task	28
No disadvantages/Don't know	19	Time taken	28
Lack of care in responses	18	No disadvantages/Don't know	26
Modification of responses to fit with earlier ones	12	Dullness/repetitiveness of task	12
Usability issues	9	Inability to change responses	11
Boredom	9	Couldn't check previous answers	9
Miscellaneous	6	Usability issues	6
		Miscellaneous	6

[a] Frequency.

recovery after an incorrect selection. The perceived disadvantage of a lack of care in responses may have been caused by distraction and may have encouraged respondents to make changes, with consequently longer completion times; alternatively, longer times-to-complete can be explained (at least partially) by participants making changes (see evidence presented above) and may also have been caused by participants taking some time to read over the questions before beginning to respond.

5.5 Summary of Results

Experiment 2 established the psychometric properties of scales for the measuring four key concepts (disorientation, perceived ease of use, flow and perceived usefulness) in the quality of users' interaction with web sites simultaneously. Overall, faster completion of questionnaire items was found using a *questionnaire layout* with direct interaction and single items and severe cross-loadings occurred with whole-form presentation. Participants felt focused and uninfluenced by other responses when using single items, whereas with whole form participants felt distracted. However, validity and sensitivity were generally similar between layouts.

6. DISCUSSION

6.1 Psychometric Properties

6.1.1 *Factor Analysis.* Four distinct factors were found—for disorientation, perceived ease of use, intensity of flow and perceived usefulness in Experiment 2 and for the first three factors in Experiment 1 when items for the last factor were not included. In Experiment 2, whole-form presentation produced a lack of simple structure with severe cross-loadings.

6.1.2 *Reliability.* The four corresponding scales, *Disorientation*, *Perceived ease of use*, *Intensity of flow* and *Perceived usefulness*, were found to be reliable in Experiment 2 and the first three in Experiment 1. In Experiment 2, reliability varied for layout with single items and whole form, but reliabilities of indirect interaction and direct interaction were similar. Furthermore, reliability values of scales with Likert format with direct interaction were very similar in both experiments.

6.1.3 *Validity.* Discriminant validity among the four scales was generally confirmed through moderate, but significant, correlations among the scales. Exceptions were a non-significant correlation in Experiment 1 between *Intensity of flow* and *Perceived ease of use* with visual analogue scale format and non-significant correlations in Experiment 2 of *Intensity of flow* with *Perceived usefulness,* and also with *Disorientation* and *Perceived ease of use* when a direct *interaction mechanism* was used. Validity of the scales was further confirmed by significant correlations between the two *response formats*, but to a lesser extent for *Perceived ease of use* than for the other scales in Experiment 1, and between the two *questionnaire layout* in Experiment 2.

Both experiments found evidence for criterion validity of *Disorientation* through significant correlations with accuracy and speed of task performance and a positive correlation with two behavioral measures (visits to the home page and, in Experiment 1, number of links before a correct answer). However, the correlation with accuracy was significantly larger with the visual analogue scale than the Likert *response format* in Experiment 1. Experiment 2 found some evidence of criterion validity for *Perceived ease of use* and, with direct *interaction mechanism*, and *Intensity of flow*, through significant correlations with accuracy of task performance.

6.1.4 *Sensitivity.* Evidence for sensitivity was found to varying degrees for the different scales. In Experiment 1, the effect size of *design principle*s was $\eta^2 = 0.21$ for *Disorientation*, 0.03 for *Intensity of flow,* and 0.03 for *Perceived ease of use*, but only the first of these three effect sizes was statistically significant. In terms of Cohen's [1988] conventions, the last two effect sizes are between a small ($\eta^2 = 0.01$) and a medium effect size ($\eta^2 = 0.059$) . However, the first effect size for *Disorientation* is considerably greater than a large effect size ($\eta^2 = 0.138$) and represents more than 20% of variability in *Disorientation* scores accounted for by *design principle*s. In Experiment 2, the effect size of *design principle*s was 0.11 for *Disorientation*, 0.11 for *Perceived ease of use*, 0.04 for *Perceived usefulness* and < 0.001 for *Intensity of flow*, with the first three statistically significant. Although some statistically significant main and interaction effects of *questionnaire layout* and *interaction mechanism* on *Disorientation* were found, their effect sizes were small, showing that scales were predominantly sensitive to *design principle*s.

Comparing statistically significant effect sizes of psychometric measures with those of objective measures, the effect of *design principle*s was extremely large on visits to the homepage in Experiment 1, followed by *Disorientation* in Experiment 1 (very large), *Disorientation* and visits to the homepage as well

as time-on-task before a correct answer in Experiment 2 (large), *Perceived ease of use* in Experiment 2 (large to medium), percentage of correct answers in Experiments 1 and 2 as well as links visited before correct and incorrect answers in Experiment 2 (medium), and *Perceived usefulness* in Experiment 2 (medium to small). These results indicate that the performance of the psychometric measures *Disorientation* and *Perceived ease of use* was similar to that of task performance- and behavioral measures in terms of sensitivity. Further similarity between *Disorientation* and *Perceived ease of use* was indicated by the findings of severe cross-loadings in the factor analyses in Experiment 2 with whole form and high correlations between the two scales.

In terms of psychometric properties, our results confirm findings of Ahuja and Webster [2001], Davis [1989] and van Schaik and Ling [2003a]. Ahuja and Webster established that *Disorientation* and *Perceived ease of use* were two separate factors and in Davis's research *Perceived ease of use* and *Perceived usefulness* were separate factors. Van Schaik and Ling [2003a] found the same factor structure for the three scales *Disorientation*, *Perceived ease of use* and *Intensity of flow*. Consistent with our results, the previous studies also found the four scales to possess reliability and discriminant validity. Our results confirmed criterion validity of *Disorientation* through correlations with accuracy as in both previous studies, but with considerably stronger evidence for validity through a higher correlation with visits to the home page than in van Schaik and Ling and also through correlations with speed and efficiency (links visited before a correct answer). Although the sensitivity of *Disorientation* to web designs was confirmed, in terms of explained variability the effect size for *Disorientation* in Experiment 1 was a factor twice as large as the effect sizes found by the previous studies ($r^2 = 0.09$ in Ahuja and Webster and $\eta^2 = 0.099$ in van Schaik and Ling). In contrast to the results of Ahuja and Webster, *Perceived ease of use* showed sensitivity in Experiment 2 (but not in Experiment 1). Although Davis and Wiedenbeck [2001] found *Intensity of flow* sensitive to interaction style, this scale did not respond to *design principle*s in the current study, showing a specificity of response. Overall, the effect of *design principle*s was strongest on *Disorientation*. Furthermore, the *design principle*s manipulated in our research were different from the manipulations included in the previous studies (navigation systems in Ahuja and Webster and orientation support in van Schaik and Ling), indicating that *Disorientation* is sensitive to a range of web page design parameters.

The finding that Davis and Wiedenbeck's [2001] control items did not emerge as separate factors in our factor analyses may be due to the nature of users' interaction with web sites. This interaction was frequently characterized by navigation through a set of pages in order to find information, where control is predominantly not a concern, rather than by the application of various "functions" to objects that are manipulated as in other software programs such as the word processor application studied by Davis and Wiedenbeck. Therefore, the extent to which the construct of control applies may vary between applications and may not apply particularly to web sites.

The findings of the current study should give more confidence regarding the quality of the four scales because a more realistic "live" web site was used

in both experiments (van Schaik and Ling used a smaller site, constructed specially for their research) and also because the four scales were analyzed together [Ahuja and Webster 2001; Davis 1989; Davis and Wiedenbeck 2001; van Schaik and Ling 2003a] all used subsets of the four scales, but did not use all four simultaneously).

6.2 Response Format

Although a large majority of respondents preferred Likert as the *response format* in Experiment 1, the two *response formats* predominantly displayed the same psychometric properties for each of the scales and there was no response bias associated with the visual analogue *response format*, which did not produce more extreme scores than the Likert format. Consistent with the results of previous research (e.g., Brunier and Graydon [1996] and Murphy et al. [1988]), we generally found significant positive correlations between a 7-point Likert and 10-cm visual analogue scale. An exception was the lower and nonsignificant correlation between *Perceived ease of use* and *Intensity of flow* for the visual analogue format in comparison with the Likert format.

In terms of sensitivity, in both *response formats* the *Disorientation* scale demonstrated sensitivity, similar to findings by Price et al. [1994] and Hayes et al. [1996]. The *Disorientation* scale was also equally sensitive to web designs, using both a 10-cm 7-point Likert format and a 10-cm visual analogue scale format, in van Schaik and Ling [2003a].

Contrary to Pfennings et al. [1995], we did not find a higher variability for visual analogue scale than for Likert, with variance ratios of visual analogue scale to Likert in the order of 0.8. This indicates, if anything, a *lower* variability for visual analogue scale.

The equivalence of the two *response formats* results can be explained in terms of the graduation of the Likert format and continually diminishing returns with increasing scale steps [Nunnally and Bernstein 1994]. A 7-point scale may have sufficient categories to produce this equivalence. However, Likert scales with fewer response alternatives or different scale lengths may produce nonequivalent results [Bellamy et al. 1999a, 1999b; Joyce et al. 1975].

The current study did not investigate the equivalence of online and paper-based administration of questionnaires; however, the similarities in psychometric properties between the current study, using online presentation, and previous research that used the same scales presented on paper [Ahuja and Webster 2001; Davis and Wiedenbeck 2001] are encouraging. This finding is consistent with results from previous research comparing paper-based and online formats [Harper et al. 1997; Slaughter et al. 1994] and comparing paper-based, online and telephone questionnaire administration [Knapp and Kirk 2003]. An equivalence of online formats with other types of format is crucial because psychometrics instruments can be usefully employed in various situations where Web access is available and the only or most convenient medium to collect data.

6.3 Interaction Mechanism and Layout

With indirect *interaction mechanism*, there was a preference for presentation of whole form, although responses were slower using this layout. Moreover,

contrary to participants' perceptions and perhaps web designers' intuition, whole-form presentation was slower than single item, with 15% (with indirect *interaction mechanism*) and 47% (direct *interaction mechanism*) of variance explained by changes made to previous answers. In particular, with direct *interaction mechanism*, more changes were made and time-to-complete was strongly associated with the number of changes made, perhaps because of higher perceived ease of making changes. The lack of distraction by responses to and the text of other items may have contributed to the advantage of single items. Indeed, the finding that despite perceived disadvantages such as the inability to change responses or check previous answers, results were better for single items is consistent with the idea that this layout gave better support for the requirements of spontaneous responses based on a first reading of each item and without long deliberation or consideration of responses to previous items.

An implication of these results is that items should be presented in online questionnaires singly rather than *en masse* using whole form. The flexibility of online questionnaires makes this a feasible option that would be prohibitive in paper-based questionnaires, due to the enormous amount of paper required with this type of presentation. Other *questionnaire layouts*, such as semantic partitions or screen-sized pages that were used by Norman et al. [2001] were not investigated in the current study, and consequently their merits in terms of psychometric properties remain unknown. Nonetheless, our results confirm that a single-items layout is more suitable for online collection of this type of data, resulting in faster completion of items and higher psychometric quality in terms of avoiding severe cross-loadings with the scales (*Disorientation*, *Perceived ease of use*, *Perceived usefulness* and *Intensity of flow*) used in this study. Indeed, this format encourages compliance with the instructions of giving a spontaneous response and makes it easier to follow them. Tourangeau et al.'s [2004, Experiment 6] finding that single items presentation leads to lower intercorrelations (between items measuring factual information rather than psychometric items) also adds substance to the argument that this form of presentation encourages participants to answer questions in isolation, rather than referring back to previous responses. Hence, the use of single-items presentation should not only improve the speed of completion, but also increase the validity of responses. One criticism participants made was that with the single-items presentation they often felt that they did not know how long the task would take to complete or how many more questions they had remaining. Such a problem could be reduced by providing participants with a visual progress indicator.

Regarding *interaction mechanism*, the advantage of a lack of distraction by the response categories of other items on the same page would, by definition, only occur when presenting multiple items on a page. However, in terms of psychometric quality, no difference in factor structure was found between direct and indirect *interaction mechanism*s in either *questionnaire layout*, reliabilities were similar between *interaction mechanism*s in either layout and no difference in the patterns of results for validity and sensitivity were found between orientations. Overall then, in the current study *interaction mechanism* had little effect on psychometric results. Furthermore, questionnaire items took longer

to complete with indirect interaction (drop-down list) than with direct interaction (radio buttons). The main cause of this difference in speed may be mainly due to the fact that, as discussed previously, more actions are required with a drop-down list. An implication of these findings is that design decisions about *interaction mechanism* with discrete response categories can be taken based on practical or design considerations rather than psychometric ones, including time-to-complete and screen design factors such as the amount of space required by different orientations (vertical or horizontal presentation of response alternatives).

6.4 Future Research

The current study found evidence for the sensitivity of *Disorientation* and *Perceived ease of use*, with responses affected by the experimental manipulation of the application of *design principle*s, but less sensitivity of *Perceived usefulness* and a lack of sensitivity of and *Intensity of flow*. Research is required to determine the design parameters (a) that individual components of the quality of interaction with web sites are uniquely sensitive to—such as disorientation, perceived ease of use, flow, aesthetic quality (e.g., Hassenzahl [2004] and Lindgaard and Dudek [2003]) and trust [Safar and Turner 2005]—and (b) to which several components are sensitive. The aim of this exercise is to establish a mapping between design parameters and quality components as measured using psychometric instruments. This research may question and refine the definition of these components and the conceptualization of quality of interaction, in order to make testable predictions about the effect of design parameters on outcomes representing quality. Important parameters are likely to include information scent [Blackmon et al. 2005; Card et al. 2001; Larson and Czerwinski, 1998] and information architecture more generally [Rosenfeld and Morville 2002]. Knowledge gained from this research regarding the differential sensitivity of components to design parameters should aid in improving the design of web sites.

7. CONCLUSION

In conclusion, despite conflicting results in other applications of different *response formats*, the *response formats* Likert and visual analogue scale resulted in essentially the same psychometric properties of scales in our evaluation of intranet pages; presentation of single items and a direct *interaction mechanism* produced faster completion of questionnaires and are therefore recommended. The *Disorientation* scale was found to have good psychometric properties in terms of factor structure, reliability, validity and sensitivity consistently in both experiments. Over the two experiments, the *Perceived ease of use* scale was also found to have good psychometric properties. Although the scale corresponded with a separate factor resulting from factor analysis and was reliable, in Experiment 2, it was strongly correlated with *Disorientation* and therefore may—at least when used in the same way as in the current study—not provide additional information on the quality of human-computer interaction. These results should give practitioners and researchers increased confidence in the online administration of both scales, in particular the *Disorientation* scale.

The *Perceived usefulness* scale—only used in Experiment 2—was good in terms of factor structure and reliability and to some extent validity and sensitivity. Although a lack of evidence for sensitivity of the *Intensity of flow* scales in the use of web sites may currently restrict its usefulness, further investigation is required to establish the conditions under which sensitivity can be demonstrated; however, the scale was found to be a distinct element in the factor structure, and was reliable and valid.

In order to satisfy the need to monitor and continually improve web site usability [Nielsen 2003], psychometric instruments are becoming increasingly important as tools to measure the quality of interaction with web sites, alongside measures of task performance and navigation behavior for monitoring and regularly improving these sites. The results from the current study should facilitate this development.

In the meantime, the results from the current study indicate that 7-point Likert and visual analogue scale, when presented one item per page, are overall equally good for the design of psychometric online questionnaires. When Likert is used, items should be presented using direct interaction—with radio buttons—and one item per page rather than using indirect interaction—with drop-down boxes—and all items on a single page. Of the four scales, *Disorientation* has the best quality of measurement.

APPENDIX

A: QUESTIONNAIRE ITEMS

Perceived ease of use

- Learning to use this site was easy
- Becoming skilful at using the site was easy
- The site was easy to navigate

Perceived usefulness

- Using the site would improve my performance in my coursework
- Using the site in my coursework would increase my productivity
- Using the site would enhance my effectiveness in my coursework
- I would find the site useful in my coursework

Disorientation

- I felt lost
- I felt like I was going around in circles
- It was difficult to find a page that I had previously viewed
- Navigating between pages was a problem
- I didn't know how to get to my desired location
- I felt disoriented
- After browsing for a while I had no idea where to go next

Intensity of flow – involvement

- I thought about other things
- I had to make an effort to keep my mind on the activity
- I was aware of distractions
- I was aware of other problems

Intensity of flow – control

- Time seemed to pass more quickly
- I knew the right things to do
- I felt like I received a lot of direct feedback
- I felt in control of myself
- I felt in harmony with the environment

Note. Experiment 1 used a (7-point) Likert scale (see Figure 1(a)) and a visual analogue scale (see Figure 1(b)) as *response formats*. Experiment 2 used (7-point) Likert scale as *response format* (see Figure 2).

B: TASKS (INFORMATION RETRIEVAL QUESTIONS)

1 What is the hand-in date of the assessment for the module Research methods 2?
2 Where is the Drop In Study Skills Centre (DISSC) located?
3 What is the telephone number of Dr. Dave Woodhouse?
4 What is PsycInfo?
5 Who should you contact to report any problems with the Psychology intranet site?
6 Which author's name is presented as an example in the library catalogue?
7 Name one of the two organisations that offer ethical guidelines.
8 What is the name of the student employment service?
9 What is the email address to find information about using the library web site?
10 Who deals with requests for coursework extensions?

REFERENCES

AHUJA, J. AND WEBSTER, J. 2001. Perceived disorientation: An examination of a new measure to assess web design effectiveness. *Interact. Comput.* 14, 15–29.

BAGOZZI, R. P., DAVIS, F. D., AND WARSHAW, P. R. 1992. Development and test of a theory of technological learning and usage. *Human Relat. 45*, 659–686.

BELLAMY, N., CAMPBELL, J., AND SYROTUIK, J. 1999a. Comparative study of self-rating pain scales in rheumatoid arthritis patients. (Retrieved November 12, 2000, from the World Wide Web: http://www.librapharm.co.uk/cmro/vol_15/2001/main.htm.)

BELLAMY, N., CAMPBELL, J., AND SYROTUIK, J. 1999b. Comparative study of self-rating pain scales in osteoarthritis patients. (Retrieved November 12, 2000, from the World Wide Web: http://www.librapharm.co.uk/cmro/vol_15/2000/main.htm.)

BLACKMON, M., KITAJIMA, M., AND POLSON, P. 2005. Tool for accurately predicting website navigation problems, nonproblems, problem severity, and effectiveness of repairs. In *Proceedings*

of the SIGCHI Conference on Human Factors in Computing Systems. ACM, New York, 31–40.

BRUNIER, G., AND GRAYDON, J. 1996. A comparison of two methods of measuring fatigue in patients on chronic haemodialysis: Visual analogue versus Likert scale. *Int. J. Nurs. Stud. 33*, 338–348.

BUCHANAN, T. 2000. Internet research: Self-monitoring and judgments of attractiveness. *Behav. Res. Meth., Instrum. Comput. 32*, 521–527.

CARD, S., PIROLLI, P., VAN DER WEGE, M., MORRISON, J., REEDER, R., SCHRAEDLEY, P., AND BOSHART, J. 2001. Information scent as a driver of web behavior graphs: Results of a protocol analysis method for web usability. In *Proceedings of CHI 2001* (Seattle, WA), ACM, New York. 498–505.

CHEN, H., WIGAND, R., AND NILAN, M. 1999. Optimal experience of Web activities. *Comput. Human Behav. 15*, 585–608.

COCKBURN, A., AND MCKENZIE, B. 2001. What do web users do? An empirical analysis of web use. *Int. J. Human-Comput. Stud. 54*, 903–922.

COHEN, J. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.) Erlbaum, Hillsdale, NJ.

COUPER, M. P., TOURANGEAU, R., CONRAD, F. G., AND CRAWFORD, S. D. 2004. What they see is what we get: Response options for web surveys. *Soc. Sci. Comput. Rev. 22*, 111–127.

DALAL, N. P., QUIBLE, Z., AND WYATT, K. 2000. Cognitive design of home pages: An experimental study of comprehension on the World Wide Web. *Inf. Proc. Manage. 36*, 607–621.

DAVIS, F. 1989. Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS Quart.* 13, 318–340.

DAVIS, F., BAGOZZI, R., AND WARSHAW, P. 1989. User acceptance of computer technology: A comparison of two theoretical models. *Manag. Sci. 35*, 982–1003.

DAVIS, S., AND WIEDENBECK, S. 2001. The mediating effects of intrinsic motivation, ease of use and usefulness perceptions on performance in first-time and subsequent computer users. *Interac. Comput.* 13, 549–580.

DILLMAN, D. A., AND CHRISTIAN, L. M. 2005. Survey mode as a source of instability in responses across surveys. *Field Meth. 17*, 30–52.

GILLAN, D., AND COOKE, N. 1995. Methods of cognitive analysis in HCI. In *Proceedings of CHI'95*. ACM, New York, 349–350.

HASSENZAHL, M. 2004. The interplay of beauty, goodness, and usability in interactive products. *Human-Comput. Interact. 19*, 319–349.

HARPER, B., SLAUGHTER, L., AND NORMAN, K. 1997. Questionnaire administration via the WWW: A validation and reliability study for a user satisfaction questionnaire. Paper presented at *WebNet 97*, Association for the Advancement of Computing in Education, Toronto, Ont., Canada. (Retrieved July 20, 2002 from the World Wide Web: http://www.lap.umd.edu/webnet/paper.html.)

HAYES, R., WALKER, S., AND KIRKPATRICK, J. 1996. Topical diclofenac relieves pain from corneal rust ring. *Eye 19*, 443–446.

INTERNET SOFTWARE CONSORTIUM. 2005. *Internet domain survey host count*. (Retrieved February 28, 2005 from the World Wide Web: http://www.isc.org/index.pl?/ops/ds/hosts.php.)

JACCARD, J. 1998. *Interaction effects in factorial analysis of variance*. Sage, London, UK.

JOYCE, C. R. B., ZUTSHI, D. W., HRUBES, V., AND MASON, R. M. 1975. Comparison of fixed interval and visual analogue scales for rating chronic pain. *Europ. J. Clini. Pharm. 8* 415–420.

KLINE, P. 2000. *The Handbook of Psychological Testing* 2nd ed. Routledge, London, UK.

KNAPP, H., AND KIRK, S. 2003. Using pencil and paper, Internet and touch-tone phones for self-administered surveys: Does methodology matter? *Comput. Human Behav. 19*, 117–134.

LARSON, K., AND CZERWINSKI, M. 1998. Web page design: Implications of memory, structure and scent for information retrieval. *Proceedings of CHI 1998* (Los Angeles, CA). ACM, New York, 25–32.

LEWIS, J. 2002. Psychometric evaluation of the PSSUQ using data from five years of usability testing. *Inter. J. Human-Comput. Interact. 14*, 463–388.

LINDGAARD, G., AND DUDEK, C. 2003. What is this evasive beast called user satisfaction? *Interac. Comput. 15*, 429–452.

MONETA, G. AND CSIKSZENTMIHALYI, M. 1996. The effect of perceived challenges and skills on the quality of subjective experience. *J. Personality* 64, 275–310.

MURPHY, D. F., MCDONALD, A., POWER, C., UNWIN, A., AND MACSULLIVAN, R. 1988. Measurement of pain: A comparison of the visual analogue with a non-visual analogue scale. *Clin. J. Pain 3*, 197–199.

NIELSEN, J. 2003. Two sigma: Usability and six sigma quality assurance. [Column posted on the World Wide Web.] Retrieved January 27, 2006 from the World Wide Web: http://www.useit.com/alertbox/20031124.html

NORMAN, K. L., FRIEDMAN, Z, NORMAN, K., AND STEVENSON, R. 2001. Navigational issues in the design of online self-administered questionnaires. *Behav. Info. Tech.* 20, 37–45.

NUNNALLY, J., AND BERNSTEIN, I. 1994. *Psychometric theory*. McGraw-Hill, London, UK.

PFENNINGS, L., COHEN, L., AND VAN DER PLOEG, H. 1995. Preconditions for sensitivity in measuring change: Visual analogue scales compared to rating scales in a Likert format. *Psych. Rep. 77*, 475–480.

PRICE, D., BUSH, F., LONG, S., AND HARKINGS, S. 1994. A comparison of pain measurement characteristics of mechanical visual analogue and simple numerical rating scales. *Pain 6*, 217–226.

ROSENFELD, L., AND MORVILLE, P. 2002. *Information Architecture for the World Wide Web* 2nd ed. O'Reilly, Sebastopol, CA.

SAFAR, J., AND TURNER, C. 2005. Validation of a two-factor structure for system trust. In *Proceedings of the HFES Annual Meeting*. Human Factors Society, Santa Monica, CA.

SAPSFORD, R. 1999. *Survey Research*. Sage, London, UK.

SLAUGHTER, L., HARPER, B., AND NORMAN, K. 1994. Assessing the equivalence of the paper and on-line formats of the QUIS 5.5. In *Proceedings of the Mid-Atlantic Human Factors Conference* (Washington, DC). 87–91.

TOURANGEAU, R., COUPER, M. P., AND CONRAD, F. G. 2004. Spacing position and order: Interpretive statistics for visual features of survey questions. *Public Opinion Quart. 68*, 368–393.

VAN SCHAIK, P., AND LING, J. 2003a. Using online surveys to measure three key constructs of the quality of human-computer interaction in Web sites: Psychometric properties and implications. *Int. J. Human-Comput. Stud. 59*, 545–567.

VAN SCHAIK, P., AND LING, J. 2003b. The effect of link colour on information retrieval in educational intranet use. *Comput. Human Behav. 19*, 553–564.

VOGT, W. 1999. *Dictionary of statistics and methodology*. Sage, Thousand Oaks, CA.

WARD, R., AND MARSDEN, P. 2003. Physiological responses to different Web page designs. *Int. J. Human-Comput. Stud. 59*, 199–212.

WARE, C. 2004. *Information visualization: Perception for design* 2nd ed. Morgan Kaufmann, San Francisco, CA.

Received June 2005; revised May 2006 and July 2006; accepted September 2006 by Marti Hearst