Using on-line surveys to measure three key constructs of the quality of human–computer interaction in web sites: psychometric properties and implications

Paul van Schaik*, Jonathan Ling

Psychology Section, School of Social Sciences and Law, University of Teesside, Middlesbrough TS1 3BA, UK

Received 31 January 2003; accepted 16 March 2003

Abstract

On-line surveys are now an important tool for data collection on the World Wide Web (the Web). Determining the psychometric properties of key constructs such as disorientation, ease of use and flow is of paramount importance in establishing the quality of users’ interactions with web sites. The current study used techniques of experimental research and on-line surveys to investigate the psychometric properties of existing instruments for measuring these constructs using two response formats: visual analogue scale and Likert scale. A 2 × 2 design with response format and orientation support as independent variables was used. Ninety participants carried out an information retrieval task using an experimental on-line shopping site before completing the scales. Factor analysis confirmed the existence of three distinct scales that possessed high reliability. Evidence for validity, and to a lesser extent, sensitivity, was found. Although psychometric results generally converged, some differences between the two response formats were found. A framework for the comprehensive investigation of response formats of on-line questionnaires is proposed as a basis for future research. Practical implications for the on-line measurement of the quality of users’ interactions with web-based systems are discussed.

© 2003 Elsevier Science Ltd. All rights reserved.

Keywords: On-line surveys; Psychometrics; World Wide Web; Intranet; Response format; Visual analogue scale; Likert scale
1. Introduction

The rate of growth in the number of World Wide Web (Web) hosts is exponential (Internet Software Consortium, 2002). Therefore, the need for more usable web-based systems, and web sites in particular, is becoming a necessity as an increasingly large proportion of the population with a decreasingly specialist computer expertise is using the Web (Nielsen, 2001).

One common function of web sites is the collection of survey data (Buchanan, 2000). Such surveys encompass an extremely broad range of topics, from single questions relating to political issues (e.g. ICM, 2002) to more complex, multi-statement surveys of drug use (e.g. www.drugsresearch.org.uk). Responses to survey questions presented on-line may take a number of forms including radio buttons, drop-down menus and sliders. Despite the wide use of such methods to collect data, little research has investigated the impact of these differing response formats on the answers given by respondents.

One of the applications of on-line questionnaires is the measurement of users’ experience of web sites (both Web and intranet sites). Many variables related to this experience, such as disorientation (Ahuja and Webster, 2001), can have an important impact on or be affected by users’ performance with and acceptance of these sites. However, a prerequisite for the use of subjective measures is that they are psychometrically sound. Psychometrics is a branch of psychology focusing on the operationalization of variables for the purposes of measurement (Vogt, 1999). No research has yet simultaneously investigated the quality of a range of key measures of interaction with web pages.

Psychometrics has been used since the early 1900s to ascertain the quality of psychological measures, in particular tests of ability and personality (Kline, 2000). However, since the late 1970s, psychometrics has also been used to develop questionnaires in order to measure the quality of human–computer interaction (Kirakowski, 1994). In the field of Human–Computer Interaction (HCI), four key aspects of the quality of psychometrics instruments are factor structure, reliability, validity and sensitivity (Lewis, 1995). The goal of factor analysis is to determine the underlying structure of a set of questionnaire items by reducing them to a smaller number that can be measured reliably. One major type of reliability is a quantitative appraisal of its consistency. The relationship between items and factors is determined and a set of items that make up a factor is tested for reliability, usually by employing Cronbach’s co-efficient \( \alpha \). If \( \alpha \) is sufficiently high (> 0.70) then the items are summed or averaged to produce a scale, thereby reducing the larger set of items scores to single scale value. A classic example is found in Davis (1989), who developed a 12-item questionnaire comprising two factors as a basis for two reliable scales: perceived ease of use (consisting of six items) and perceived usefulness (consisting of six items). Reliability is a pre-requisite for validity, which is the degree to which an instrument measures what it claims to measure. Two major types are discriminant validity and

\[ ^1 \text{A distinction is made between the process of human-computer interaction (lowercase) and the academic discipline or field of Human-Computer Interaction (uppercase).} \]
criterion-related validity (Bagozzi et al., 1992), both usually established with Pearson’s correlation co-efficient \( r \). Discriminant validity ascertains the level of differentiation between measures of distinct constructs. Criterion validity appraises the relationship between one indicator of a construct (e.g. a workload inventory such as the Task Load Index, NASA-TLX; Hart and Staveland, 1988) 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 different systems, user populations or tasks. Typically, analysis of variance (ANOVA) is used to test sensitivity, for instance with system and user type as independent variables.

The psychometric properties of questionnaires may depend on the response format that is used in scale items. Psychological and health research have used two types of response format: discrete (typically a Likert scale, Likert for short; see Fig. 1a) and analogue (typically a visual analogue scale; see Fig. 1b). In HCI, discrete response formats are normally used rather than analogue formats (Gillan and Cooke, 1995); however, a scientific justification for the choice of response format in HCI is lacking. Discrete response formats such as Likert typically have seven graduated categories to choose from, anchored with descriptive phrases representing the lowest and highest response categories. Respondents select the category most representative of, for example, their perceived quality of human–computer interaction. Continuous response formats such as visual analogue scale are typically presented as a 10cm horizontal line, anchored with two verbal descriptors at the extremes (e.g. strongly agree and strongly disagree) on which respondents indicate their perceived status by placing a mark along the horizontal line at the most appropriate point. Reported advantages of Likert and visual analogue scale (Flynn et al., forthcoming) are presented in Table 1.

A small number of studies has directly compared Likert and visual analogue scale formats, resulting in several findings. Significant positive correlations for ratings of acute pain have been reported for a 10-point Likert and a 10-cm visual analogue scale (Murphy et al., 1988), and ratings of fatigue using a 5-point Likert and a 10-cm visual analogue scale (Brunier and Graydon, 1996). Studies have found that

![Fig. 1. Response formats as used in the current study. (a) Likert response format, (b) visual analogue scale response format.](image-url)
self-ratings of pain intensity are rated higher with a 10-point Likert than with a 10-cm visual analogue scale (Murphy et al., 1988) and an 11-point Likert produced higher ratings of level of pain and pain unpleasantness than a 15-cm visual analogue scale (Price et al., 1994); conversely, visual analogue scale ratings obtained for the perceived blackness of squares were higher than perceptions obtained from graduated Likert (Neely and Borg, 1995). Furthermore, Pfennings et al. (1995) reported significantly more variability in 10cm visual analogue scale responses to a general health questionnaire, compared to measurements using 10-point Likert. Researchers have also reported both ceiling and floor effects of varying length for measurements of subjective phenomena using visual analogue scale (Gift, 1989; Neely et al., 1992; Neely and Borg, 1995), although such effects were not found by Murphy et al. (1988) for either 10-point Likert or 10-cm visual analogue scales for self-ratings of acute pain.

Nyren et al. (1987) found greater reproducibility in 7-point Likert than in 10 cm visual analogue scale for self-ratings of abdominal pain; however, other research found that 10 cm visual analogue scale was more reproducible than 5-point Likert for assessments of breathlessness and general fatigue after exercise (Grant et al., 1999).

Direct comparisons between Likert and visual analogue scale have shown no significant differences in responsiveness for ratings of chronic pain (Guyatt et al., 1987; Nyren et al., 1987; Hayes et al., 1996), ratings of intensity and experience of pain in response to thermal stimuli (Price et al., 1994) or breathlessness in patients surviving chronic heart failure (Jaeschke et al., 1990). However, other studies have demonstrated that 10-cm visual analogue scale was more responsive than 4-point Likert for self-ratings of arthritic and chronic pain (Joyce et al., 1975; Bellamy et al., 1999a, b).

In summary, the research reviewed above indicates that Likert and visual analogue scales may differ across a number of properties, including level of rating,
reproducibility and responsiveness. These criteria are all important for the evaluation of human–computer interaction. However, the majority of direct comparisons have focused on subjective assessments of pain or fatigue. It is not clear from the literature if Likert and visual analogue scale would yield equivalent measurement of the quality of human–computer interaction because no empirical research has compared the properties of different response formats for measuring this quality.

In the context of web-based systems, recent research has studied the psychometric properties of some key constructs previously investigated in mainstream HCI and hypermedia research.

Davis (1989) provided the definition and measurement of two key concepts, perceived ease of use and perceived usefulness, and investigated their role within the framework of Technology Acceptance. Davis et al. (1989, p. 985) defined perceived ease of use as the extent to which an individual believes that using a computer system will be free of effort. Within mainstream HCI, Davis and Wiedenbeck (2001) established that perceived ease of use possessed reliability and criterion validity. In the context of Web use, Ahuja and Webster (2001) confirmed the reliability of a 3-item perceived ease of use scale, but a lack of sensitivity for navigation support in web site designs.

Ahuja and Webster (2001) reviewed existing research on the concept of disorientation, ‘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’ (p. 16). They subsequently developed a new scale for measuring disorientation. Factor analysis showed that disorientation and perceived ease of use were two distinct factors and that the disorientation scale possessed consistency reliability, discriminant validity and sensitivity for navigation support.

Flow is psychological state in which a person feels cognitively efficient, motivated and happy (Moneta and Csikszentmihalyi, 1996, p. 277). In the state of flow, people become absorbed in their activities, and irrelevant thoughts and perceptions are screened out (Chen et al., 1999). Chen et al. identified various factors associated with flow on the Web, suggesting their results could be used to develop a ‘situational measuring instrument’. In the context of mainstream HCI, but relevant to the present research, Davis and Wiedenbeck (2001) developed a new scale for measuring flow. They established that their nine flow items formed a two-factor structure. The flow scale exhibited reliability and sensitivity to both training and interaction style. The scales used in all the HCI and web research described above used 7-point Likert as the response format, except Davis and Wiedenbeck, who measured perceived ease of use with a 5-point Likert format.

Other measures of the quality of human–computer interaction on the Web include task performance and navigation behaviour. These are typically quantitative such as speed (often measured as time-on-task) and navigation behaviour (frequently measured as percentage of correct answers). Usual measures of accuracy include number of pages visited, revisitation rate (which can, incidentally, be considered as behavioural measures of disorientation), pages visited once, and number and percentage of visits to search pages (Cockburn and McKenzie, 2001).
The preceding review of research has indicated first, an absence of evidence for the psychometric properties of psychometric instruments when measuring multiple key concepts in the quality of experience of web sites simultaneously, and second, a lack of empirical evidence for the best response format for measuring quality of HCI. The current study therefore aimed to first, establish psychometric properties of three existing psychometric instruments to measure three key concepts in the quality of experience of web sites (disorientation, perceived ease of use and flow), and second, compare the relative merits of two response formats (7-point Likert and 10-cm visual analogue scale).

2. Method

2.1. Experimental design

This study used a $2 \times 2$ mixed measures design with two factors. The first, independent measures factor, was level of orientation support. This factor had two levels: orientation support (page titles) and no support (no page titles). The second, repeated measures, factor was the response format. This factor also had two levels: Likert (7-point) scale and visual analogue scale. Participants were allocated to either support or no support conditions, and then responding to either Likert or visual analogue scale first, before then responding to the other set of statements (i.e. response format was counterbalanced). Therefore participants were allocated into one of four separate conditions: support and Likert followed by visual analogue scale, support and visual analogue scale followed by Likert, no support and Likert followed by visual analogue scale, and no support and visual analogue scale followed by Likert.

Participants responded to three scales: disorientation and perceived ease of use scales developed by Ahuja and Webster (2001), and the flow scale used by Davis and Wiedenbeck (2001).

2.2. Participants

There were 90 participants, consisting of 74 females and 16 males. They were all undergraduate psychology students and took part in the experiment as a course requirement. Of these, 47 took part in the orientation support conditions (with 24 completing the Likert statements first, and 23 completing the visual analogue scale statements first) and 43 took part in the no-support conditions (with 21 completing Likert statements first, and 22 completing visual analogue scale statements first). Participants had a mean age of 24 years (s.d. = 8.28).

Almost all participants had experience of using the Web (see Table 2) and the vast majority had been using the Web for more than 1 year (91.1%). Frequency of using the Web varied from more than once a day to less than once a month, with a majority (56.7%) using the Web at least once a day.
2.3. Materials

The study was based around a bespoke experimental web site that had the theme of an on-line toy shop (see Fig. 2). This site consisted of a range of pages that contained information about products offered by the shop, a word game, an orders page and details of special offers. Two versions of the web site were produced: one incorporated orientation support in the form of titles for each of the pages (e.g. ‘Order form’), while the other version used the same web pages but removed the page titles.

In addition, participants also completed a series of three on-screen questionnaires. The first scale examined feelings of disorientation when navigating around the site. The scale consisted of seven statements, such as ‘I felt like I was going round in circles’. The second scale focused on perceived ease of use. This instrument comprised three statements such as ‘learning to use the site was easy’. Both the disorientation scale and perceived ease of use scale were taken from Ahuja and Webster (2001). The third questionnaire assessed participants’ perception of their flow through the site. This instrument was comprised of nine statements such as ‘I had to keep my mind on the activity’. This questionnaire was taken from Davis and Wiedenbeck (2001).

Table 2
Experience with the Web and frequency of Web use

<table>
<thead>
<tr>
<th>Length of time</th>
<th>%</th>
<th>Frequency</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 1 Year</td>
<td>91.1</td>
<td>&gt; 1/day</td>
<td>16.7</td>
</tr>
<tr>
<td>&gt; 6 Months</td>
<td>4.4</td>
<td>1/day</td>
<td>40.0</td>
</tr>
<tr>
<td>≤ 1 Month</td>
<td>3.3</td>
<td>&gt; 1/week</td>
<td>35.6</td>
</tr>
<tr>
<td>Not used the Web</td>
<td>1.1</td>
<td>&gt; 1/month</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt; 1/month</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Fig. 2. Typical web pages used in the experiment. (a) Orientation support (page title included), (b) lack of orientation support (page title not included).
Each questionnaire was developed in two response formats: Likert scale (Likert) and visual analogue scale. The Likert response format was presented on a 7-point scale, which included ‘strongly agree’ and ‘strongly disagree’ at either end of the scale (see Fig. 1a). Participants responded to statements presented in this format by clicking on a button beneath one of these points. Participants responded to the visual analogue scale statements by dragging a slider along the scale (see Fig. 1b). There were no subdivisions along the scale, although ‘strongly agree’ and ‘strongly disagree’ were presented at either end of the scale. The slider always started at the midpoint of the scale.

2.4. Procedure

The experiment consisted of an information retrieval task, followed by a questionnaire. Participants were instructed to work independently; any questions raised were answered. In the information retrieval task, participants first completed a practice run consisting of three trials to familiarize them with the site before moving on to the main experiment, which had 15 trials. Participants were told that a question would appear at the top of the screen. After reading the question, they had to click on a button labelled ‘Show web site’. The home page of the web site appeared on the screen and participants were asked to find the answer to the question using the web site. Participants were instructed to take the most direct route possible to locate the answer. Once they found the answer, they clicked on a button labelled ‘Your answer’. A dialog box appeared at the bottom of the screen. Participants typed in their answers in the box and then clicked on ‘OK’ to move on to the next question. At the end of the series of 15 questions related to the web site, participants then completed the questionnaires using Likert and visual analogue scale as response formats. The experiment took approximately 40 min to complete.

3. Results

Psychometric properties and preference for response formats were analysed followed by task performance in the information retrieval task.

3.1. Psychometric properties

According to Lewis (1995), psychometric instruments for the evaluation of human–computer interaction need to have a well-defined factor structure and show reliability, validity and sensitivity. We sought to confirm these characteristics in the three scales used in this study.

3.1.1. Factor structure

The factor structures of (1) the disorientation and the perceived ease of use scales and (2) the flow scale had been investigated separately in previous research using Likert formats (the first two scales by Ahuja and Webster (2001) and the third scale
by Davis and Wiedenbeck (2001)). The current study examined the factor structure of the three scales together for Likert and visual analogue scale formats expecting to find four factors: (1) disorientation, (2) perceived ease of use (both identified by Ahuja and Webster), (3) first flow factor and (4) second flow factor (both identified by Davis and Wiedenbeck).

The correlations between scale items form the basis for factor analysis (Tabachnick and Fidell, 2001). Therefore the correlations between items on each of the three scales were examined prior to factor analysis. Correlations among the seven disorientation items, among the three perceived ease of use items and among the first set of four disorientation items (the second set of four flow items in Davis and Wiedenbeck (2001)) were all >0.3 and statistically significant (see Table 3) for both response formats. However, of the correlations within the second set of five

Table 3
Correlations among questionnaire items: (a) intensity of flow; (b) perceived ease of use; (c) Disorientation

(a) Intensity of flow

<table>
<thead>
<tr>
<th></th>
<th>FLO1V</th>
<th>FLO2V</th>
<th>FLO3V</th>
<th>FLO4V</th>
<th>FLO5V</th>
<th>FLO6V</th>
<th>FLO7V</th>
<th>FLO8V</th>
<th>FLO9V</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLO1L</td>
<td>0.609**</td>
<td>0.717**</td>
<td>0.475**</td>
<td>0.501**</td>
<td>−0.486**</td>
<td>0.009</td>
<td>−0.078</td>
<td>−0.063</td>
<td>0.075</td>
</tr>
<tr>
<td>FLO2L</td>
<td>0.549**</td>
<td>0.679**</td>
<td>0.522**</td>
<td>0.556**</td>
<td>−0.572</td>
<td>−0.008</td>
<td>−0.116</td>
<td>−0.200</td>
<td>−0.072</td>
</tr>
<tr>
<td>FLO3L</td>
<td>0.436**</td>
<td>0.516**</td>
<td>0.778**</td>
<td>0.567**</td>
<td>−0.178</td>
<td>−0.033</td>
<td>−0.086</td>
<td>−0.093</td>
<td>−0.055</td>
</tr>
<tr>
<td>FLO4L</td>
<td>0.457**</td>
<td>0.470**</td>
<td>0.657**</td>
<td>0.693**</td>
<td>−0.222*</td>
<td>−0.234*</td>
<td>−0.007</td>
<td>−0.269*</td>
<td>−0.176</td>
</tr>
<tr>
<td>FLO5L</td>
<td>−0.339**</td>
<td>−0.372**</td>
<td>−0.203</td>
<td>−0.190</td>
<td>0.727**</td>
<td>0.087</td>
<td>0.080</td>
<td>0.238*</td>
<td>0.078</td>
</tr>
<tr>
<td>FLO6L</td>
<td>0.106</td>
<td>0.104</td>
<td>0.008</td>
<td>−0.026</td>
<td>0.229*</td>
<td>0.567**</td>
<td>0.191</td>
<td>0.557**</td>
<td>0.266**</td>
</tr>
<tr>
<td>FLO7L</td>
<td>−0.209*</td>
<td>−0.200</td>
<td>0.047</td>
<td>−0.092</td>
<td>0.123</td>
<td>0.239*</td>
<td>0.737**</td>
<td>0.296*</td>
<td>0.343**</td>
</tr>
<tr>
<td>FLO8L</td>
<td>−0.207*</td>
<td>−0.282**</td>
<td>−0.051</td>
<td>−0.175</td>
<td>0.210*</td>
<td>0.339**</td>
<td>0.391**</td>
<td>0.750**</td>
<td>0.484**</td>
</tr>
<tr>
<td>FLO9L</td>
<td>−0.006</td>
<td>−0.157</td>
<td>−0.020</td>
<td>−0.136</td>
<td>0.204</td>
<td>0.043</td>
<td>0.323**</td>
<td>0.511**</td>
<td>0.844**</td>
</tr>
</tbody>
</table>

(b) Perceived ease of use

<table>
<thead>
<tr>
<th></th>
<th>EOU1V</th>
<th>EOU2V</th>
<th>EOU3V</th>
</tr>
</thead>
<tbody>
<tr>
<td>EOU1L</td>
<td>0.064</td>
<td>0.775**</td>
<td>0.622**</td>
</tr>
<tr>
<td>EOU2L</td>
<td>0.830**</td>
<td>0.079</td>
<td>0.789**</td>
</tr>
<tr>
<td>EOU3L</td>
<td>0.577**</td>
<td>0.720**</td>
<td>0.228*</td>
</tr>
</tbody>
</table>

(c) Disorientation†

<table>
<thead>
<tr>
<th></th>
<th>DIS1V</th>
<th>DIS2V</th>
<th>DIS3V</th>
<th>DIS4V</th>
<th>DIS5V</th>
<th>DIS6V</th>
<th>DIS7V</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIS1L</td>
<td>0.568</td>
<td>0.613</td>
<td>0.451</td>
<td>0.552</td>
<td>0.704</td>
<td>0.759</td>
<td>0.699</td>
</tr>
<tr>
<td>DIS2L</td>
<td>0.395</td>
<td>0.786</td>
<td>0.395</td>
<td>0.438</td>
<td>0.446</td>
<td>0.558</td>
<td>0.401</td>
</tr>
<tr>
<td>DIS3L</td>
<td>0.324</td>
<td>0.339</td>
<td>0.839</td>
<td>0.766</td>
<td>0.555</td>
<td>0.440</td>
<td>0.406</td>
</tr>
<tr>
<td>DIS4L</td>
<td>0.474</td>
<td>0.419</td>
<td>0.828</td>
<td>0.754</td>
<td>0.642</td>
<td>0.517</td>
<td>0.441</td>
</tr>
<tr>
<td>DIS5L</td>
<td>0.664</td>
<td>0.345</td>
<td>0.483</td>
<td>0.642</td>
<td>0.867</td>
<td>0.605</td>
<td>0.757</td>
</tr>
<tr>
<td>DIS6L</td>
<td>0.685</td>
<td>0.433</td>
<td>0.404</td>
<td>0.569</td>
<td>0.766</td>
<td>0.686</td>
<td>0.655</td>
</tr>
<tr>
<td>DIS7L</td>
<td>0.512</td>
<td>0.369</td>
<td>0.366</td>
<td>0.477</td>
<td>0.771</td>
<td>0.707</td>
<td>0.742</td>
</tr>
</tbody>
</table>

Note: Left: Likert; top: visual analogue scale.
*Correlation significant at 0.05 level (2-tailed).
**Correlation significant at 0.01 level (2-tailed).
†Correlations all significant at 0.01 level (2-tailed).
disorientation items (Davis and Wiedenbeck’s first set of five items), only four out of 10 correlations for the Likert format and three out of 10 correlations for the visual analogue scale response formats were >0.3; this finding indicated that this set was unsuitable for factor analysis. In fact, separate factor analyses for the two response formats, seeking to confirm a four-factor solution, showed that the second set of items did not consistently load on the same factor and showed cross-loadings with other factors. A three-factor solution was therefore pursued.

When extracting three factors, the results of factor analyses using \( z \) factoring and direct oblimin (oblique) rotation showed simple structure 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 4). The factor structure was essentially the same between the two response formats and the percentage of explained variance was similar (unrotated three-factor solutions explained 61.6% and 59.4% of variance with visual analogue scale and Likert format, respectively). Subsequent analyses investigated the psychometric properties of the scales Disorientation, Perceived ease of use and Intensity of flow defined by the items that loaded highly on each of the three factors.

### 3.1.2. Reliability

Using Cronbach’s \( z \) it was found that the scales Disorientation (\( z = 0.87, 0.89 \)), Perceived ease of use (\( z = 0.88, 0.89 \)) and Intensity of flow (\( z = 0.81, 0.83 \)) were all reliable for the Likert and visual analogue scale formats, respectively.

---

**Table 4**

<table>
<thead>
<tr>
<th>Item</th>
<th>Likert</th>
<th>Visual analogue scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Flow</td>
<td>Ease of use</td>
</tr>
<tr>
<td>FLO1</td>
<td>0.196</td>
<td>-0.689</td>
</tr>
<tr>
<td>FLO2</td>
<td>0.236</td>
<td>-0.734</td>
</tr>
<tr>
<td>FLO3</td>
<td>0.116</td>
<td>-0.726</td>
</tr>
<tr>
<td>FLO4</td>
<td>0.192</td>
<td>-0.717</td>
</tr>
<tr>
<td>EOU1</td>
<td>0.074</td>
<td>0.250</td>
</tr>
<tr>
<td>EOU2</td>
<td>-0.083</td>
<td>0.311</td>
</tr>
<tr>
<td>EOU3</td>
<td>-0.210</td>
<td>0.395</td>
</tr>
<tr>
<td>DIS1</td>
<td>0.712</td>
<td>-0.129</td>
</tr>
<tr>
<td>DIS2</td>
<td>0.513</td>
<td>-0.271</td>
</tr>
<tr>
<td>DIS3</td>
<td>0.627</td>
<td>-0.174</td>
</tr>
<tr>
<td>DIS4</td>
<td>0.825</td>
<td>-0.119</td>
</tr>
<tr>
<td>DIS5</td>
<td>0.857</td>
<td>-0.179</td>
</tr>
<tr>
<td>DIS6</td>
<td>0.829</td>
<td>-0.194</td>
</tr>
<tr>
<td>DIS7</td>
<td>0.726</td>
<td>-0.262</td>
</tr>
</tbody>
</table>

*Note: Extraction method: \( z \) factoring; rotation method: direct oblimin.*

The factor extraction technique \( z \) factoring estimates communalities that maximise Cronbach’s reliability co-efficient \( z \) for the factors (Tabachnick and Fidell, 2001); essentially the same factor solution was obtained for both response formats using principal components as an extraction technique.
3.1.3. Validity

Validity was established using Pearson’s correlation (see Table 5). Regarding discriminant validity, for the visual analogue scale format all three correlations between pairs of *Intensity of flow*, *Perceived ease of use* and *Disorientation* were moderate and significant. However, for the Likert format *Intensity of flow* showed moderate, but statistically significant correlations with *Perceived ease of use* and *Disorientation*, but there was a lack of correlation between *Perceived ease of use* and *Disorientation*. Related to this finding, correlations between the two response formats were substantial and significant for *Intensity of flow* and *Disorientation*, but not for *Perceived ease of use*.

Regarding criterion validity, correlations between the three scales and task performance measures of accuracy (number of correct answers) and speed (time-on-task) were not significant except between *Disorientation* and number of correct answers, $r = -0.375$, $p = 0.001$, and $r = -0.330$, $p = 0.005$, for Likert and visual analogue scale format, respectively. Correlations between behavioural measures (measures of page visits) and scales were not significant except between *Disorientation* for the Likert format and visits to the home page, $r = 0.209$, $p < 0.05$.

3.1.4. Sensitivity

First, tests were conducted to determine if presentation of order of response format (first Likert, then visual analogue scale or first visual analogue scale then Likert) had an effect on levels of *Intensity of flow*, *Perceived ease of use* and *Disorientation*. For each scale a $2 \times 2$ ANOVA was conducted with order and scale as independent variables. The effects of order and scale and the interaction effect between order and scale were not significant for *Intensity of flow*, all $F < 1$. For

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Correlations between scales</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Perceived ease of use, Likert</td>
</tr>
<tr>
<td><em>Intensity of flow</em>, Likert</td>
<td>-0.349***</td>
</tr>
<tr>
<td><em>Perceived ease of use</em>, Likert</td>
<td>-0.096</td>
</tr>
<tr>
<td><em>Disorientation</em>, Likert</td>
<td>0.241*</td>
</tr>
<tr>
<td><em>Intensity of flow</em>, visual analogue scale</td>
<td>-0.283**</td>
</tr>
<tr>
<td><em>Perceived ease of use</em>, visual analogue scale</td>
<td></td>
</tr>
</tbody>
</table>

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

**Correlation significant at the 0.01 level (2-tailed).
Perceived ease of use the effects of order, $F(1, 88) = 1.833$, $p > 0.05$, $\text{MS}_{\text{order}} = 1690.075$, and scale, $F(1, 88) = 2.026$, $p > 0.05$, $\text{MS}_{\text{scale}} = 1201.824$, were not significant; however, the interaction effect was significant, $F(1, 88) = 5.404$, $p < 0.05$, $\text{MS}_{\text{interaction}} = 3205.137$. For Disorientation the effects of order and scale were not significant, both $F < 1$; however, the interaction effect was significant, $F(1, 88) = 7.937$, $p < 0.05$, $\text{MS}_{\text{interaction}} = 257.830$.

Next, the effects of scale and orientation support were investigated using a $2 \times 2$ ANOVA with orientation support and scale as independent variables. For Intensity of flow (see Table 6 and Fig. 3), the effects of orientation support and scale were not significant, both $F < 1$; however, the interaction effect was significant, $F(1, 88) = 4.116$, $p < 0.05$, $\text{MS}_{\text{interaction}} = 355.303$. Simple effect tests showed that the effect of orientation support was not significant for either response format, $p > 0.05$.

Because of the significant interaction effect between order and scale for Perceived ease of use and Disorientation, separate analyses were conducted for Likert as first response format and visual analogue scale as first response format. With Likert as the first and visual analogue scale as the second response format (see Table 7), the effect of scale was significant for Perceived ease of use, $F(1, 43) = 7.189$, $p = 0.01$, $\text{MS}_{\text{scale}} = 4288.539$; however the effect of orientation support and the interaction

<table>
<thead>
<tr>
<th>Orientation support</th>
<th>Response format</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Likert</td>
<td>Visual analogue scale</td>
<td>Overall</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M (S.D.)</td>
<td>M (S.D.)</td>
<td>M (S.D.)</td>
<td></td>
</tr>
<tr>
<td>No page title</td>
<td>46.71 (25.04)</td>
<td>48.62 (24.48)</td>
<td>47.66 (24.23)</td>
<td></td>
</tr>
<tr>
<td>Page title</td>
<td>47.25 (19.57)</td>
<td>43.54 (17.39)</td>
<td>45.39 (16.86)</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>46.99 (22.22)</td>
<td>45.96 (21.11)</td>
<td>46.48 (20.62)</td>
<td></td>
</tr>
</tbody>
</table>

Table 6

Effects of scale and orientation support on flow

Fig. 3. Interaction between response format and order of presentation for ease of use.
effect were not, both $F < 1$. With visual analogue scale as the first and Likert as the second response format, the effects of orientation support and scale were not significant, both $F < 1$; furthermore, the interaction effect was not significant, $F(1,43) = 1.717$, $p > 0.05$.

With Likert as the first and visual analogue scale as the second response format (see Table 8), the effect of orientation support was significant for Disorientation, $F(1,43) = 4.978$, $p < 0.05$, $MS_{orientation\ support} = 3267.794$; however the effect of scale, $F(1,43) = 1.956$, $p > 0.05$, $MS_{interaction} = 59.750$, and the interaction effect were not, $F(1,43) = 1.450$, $p > 0.05$, $MS_{interaction} = 44.310$. With visual analogue scale as the first and Likert as the second response format (see Table 9), the effect of scale was significant, $F(1,43) = 6.014$, $p < 0.05$, $MS_{scale} = 209.419$; however, the effect of orientation support and the interaction effect were not significant, both $F < 1$.

Correlations between Disorientation and experience in using the Web were significant for both the Likert and visual analogue scale formats, $\rho = -0.249$, $p < 0.05$ and $\rho = -0.269$, $p = 0.01$, respectively. The correlation between experience and Perceived ease of use was also significant for the visual analogue scale

---

**Table 7**

Effects of scale and orientation support on ease of use (Likert first, visual analogue scale second)

<table>
<thead>
<tr>
<th>Orientation support</th>
<th>Response format</th>
<th>Likert M (S.D.)</th>
<th>Visual analogue scale M (S.D.)</th>
<th>Overall M (S.D.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No page title</td>
<td></td>
<td>45.50 (31.31)</td>
<td>62.78 (25.32)</td>
<td>54.14 (24.15)</td>
</tr>
<tr>
<td>Page title</td>
<td></td>
<td>52.55 (30.12)</td>
<td>62.94 (29.66)</td>
<td>57.75 (23.10)</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>49.26 (30.53)</td>
<td>62.87 (27.41)</td>
<td>56.06 (23.39)</td>
</tr>
</tbody>
</table>

**Table 8**

Effects of scale and orientation support on disorientation (Likert first, visual analogue scale second)

<table>
<thead>
<tr>
<th>Orientation support</th>
<th>Response format</th>
<th>Likert M (S.D.)</th>
<th>Visual analogue scale M (S.D.)</th>
<th>Overall M (S.D.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No page title</td>
<td></td>
<td>34.58 (18.97)</td>
<td>34.35 (18.60)</td>
<td>34.47 (18.26)</td>
</tr>
<tr>
<td>Page title</td>
<td></td>
<td>23.91 (18.81)</td>
<td>20.87 (17.80)</td>
<td>22.39 (17.99)</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>28.89 (19.43)</td>
<td>27.16 (19.22)</td>
<td>28.03 (18.92)</td>
</tr>
</tbody>
</table>
Experience was not significantly correlated with performance measures (accuracy and speed) or behavioural measures (pages visited), \( p > 0.05 \).

3.2. Preference of response format

Using Fisher’s exact test, preference of response format was found to be independent of orientation support and first response format, both \( p = 1.00 \). A chi-square test showed that a statistically significant majority (79%) preferred the Likert response format, \( \chi^2(1) = 30.044, p < 0.001 \). Participants identified various advantages and disadvantages of both response formats (see Table 10). Overall, participants cited clarity of response as the main advantage of Likert, stating that the categories allowed them to give more consistent and precise reactions to the statements. Participants also highlighted ease of use as a benefit, citing the high degree of control over responses. In contrast, participants stated that the fluidity of the visual analogue scale response format allowed them a greater range of responses without being restricted to numbers, which helped them to reflect their opinions.
more accurately. In addition, some participants found that mouse operation of the slider on the items with visual analogue scale was quick and easy to use.

In terms of disadvantages, participants pointed to the issue of the difficulty of categorization using Likert. For example, participants identified that they had difficulty in choosing a response when they felt their opinion lay between ‘agree’ and ‘strongly agree’. Participants also stated that they were inclined to make responses from within a limited range of responses, in particular by avoiding extreme responses, and they also believed that Likert might lead to a bias towards neutral answers. Visual analogue scale was identified as leading to a lack of precision in responses due to a lack of category markers. Linked to this, problems with consistency of responses between questionnaire items were also identified. For example, participants said that although they wanted to give the same response to different statements, they found it difficult to replicate answers from one statement to the next. Participants also had problems with the usability of the visual analogue scale in terms of coordinating their control of the mouse; several participants found the visual analogue scale to be annoying.

Although participants believed that first, visual analogue scale allowed a greater range and Likert a more restricted range, and second, visual analogue scale made it more difficult to give consistent answers between statements, in reality the variability in scores was similar between the two response formats, with variance ratios of visual analogue scale to Likert of approximately 0.9 (0.90, 0.85 and 0.89 for Intensity of flow, Perceived ease of use and Disorientation, respectively). Even though participants believed that Likert might lead to a bias towards neutral answers, after converting overall scores on the three scales back to a 7-point range, differences in frequencies for the middle neutral response category were within 6% between the two response formats (6%, 1% and 4% more neutral answers with visual analogue scale than Likert for Intensity of flow, Perceived ease of use and Disorientation respectively). 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 1% between the two response formats (no difference in, 1% more and 1% fewer extreme lowest answers with visual analogue scale compared to Likert for Intensity of flow, Perceived ease of use and Disorientation respectively). Similarly, differences in frequencies for the extreme highest response category were within 4% between the two response formats (1% fewer, 4% more and no difference in extreme lowest answers with visual analogue scale compared to Likert for Intensity of flow, Perceived ease of use and Disorientation respectively).

3.3. Task performance and navigation behaviour

Unrelated \( t \)-tests showed that the effect of orientation support on time-on-task when giving a correct answer, \( t(88) = -2.09, p < 0.05 \), was significant, with mean = 41.08 s (s.d. = 168.62) and mean = 35.15 s (s.d. = 111.26) for pages with and without orientation support, respectively. However, the effect on other measures (number of correct answers, number of web pages visited when giving a correct
answer, and time-on-task and number of pages visited when giving an incorrect answer) was not significant (all $p > 0.05$).

4. Discussion

The present study is the first to investigate the psychometric properties of three existing quality of interaction scales simultaneously including their sensitivity to a difference in orientation support. In addition, we examined the influence of response format on the quality of interaction with web pages.

Results from the current study confirm the existence of three factors representing disorientation, perceived ease of use and flow that had previously been identified in separate data sets. Three scales, Disorientation, Perceived ease of use and Intensity of flow, corresponding with the factors were found to be reliable. With visual analogue scale as the response format, validity among the three scales was confirmed through moderate, but significant, correlations among the scales. With Likert as the response format, Perceived ease of use lacked validity due to a non-significant correlation with Disorientation, although the other two correlations among scales were significant. Validity of Disorientation and Intensity of flow was further confirmed by high and significant correlations between the two response formats. However, a lack of validity of Perceived ease of use for Likert was further confirmed by a non-significant correlation between the two response formats for this scale. Additional evidence for validity of Disorientation, but not the other two scales, was found through a significant negative correlation with accuracy of task performance for both response formats and a positive correlation with a behavioural measure (visits to the home page) for Likert format. Some evidence was found for the sensitivity of the scales. In particular, the Intensity of flow scale showed an interaction between orientation support and response format, with a larger difference between the two levels of orientation support with visual analogue scale compared to Likert. More importantly, with Likert as first response format, the effect of orientation support was significant, irrespective of response format, with higher levels of Disorientation without orientation support than with orientation support (that is page title on each page in the web site). Furthermore, a negative correlation was found between Disorientation and Web experience for both response formats. Time-on-task was also sensitive to orientation support, with more time spent on pages with than without orientation support. Further results showed a tendency of visual analogue scale to produce higher scores than Likert. In particular, with Likert as first response format, visual analogue scale produced higher Perceived ease of use scores than Likert. With visual analogue scale as first response format, visual analogue scale produced higher Disorientation scores than Likert.

Overall, the advantages and disadvantages cited in the literature (see Table 1) were confirmed by participants’ comments. A large majority of participants preferred Likert as the response format. However, various advantages and disadvantages of both response formats were identified. Perceived clarity of response for Likert was matched with a perceived lack of clarity and consistency of response for visual
analogue scale, although neither of these was confirmed by differences in variability of scores on the three scales between response formats. The advantage of degree of choice for visual analogue scale was matched with both difficulty of mapping judgement to a numerical response format and a perceived response bias, even though none of these was confirmed by our findings. The reported experience of ease of use for Likert was matched with a perceived lack of ease of use for visual analogue scale; however, ease and speed of use were also mentioned as an advantage of visual analogue scale, although less frequently. These results indicate that first, experience with using mouse-operated sliders may make visual analogue scale easier to use, and second, the usability of the slider could be improved. A difficulty with the use of a slider is that when positioning the mouse, users may inadvertently release the left mouse button that is used to control the slider before having arrived at the response of their choice. In order to avoid this, the current experiment required participants to confirm their answer after releasing the left mouse button.

The alleged advantage of Likert of easy interpretation of relevant changes compared to visual analogue scale appears to be a false one; correctness, rather than ease of interpretation can be achieved through standardization, as for example in tests of intelligence (Kline, 2000). The alleged disadvantage of having to convert visual analogue scale responses to numbers did not apply in the current study because, as data were collected by computer, visual analogue scale scores were already in numeric format and thus no conversion to numeric format was required.

Several of these results confirm and extend the findings of the landmark study by Ahuja and Webster (2001). In particular, two separate scales for disorientation and perceived ease of use (through factor analysis) and reliability of the scales Disorientation and Perceived ease of use were confirmed, but now also in combination with a flow factor, and further extended by the use of both discrete (Likert) and continuous (visual analogue scale) response formats. Theoretically, cognitive problems in wayfinding in an information space (Smith, 1996), should be negatively related to ease of use and positively to flow and should increase with causes of these problems, such as a lack of orientation support. Indeed, the validity of Disorientation was confirmed for both response formats with Intensity of flow and Perceived ease of use as predicted (except for a non-significant correlation with Perceived ease of use presented in Likert format), task performance and, only to a limited extent, navigation behaviour. Evidence for sensitivity of the scales was found. In particular, orientation support did have an effect on Disorientation (when Likert was the first response format) with more disorientation experienced without orientation support, but not Perceived ease of use, confirming that Disorientation is a more sensitive measure than Perceived ease of use. The existence of a significant positive correlation of Disorientation with web experience further confirms the sensitivity of the scale. The finding that time-on-task increased with orientation support reflects the extra time required for reading page titles.

Our results confirm and extend results from previous research by Davis and Wiedenbeck (2001). The existence of one of the two flow factors (‘involvement’, but not ‘control’) identified by these authors was confirmed in the present study, this time in combination with the factors of disorientation and perceived ease of use and
formed a reliable *Intensity of flow* scale. Theoretically, flow should correlate positively with perceived ease of use and negatively with perceived disorientation. Indeed, *Intensity of flow* showed validity through significant correlations with *Perceived ease of use* (as in previous research) and *Disorientation*. However, there was a lack of evidence for a control factor within Davis and Wiedenbeck’s set of flow items, which is a pre-requisite of reliability and validity. Arguably, the concept of control may have limited applicability to web sites, which are mainly characterized by navigation, following hypermedia links, as opposed to other types of application such as computer games, where users are controlling the gaming environment.

In comparing the two response formats, we found that most of their psychometric properties were essentially the same in terms of factor structure and reliability. Regarding validity, the response formats were broadly similar. However, the Likert form of *Perceived ease of use* appeared to show a lack of validity through non-significant correlations with *Disorientation* and with *Perceived ease of use* in visual analogue scale format. This apparent, but unexpected, lack of validity requires additional research. Furthermore, *Disorientation* presented in Likert format—but not when presented in visual analogue scale format—showed a small, but significant correlation with home page visits. In terms of sensitivity, the significant effect of orientation support on *Disorientation* occurred irrespective of response format. The interaction between response format and orientation support on *Intensity of flow* also provides some support for validity of visual analogue scale rather than Likert.

We found significant positive correlations between a 7-point Likert and 10-cm visual analogue scale, but not for *Perceived ease of use*, just as previous research 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). The level of ratings was higher with visual analogue scale than Likert for *Perceived ease of use*, when Likert was the first response format, and for *Disorientation*, when visual analogue scale was the first response format. These results confirm the results of Neely and Borg (1995), who found higher scores for 10-cm visual analogue scale than 11-point Likert (ranging from 0 to 10) when rating the blackness of squares, but contrast with the results that found higher scores for Likert. Our results are different from Murphy et al. (1988), Price et al. (1994), Flynn et al. (2003), who all found higher ratings on Likert scales than visual analogue scale. Although Pfennings et al. (1995) found a higher variability of scores when using 10-cm visual analogue scale compared to 10-point Likert, our results showed no difference in variability between Likert and visual analogue scale with variance ratios of visual analogue scale to Likert in the order of 0.9. In terms of sensitivity, the two response formats produced essentially similar results, with a lack of sensitivity to differences in orientation support, except for *Disorientation* when Likert for first response format. The only scale where the two response formats may not have been equally sensitive was *Intensity of flow*, where there was some evidence of visual analogue scale being more sensitive to differences in orientation support. Previous research by 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. Results of Hayes et al. (1996) showed 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.

The results can largely be explained by the graduation of the Likert format. A 7-point scale may have sufficient categories to produce satisfactory psychometric results. In contrast, research using Likert with a smaller number of response alternatives (4-point Likert) resulted in a lack of responsiveness for Likert compared to visual analogue scale (Joyce et al., 1975; Bellamy et al., 1999a, b). Although a 7-point Likert appeared to have similar psychometric properties as visual analogue scale in our study, the lack of validity for Perceived ease of use with Likert requires replication and the sensitivity of both Likert and visual analogue scale needs further investigation with stronger experimental manipulation of web site designs in order to reveal significant differences between design alternatives on the three scales.

Although the current study did not investigate the equivalence of on-line and paper-based administration of questionnaires, the similarities in psychometric properties between the current study and previous research that used the same scales (Ahuja and Webster, 2001; Davis and Wiedenbeck, 2001) are encouraging. This finding should be considered light of previous research. First, direct comparisons of administration formats found positive evidence for the equivalence of paper-based and on-line formats (e.g. Slaughter et al., 1994), including an apparently higher quality of answers to open-ended questions gained from on-line administration, with better clarification of problems, strengths and often examples. Second, a successful validation of a well-established questionnaire (Questionnaire for User Interaction Satisfaction, QUIS) administered through the Web (Harper et al., 1997) indicated the increased scope of psychometrically sound measurement in situations that are not or cannot be experimentally controlled. This is very important because numerous potential applications of the administration of psychometric instruments occur with respondents who are remote in place and time from the research team, but who do have access to the Web.

Further research should investigate other response formats and experimental manipulations of web site parameters to demonstrate sensitivity of measures.

4.1. Response formats

A major distinction in response format for rating scales is between discrete (in particular Likert) and continuous (in particular visual analogue scale). However, within both types of format, several further dimensions exist that have implications for screen layout in the on-line administration of questionnaires and these dimensions may affect the psychometric properties of scales. A first dimension that applies to both types is the orientation of the scale. Usually scales are presented with the response range presented horizontally. However, a vertical presentation is possible and this is sometimes found in questionnaires on the Web, using a list. Within the vertical format, a parameter is visibility of response alternatives. This can
be particularly important when several questions are presented on a page. If a drop-
down list box is used, users cannot see the response range of previously answered
and still unattempted questions; at any moment in time, they can only see the range
for the question that they are answering at that moment. Another alternative is a list
box, where all response alternatives are visible at any time. Visibility will inevitably
have an impact on screen design and potentially the quality of the layout of on-line
questionnaires as perceived by respondents and users' interactions with question-
naires, and these may affect psychometric properties.

A second dimension is scale range. In the case of continuous scales, research using
visual analogue scale administered with pencil and paper typically uses a line length
of 10 cm. In the case of discrete scales, typically 7-point and 5-point scales are used
(in many studies, the physical length of scales is not reported), although wider ranges
are also used in instruments that are administered using pencil and paper (for
example Task Load Index, Hart and Staveland, 1988). Scale range has by definition
an impact on screen layout of discrete and continuous scales and may also affect the
quality of perceived layout of on-line questionnaires when presenting scales. Both of
these may affect the psychometric properties of scales. In particular, continuous
scales that are wide and discrete scales with many response categories may lead to
increased variability in scores due to measurement error if raters cannot use the
range effectively. However, increased variability may also be an indication of higher
responsiveness (Pfennings et al., 1995). On the other hand, scales that are narrow or
have few response categories may result in reduced variability because raters cannot
sufficiently express differences in judgement (Kline, 2000), with a resultant lack of
sensitivity.

A third dimension is the representation of discrete response categories: using
standard user interface elements, such as radio buttons (for horizontal and vertical
formats) and lists (for vertical format) or bespoke interface elements. Again, the type
of interface element used will have an impact on screen design and potentially the
perceived quality of layout of on-line questionnaires and this may also affect
psychometric properties.

4.2. Sensitivity

The current study found some evidence of the sensitivity of Disorientation with
responses being affected by the experimental manipulation of orientation support
with Likert as first response format. However, with a stronger appropriate
experimental manipulation of web site designs in terms of information architecture
(Rosenfeld and Morville, 2002), evidence of sensitivity would be expected across a
range of measures. In particular, more dramatic differences in outcomes would be
expected with the experimental manipulation of a web site's navigation systems, as
found by, for example, Ahuja and Webster (2001). In theory, measures of perceived
disorientation and flow would be expected to be sensitive to manipulations because
first, navigation support should enhance orientation, and second, navigation support
should avoid disrupting a user's engagement with a site. However, the motivational
value and degree to which a site is engaging will depend on both the nature of the site
and the task carried out. Task performance may be less sensitive to experimental manipulations because participants will put more effort in when tasks become more difficult, and as a consequence, error rates or performance speeds will remain fairly constant over a range of experimental manipulations (Spenkelink and Besuijen, 2003). More generally, as Spenkelink and Besuijen note ‘it seems that, in general, performance tasks are just not sensitive enough for…individual effects to show’. However, with increasing differences in experimental manipulation, fatigue should increase and task performance should ultimately also show sensitivity to differences. Although Ahuja and Webster did not find evidence for sensitivity of behavioural measures in response to experimental manipulation, behavioural measures and perceived ease of use should eventually show sensitivity to large differences in experimental manipulation. Other components of information architecture that should also affect the perceived quality of a web site are organization systems and labelling systems as well as user interface design.

In conclusion, the factor structure, reliability, validity and, to some extent, sensitivity of the Disorientation instrument was confirmed for both Likert and visual analogue scale response formats. The results on Disorientation should give practitioners and researchers increased confidence in the on-line administration of this scale in addition to other methods of assessing the relative usability of web site designs. However, further investigation of its psychometric properties in general and in particular with different response formats is required. Standardization that has been achieved with other psychometric measures, for example the SUMI (Kirakowski, 1994), should be a goal for future research because this will allow a judgement about the absolute degree of perceived disorientation of a web site design rather than merely a comparison between designs. As suggested by Ahuja and Webster (2001), perceived ease of use ‘will remain a key construct in the evaluation of information systems’ (p. 27). However, a lack of validity of Perceived ease of use with Likert format and a lack of sensitivity found in this study and by Ahuja and Webster currently limit its usefulness for the on-line evaluation of web-based systems. The construct of flow, in particular the involvement component, can be important in the experience and use of web sites. Nonetheless, the lack of sensitivity, especially with Likert scales, observed here currently restricts usefulness of Intensity of flow when applied to the on-line evaluation of web sites.
References


