Questionnaire layout and national culture in online psychometrics

Paul van Schaik[a*], Su Luan Wong[b] and Timothy Teo[c]

To appear in International Journal of Human-Computer Studies

[a]School of Social Sciences and Law, Teesside University, United Kingdom
[b]Department of Science and Technical Education, Universiti Putra Malaysia, Malaysia
[c]Faculty of Education, University of Macau, Macau, People's Republic of China

[*]Corresponding author. School of Social Sciences and Law, Teesside University, Middlesbrough, TS1 3BA, United Kingdom
T +44 1642 342320. F +44 1642 342301. E p.van-schaik@tees.ac.uk

Questionnaire layout and national culture in online psychometrics

Abstract. Given the proliferation of online psychometric questionnaires on the Internet and other platforms, the design of online psychometrics becomes increasingly important to ensure good measurement properties. The current study reports four experiments testing how questionnaire layout and national culture affect people's responses to online questionnaires. Flow- and the disorientation scales for web navigation had good psychometric quality overall and across experimental manipulations of questionnaire layout, field dependence, national culture, response correction and question grouping. However, single-item layout (presenting one question at a time) had the advantage of faster completion than whole-form layout (presenting each page filled with items). Support was also found for the idea that whole form divides and thereby requires more attention to respond. Future research should be directed at online psychometrics on small and large displays.

Keywords: psychometrics; online questionnaire; questionnaire layout; culture; dual tasking.

Additional keywords: human-computer interaction, website, screen design.

Highlights

Disorientation and flow scales for web navigation have good measurement quality

Single-item layout leads to faster psychometric-questionnaire completion

Single-item layout facilitates focusing attention in questionnaire completion

National culture does not affect measurement quality

1. Introduction

The advent of popular online survey tools facilitates the creation of psychometric and other online questionnaires that can be administered on different platforms, including desktop and laptop computers, but also tablets and smartphones. As a result, the design of online psychometrics becomes increasingly important to ensure good measurement properties of the data that are collected (Evans et al., 2009). Indeed, research into online psychometrics to inform this design is especially timely, given the widespread use of online questionnaires internationally. This research has potential implications for software developers and, more generally, private-sector, public-sector and third-sector organisations that use online questionnaires. A distinction needs to be made between surveys to collect factual information (e.g. census information; Norman et al., 2001) and psychometric questionnaires to measure people's abilities, attitudes or perceptions in various domains, including human-computer interaction (e.g. disorientation experienced by the users of a website; Ahuja & Webster, 2001). When factual information is collected, the aim is to collect responses that are deliberate and accurate, but in psychometric measurement the aim is to collect responses that are spontaneous (van Schaik & Ling, 2007); the latter normally involves giving speedy, but not careless answers.

1.1. Questionnaire layout in psychometrics

Despite the ubiquity of online psychometrics, research into human-computer-interaction design of online psychometrics is still scarce (but see van Schaik & Ling, 2003; 2007). Existing research on attitude surveys that have examined the context in which an item is presented and how an item is combined with others in a scale (Moore, 2002; Schwarz et al., 1991; Schwarz, 1999). However, the research reported here is directed at a different important consideration in the design of online

psychometrics: questionnaire layout. Norman et al. (2001) studied different questionnaire layouts to present and administer an online factual survey, at least two of which can also be used in online psychometric questionnaires. First, single-item layout involves presenting one questionnaire item (question) at a time. Respondents move to the next question automatically when they have answered the current question (as in Figures 1a, 2a and 3a) or by pressing a control (e.g. a button labelled Next). Second, in whole-form layout the screen is filled with items and respondents can move through the items by pressing controls (e.g. Up and Down, as in Figures 1b, 2b and 3b). Van Schaik and Ling (2007) found that single-item layout was faster and had some advantage in terms of psychometric structure. However, this research studied neither the role of attention nor individual-difference variables in relation to questionnaire layout, and only one type of task and one type of website were investigated, and used traditional factor analysis to assess psychometric structure.

A justification for using single-item layout is that it precludes distraction by simultaneously presented items that can occur with other layouts (such as whole form) (van Schaik & Ling, 2007). This layout supports the aim of collecting spontaneous responses from respondents because there is only one presented item to attend to at any one time. Whole-form layout, by contrast, facilitates extensive deliberation by making responses to previous questions and future questions visible at the same time as the next question to be answered. In other words, single-item layout promotes focused attention on one item at a time. By contrast, whole-form layout can break attention because it allows many presented items to compete for attention. Thus, respondents' attention may inadvertently switch back and forth

between the next question to be answered and simultaneously presented previous and following items. Therefore,

Hypothesis 1: time to complete psychometric items is longer under whole-form layout than under single-item layout.

Because whole-form layout is expected to divide the attention of respondents to online questionnaires among items, more attention may be needed to focus on the next question to be answered. To test this idea, it becomes important to measure the amount of attention that is needed for performing the task of responding to psychometric items. One common research technique to measure this is dual tasking (Wickens & McCarley, 2008). Participants are asked to perform two tasks simultaneously, generally one primary task (e.g. responding to a psychometric questionnaire) and a secondary task (e.g. responding to a specific number in a continual series of random numbers). If secondary-task performance deteriorates more in one questionnaire layout (e.g. whole-form layout) than in another (e.g. single-item layout) then this is an indication that the former requires more attention (Wulf & Lewthwaite, 2010). Given the potential effect of whole-form layout in dividing attention and thereby requiring more attention to answer psychometric items, and dual tasking as a technique to measure the amount of attention needed for primary-task performance, follows

Hypothesis 2: secondary-task performance is better with single-item layout than with whole-form layout.

When the primary task (here responding to psychometric items) requires more attention (as established with dual tasking), people may be able to compensate by putting in additional effort to maintain task performance (Spenkelink & Besuijen, 1999) and produce data with the same measurement properties; however, if too

much attention is needed, primary-task performance (and here, consequently, psychometric quality) will be degraded. For example, people may follow simple strategies such as answering randomly, repeating the previous answer or consistently selecting the middle rating-scale response; such strategies will potentially result in, for instance, poor factor structure and a lack of discriminant validity. Here, single-item presentation may have an advantage over whole-form layout, as the latter will inherently divide attention and therefore require more attention to focus on the next question to be answered, whereas the former will require less attention. Thus, psychometric quality may not be equivalent between questionnaire layouts if whole form requires more attention than single item.

1.2. Field dependence, national culture and psychometrics

In addition to design factors, such as questionnaire layout, individual-difference variables may influence how people process information when responding to online psychometrics. In relation to focusing attention on the next questionnaire item to be answered, a distinction needs to be made between those who have a propensity to process information holistically (processing together several simultaneously presented items) rather than analytically (processing separately simultaneously presented items), as they have a tendency to focus their attention less on individual items and their characteristics separately. This dimension of information-processing is known as field dependence, where holistic processors are field-dependent and analytic processors are field-independent (Nisbett et al., 2001). In reality, field dependence is not a matter of all or nothing, but graded, so people possess varying degrees of field dependence. Moreover, research in cognitive psychology has proposed and found evidence for the idea that field-independence makes people less susceptible to distraction from irrelevant information (Meys & Sanderson, 2013;

Miyake et al., 2001; Ray, 1974). Given this research and given that field dependence may divide attention between the next question to be answered and other items on the screen in online psychometric items, follows

Hypothesis 3: field independence reduces susceptibility to distraction by a secondary-task.

An extensive body of research has examined 'cultural invariance' of psychometrics. For example, researchers have considered the measurement invariance of instruments ranging from teachers' perception of grading practices scale (TPGP) between US and Chinese students (Liu, 2011) to the basic psychological needs in exercise scale (BPBES) among Greek, Spanish, Portuguese, and Turkish students (Vlachopoulos et al., 2013) and the Educational Needs Assessment Tool (ENAT) among Dutch, Finnish, Norwegian, Portuguese, Spanish and Swedish (Ndosi et al., 2011). In these studies, the focus was on participants' responses to the content in the scale items and how these items had contributed to construct validity and reliability. However, research on cultural invariance has not examined online psychometrics in relation to design parameters (such questionnaire layout) and, notably, field dependence in relation to psychometrics. Research on attention and perception has demonstrated that field dependence is linked to national culture: "Westerners tend to engage in context-independent and analytic perceptual processes by focusing on a salient object independently of its context, whereas [East] Asians tend to engage in context-dependent and holistic perceptual processes by attending to the relationship between the object and the context in which the object is located" (Nisbett & Miyamoto, 2005, p. 467) and "they should find it more difficult than Americans to isolate and analyze an object while ignoring the field in

which it is embedded" (Nisbett et al., 2001, p. 296). Therefore and if field dependence is linked to national culture (Western versus East-asian), follows *Hypothesis 4*: national culture is a determinant of susceptibility to distraction by a secondary-task, with Western cultures being less susceptible than East-asian.

A similar argument as that made above regarding questionnaire layout in relation to psychometric quality can be made regarding field dependence and national culture. First, psychometric quality may not be equivalent between those who are more field-dependent and those who are less field-dependent if the former require more attention in focusing on the primary task of responding to psychometrics. Second, likewise, psychometric quality may not be equivalent between those from East-asian cultures and those from Western cultures if the former are more field-dependent.

1.3. Outline of experiments

As Davis and Venkatesh (1996) conducted a series of experiments to clarify the merits and demerits of question grouping, so we conducted a series of experiments to clarify the benefits and drawbacks of questionnaire layouts. The aim of the current study is to test the effects, according to the hypotheses, of psychometric questionnaire layout, field dependence and national culture on questionnaire completion and secondary-task performance. In each of a series of computer-based experiments, participants first navigated a website and then, using psychometric questionnaires, rated the quality of their experience with the website. The presentation of the questionnaires was manipulated to analyse the effect of questionnaire layout on time to respond and secondary-task performance. A pilot experiment tested Hypotheses 1 and 2, examining the effect of questionnaire layout under single- and dual-task conditions. Experiments 1 and 2 tested Hypotheses 1 through 4, also examining the effects of field dependence and national culture.

Using a different type of website and a different type of web navigation task, Experiment 3 tested Hypotheses 1 and 2 and also analysed the sensitivity of psychometric measurement.

2. Pilot experiment

2.1. Method

2.1.1. Design

A 2×(2) experimental design was used. The first within-subjects independent variable was rating-scale task condition (single or dual) and the second between-subjects variable was questionnaire layout (single item and whole form) (see Figure 1). Dependent variables were time to complete rating-scale items and secondary-task performance.

2.1.2. Participants

Students from Kingston University, England, (*N* = 42; 25 female) took part. Mean years of experience with using Internet sites was 8.40 (*SD* = 2.64). Mean frequency per week of using Internet sites was 28.31 hours (*SD* = 20.02).

2.1.3. Materials and equipment

The experiment was programmed using Visual Basic 6.0. The experiment ran on personal computers (Intel Pentium, 2.8 GHz, 512 Mb RAM, Microsoft Windows XP operation system) with 17-inch monitors and a screen resolution of 1024×768 pixels. Contrast and brightness were set to optimal levels. Guo and Poole's (2009) 30-item

psychometric flow scale for websites was used to measure flow experience[1] (see Appendix), with a 7-point Likert scale.[2]

2.1.4. Procedure

Participants took part individually. As an introduction to the website, they first viewed 10 typical *Wikipedia* pages (*http://en.wikipedia.org/*). They then explored the website for 20 minutes. The mean number of Wikipedia pages visited was 22.64 (*SD* = 17.31). Finally, they completed the flow scale in two different layouts (single item and whole form) in counterbalanced order. In the rating-scale task conditions with single tasking, participants only completed the flow scale. In the conditions with dual tasking, participants' primary task was to complete the flow scale; their secondary task was to respond to a random digit (ranging from 0 to 9), presented each second, when its value equalled 0 (see Figure 1c, 1d). In the whole-form conditions, the psychometric scale had seven items presented on the first page (see Figure 1), eight on the following three pages and three on the last page. After the first page, on each page the last item of the previous page was repeated. The experiment took approximately 35 minutes to complete.

---

[1] the feeling of being fully involved in a particular activity, in this case browsing a website (van Schaik & Ling, 2012b)

[2] The flow scale has previously been shown to have good psychometric properties (Guo & Poole, 2009; van Schaik & Ling, 2012a, 2012b). This flow scale provides conceptually considerably more comprehensive measurement than the flow scale previously used by van Schaik and Ling (2003, 2007).

2.2. Results[3]

2.2.1. Psychometric analysis

Psychometric analysis is presented in Online Appendix A0. The flow subscales were found to possess good internal-consistency reliability and evidence was found for convergent and discriminant validity.

2.2.2. Time to complete rating-scale items

Overall, descriptives indicate that time to complete was shorter under the single-tasking conditions than under the dual-tasking conditions, in particular with whole-form layout (see Table 1). Mixed analysis of variance (ANOVA) confirmed that the interaction effect of questionnaire layout and tasking was significant, F (1, 40) = 4.18, $\varepsilon^2 = 0.03$[4], $p < 0.05$, but the main effects of questionnaire layout, F (1, 40) = 1.17, $p > 0.05$, and tasking, F (1, 40) = 2.13, $p > 0.05$, were not.[5] Furthermore, simple-effect tests showed that the effect of questionnaire layout was significant under dual tasking, t (17) = 3.25, $r = 0.62$, $p < 0.01$, but not under single tasking, $|t| < 1$. Additional simple effects showed that the effect of tasking was significant for whole form, t (40) = 2.49, $r = 0.37$, $p < 0.05$, but not for single item, $|t| < 1$.

2.2.3. Secondary-task performance

According to descriptives, whole-form layout produced more false alarms per secondary-task stimulus (mean [SD]: 0.37 [0.24]) than single-item layout (0.25 [0.18]). Related t tests confirmed that the difference was significant, t (17) = 2.86, r =

---

[3] For the sake of brevity, throughout this paper statistically non-significant results are not always reported. For example, in analysis of variance, the results for effects that are statistically significant are always reported, but not always the results for effects that are not.

[4] An approximately unbiased estimate, ε^2, is used in favour of η^2 (Jaccard, 1998).

[5] Time to complete was positively skewed and therefore for this analysis the distribution was improved by a logarithmic transformation.

0.57, $p < 0.05$. The correlations between time to complete (for both questionnaire layouts) and secondary-task performance (false alarms and hits) were positive (average $r = 0.25$), but given the small sample and therefore low statistical power not significant.

2.3. Summary of results

First, the flow scale was reliable in both questionnaire layouts and showed evidence of convergent and discriminant validity. Second, with whole-form layout, time to complete was faster, supporting *Hypothesis 1*, but only under single tasking. Third, single-item layout produced fewer alarms on the secondary task, in support of *Hypothesis 2*. Overall, the results indicate that single-item layout is advantageous. Experiment 1 expanded on the pilot experiment by also testing the effects of field dependence (*Hypothesis 3*) and national culture (*Hypothesis 4*) on secondary-task performance.

3. Experiment 1

3.1. Method

3.1.1. Design

A 2×2×2×2×2 independent-measures experimental design was used. The independent variables were (1) questionnaire layout (single item or whole form; see Figure 2); (2) rating-scale task condition (single task [answering questions only] or dual task [simultaneously performing a secondary task]; see Figure 2); (3) national culture (English [Western] or Malaysian [East-asian]); (4) correction facility (answers given can be corrected or not); (5) question grouping (questions grouped by scale or

mixed). Correction facility and question grouping were included to establish the generality of the effects of the other independent variables.[6]

Dependent variables were time to complete rating-scale items and secondary-task performance. Field dependence was measured as a potential covariate. For Experiment 1, pilot testing (n = 16) was conducted to confirm that the procedure, materials and equipment worked correctly across the various experimental conditions.

3.1.2. Participants

One hundred and thirty-eight participants were students at Teesside University, England, (n = 66, 32 female) and Universiti Putra Malaysia, Malaysia (n = 72, 42 female). Mean years of experience with using Internet sites was 8.80 (SD = 3.88).

3.1.3. Materials and equipment

The experiment was programmed as a dynamic database-driven website created with Visual Studio 2008 and using C# for server-side programming. The experimental conditions were modelled as relational data tables and the Wikipedia site ran integrated in the experiment program. In England, the experiment ran on personal computers (Intel Pentium, 1.86 GHz, 2 GB RAM, Microsoft Windows XP operating system, 17-inch monitors) with screen dimensions of 1280×1024. In Malaysia, computers of different configurations were used, similar to those used in England. Contrast and brightness were set to optimal levels.

---

[6] The presence of a correction facility would be expected to slow down responding to psychometric items, as it requires a motor action, pressing an additional next button on each screen with questions, to proceed to the next screen (Card, Moran & Newell, 1983); question grouping would be expected to speed up responding as respondents do not have to switch between different conceptual dimensions corresponding with different (sub)scales (Marí-Beffa et al., 2000).

Guo and Poole's (2009) scale was used to measure flow experience, and Ahuja and Webster's (2001) disorientation scale (see Appendix) to measure disorientation in web navigation with Wikipedia, both with a 7-point Likert scale. Field dependence was measured with the Group Embedded Figures Test (GEFT) (Witkin et al., 1971/2002).[7]

3.1.4. Procedure

The experiment ran in a computer laboratory with groups of 15-20 participants who worked independently. Participants first completed the GEFT on paper. Next, they explored the Wikipedia site within the experiment program for 10 minutes. The mean number of reported Wikipedia pages visited was 8.73 (*SD* = 6.49). They then completed the scales of disorientation and flow. In the whole-form conditions, for the flow scale eight items were presented on each of the first three pages (see Figure 2) and six on the last page; all seven disorientation items were presented on one page. The experiment took about 50 minutes to complete.

3.2. Results

3.2.1. Psychometric analysis

The psychometric analysis of scales is presented in Online Appendix A1. Overall, psychometric properties were good for the whole sample and for subgroups according to levels of the independent variables. However, for single-item layout average variance extracted (AVE) for the Flow Subscale Concentration was significantly higher than whole-form layout. Moreover, for Concentration and

[7] In addition to the flow scale (see above), the disorientation scale (Ahuja & Webster, 2001; van Schaik & Ling, 2003, 2005a, 2005b, 2007, 2012a) and the GEFT (Witkin et al., 1971/2002) have previously been shown to have good psychometric properties. The GEFT scale scores range from 0 to 18.

Mergence of Action and Awareness AVE was statistically higher for England than for Malaysia. These results were consistent with significant differences on loadings for some subscales between the two layouts (CT1 and CT4) and the two nations (CT1, CT4 and MA2).[8] In addition, there were statistically significant differences on some flow subscales between the two layouts (FB1, in favour of whole form, and FB2, in favour of single item) and the two nations (CG4, CN2 and TS2 in favour of England).

3.2.2. Field dependence

Total scores on the GEFT were calculated per participant. Original scores were negatively skewed, with mean = 11.90, $CI(mean)_{0.95}$ = [11.14; 12.65]. Scores were transformed through reflection and then a square-root transformation to achieve an approximately normal distribution. Participants from the two nations did not differ significantly on field dependence, t (138) = 1.39, $p > 0.05$. Level of disorientation was positively correlated with transformed GEFT scores, $r = 0.26$, $p < 0.01$. Therefore, with increasing field-dependence, disorientation in navigating the website increased.

3.2.3. Time to complete rating-scale items

For both scales, time to respond per item was positively skewed. After an inverse transformation scores were approximately normally distributed. Descriptives indicate no effect of questionnaire layout (see Table 2). Analysis of variance (ANOVA) was conducted with the independent variables questionnaire layout, correction facility, item grouping, rating-scale task condition (with or without secondary, vigilance task) and nation. For the disorientation scale, there were significant main effects of rating-

[8] AVE should be seen as more important because (sub)scale items are used together, rather than in isolation, to measure a particular construct and therefore an evaluation of the combined items through AVE is more useful.

scale task condition, F (1, 106) = 4.37, ε^2 = 0.02, p < 0.05, correction facility, F (1, 106) = 7.62, ε^2 = 0.04, p < 0.01 and nation, F (1, 106) = 51.25, ε^2 = 0.27, p < 0.001. Completion was faster without secondary task (mean items completed per second = 0.15, *SD* = 0.08) than with (mean = 0.14, *SD* = 0.06). Completion was faster without correction (mean = 0.16, *SD* = 0.06) than with (mean = 0.13, *SD* = 0.07). English participants completed faster (mean = 0.18; *SD* = 0.06) than Malaysian (mean = 0.11; *SD* = 0.06).

For the flow scale, there were significant main effects of nation, F (1, 106) = 40.74, ε^2 = 0.22, p < 0.001, and correction facility, F (1, 106) = 8.31, ε^2 = 0.04, p < 0.01, and of the four-way interaction of correction facility, nation, questionnaire layout and rating-scale task condition, F (1, 106) = 4.63, ε^2 = 0.02, p < 0.05. Simple-effect analysis for whole-form layout showed significant effects of correction, F (1, 53) = 4.20, ε^2 = 0.04, p < 0.05, and nation, F (1, 53) = 26.12, ε^2 = 0.28, p < 0.001. Simple effect analysis for single-item layout showed significant effects of correction, F (1, 53) = 4.12, ε^2 = 0.04, p < 0.05, nation, F (1, 53) = 15.47, ε^2 = 0.17, p < 0.05, and the interaction between rating-scale task condition and nation, F (1, 53) = 4.21, ε^2 = 0.05, p < 0.05.[9] English participants completed faster (mean = 0.17; *SD* = 0.05) than Malaysian (mean = 0.11; *SD* = 0.05). Completion was faster without correction facility (mean = 0.15, *SD* = 0.05) than with (mean = 0.13, *SD* = 0.06).

[9] Further simple effects showed that the main effect of nation was significant with single-item layout and dual tasking, F (1, 53) = 10.58, ε^2 = 0.24, p < 0.01, and with single-item layout and single tasking, F (1, 29) = 5.35, ε^2 = 0.10, p < 0.05.

3.2.4. Secondary-task performance

Performance on the secondary task was measured as proportion of responses per stimulus for participants who took part in the experimental conditions with secondary task.[10] For both scales, performance was negatively correlated with transformed GEFT scores: $r = -0.31$, $p < 0.05$, for disorientation and $r = -0.46$, $p < 0.001$, for flow. Therefore, with increasing field-dependence respondents were more detracted by the secondary task.

Given that secondary-task stimulus presentation was randomised, on the performance measure some of the cells for the research design for the flow scale had no variance because there were zero or one participants to whom secondary-task stimuli had been presented. Therefore, a reduced design was analysed consisting of all main effects of the independent variables, their two-way interactions, the covariate field dependence and all interactions between the main effects and the covariate. Analysis of covariance (ANCOVA) was conducted with the independent variables questionnaire layout, correction facility, item grouping and nation, and the covariate. For the disorientation scale, there were no significant main effects, but the interaction effect between nation and correction facility approached significance, $F(1, 29) = 4.15$, $\varepsilon^2 = 0.05$, $p = 0.051$. Without correction facility, adjusted mean scores (*SE*) were 0.67 (0.08) for England and 0.16 (0.11) for Malaysia, but with correction facility, adjusted mean scores (*SE*) were 0.58 (0.08) for England and 0.47 (0.11) for

[10] This measure was used rather than hits, false alarms and misses, as many of the participants' responses to a stimulus were delayed beyond the start of the next stimulus.

Malaysia.[11] The correlation between time to complete and secondary-task performance was significant, $r = 0.38$, $p < 0.05$.

For the flow scale, in ANCOVA the effect of the covariate was significant, F (1, 37) = 14.81, $\varepsilon^2 = 0.16$, $p < 0.001$, as was the interaction effect of correction facility by nation, F (1, 37) = 11.06, $\varepsilon^2 = 0.11$, $p < 0.01$.[12] Without correction facility, adjusted mean scores (*SE*) were 0.55 (0.06) for England and 0.20 (0.07) for Malaysia, but with correction facility, mean scores (*SE*) were 0.41 (0.05) for England and 0.51 (0.08) for Malaysia.

3.3. Summary of results and discussion

First, the psychometric properties of the scales were good overall and for different subgroups according to the independent variables. Second, no evidence was found for *Hypothesis 1* or *Hypothesis 2*. Third, in support of *Hypothesis 3*, field dependence was associated with responding more to the secondary task. Fourth, the results do not allow *Hypothesis 4*, as stated, to be tested because this hypothesis was made conditional on a significant correlation between field dependence and national culture, but this was not significant. Nevertheless, the results show that English (Western) responded more frequently on the secondary task than Malaysian (East-asian) respondents, when there was no correction facility.

---

[11] As a follow-up, ANCOVA by correction facility showed that without correction facility, the effect of nation, F (1, 22) = 15.80, $\varepsilon^2 = 0.41$, $p < 0.01$, remained significant, but not the covariate, F (1, 22) = 1.03, $p > 0.05$. However, with correction facility, the effect of nation approached significance, F (1, 23) = 4.82, $\varepsilon^2 = 0.10$, $p = 0.053$, but the covariate was not significant, F (1, 23) = 1.16, $p > 0.05$.

[12] As a follow-up, ANCOVA by correction facility showed that without correction facility, the effects of the covariate, F (1, 17) = 4.58, $\varepsilon^2 = 0.09$, $p < 0.05$, and nation, F (1, 17) = 11.04, $\varepsilon^2 = 0.26$, $p < 0.01$, remained significant. However, with correction facility, the effect of the covariate remained significant, F (1, 18) = 7.92, $\varepsilon^2 = 0.19$, $p < 0.05$, but not nation, $F < 1$.

The results differ from those of previous research. In particular, the effect of questionnaire layout that was found by previous research (van Schaik and Ling [2007] and the current pilot experiment using a dual-tasking paradigm) was not confirmed here. A likely cause of the difference in results is that in the current experiment questionnaire layout was used with independent measures (presumably creating too much variation between participants to detect an effect of questionnaire layout), whereas previous research used repeated measures (eliminating variation among participants in tests of the effect of layout). Therefore, the effect of questionnaire layout may have remained undetected due to relatively large individual differences among participants. Consequently, it was decided to design and run Experiment 2, using a mixed design (as in the pilot experiment), with questionnaire layout manipulated within subjects.

4. Experiment 2

4.1. Method

4.1.1. Design

A 2×2×(2) experimental design was used. The first independent variable was rating-scale task condition (single or dual), the second was national culture (English [Western] or Malaysian [East-asian]) (both used with independent measures) and the third (within-subjects variable) was questionnaire layout (single item and whole form) (see Figure 1). Dependent variables were time to complete rating-scale items and secondary-task performance. Field dependence was measured as a potential covariate.

4.1.2. Participants

Ninety-five students from Teesside University, England, (n = 45; 23 female) and Universiti Putra Malaysia, Malaysia, (n = 50; 23 female) took part. Mean years of

experience with using Internet sites was 12.29 (*SD* = 3.54). Mean frequency per week of using Internet sites was 18.39 (*SD* = 16.44).

4.1.3. Materials and equipment

The experiment was programmed using Visual Basic 6.0. For the flow scale, questionnaire layout was the essentially the same as in the pilot experiment (see Figure 1). In the whole-form conditions, all seven disorientation items were presented on one page. In England, the experiment ran on personal computers (Intel Pentium, 1.86 GHz, 2 GB RAM, Microsoft Windows XP operating system, 17-inch monitors) with screen dimensions of 1280×1024. In Malaysia, computers of different configurations were used, similar to those used in England. Contrast and brightness were set to optimal levels. The same scales were used as in Experiment 1 to measure disorientation and flow.

4.1.4. Procedure

The experiment ran in a computer laboratory with groups of 15-20 participants who worked independently. They first completed the GEFT on paper. The remainder of the procedure was the same as in the pilot experiment, except that participants completed both the disorientation scale (first) and the flow scale (second) in each questionnaire layout with counterbalanced order of questionnaire layout. The mean number of Wikipedia pages visited was 16.80 (*SD* = 10.95). The experiment took about 50 minutes to complete.

4.2. Results

4.2.1. Psychometric analysis

The psychometric analysis of scales is presented in Online Appendix A2. The results showed that psychometric properties were good for each of the two questionnaire layouts. There were no statistically significant differences on AVE

between the two layouts. However, there were statistically significant differences on the loadings of some flow subscale items between the two layouts, in favour of single item (for Item CT4) and whole form (for Items SC2 and CT1).

4.2.2. Field dependence

Total scores on the GEFT were calculated per participant. Original scores were negatively skewed, with mean = 11.81, $CI(mean)_{0.95}$ = [10.79; 12.83]. Scores were square-root transformed as in Experiment 1. Participants from the two nations did not differ significantly on field dependence, t (93) = 1.54, $p > 0.05$.

4.2.3. Time to complete rating-scale items

For the disorientation scale, descriptives indicate that time to complete was shorter under the single-item conditions than under the whole-form conditions and under the single-tasking conditions than under the dual-tasking conditions (see Table 3). Mixed ANOVA confirmed that the effects of questionnaire layout, F (1, 91) = 12.48, $\varepsilon^2 = 0.08$, $p < 0.001$, and rating-scale task condition, F (1, 91) = 21.01, $\varepsilon^2 = 0.18$, $p < 0.001$, were significant, but not the effect of nation, $F < 1$.[13]

For the flow scale, again according to descriptives time was shorter under the single-item conditions (mean = 128.61, *SD* = 61.88) than under the whole-form conditions (mean = 162.34, *SD* = 64.90), and under the single-tasking conditions (mean = 136.46, *SD* = 53.04) than under the dual-tasking conditions (mean = 154.30, *SD* = 46.53). In addition, Malaysians completed the scale items quicker (mean = 136.03, *SD* = 50.78) than English participants (mean = 155.97, *SD* = 48.38). Mixed ANOVA confirmed that the effects of questionnaire layout, F (1, 91) = 19.87, $\varepsilon^2 = 0.07$, $p <$

[13] Time to complete was positively skewed and therefore for this analysis the distribution was improved by a logarithmic transformation for both the disorientation scale and the flow scale.

0.001, rating-scale task condition, F (1, 91) = 5.22, ε^2 = 0.02, p < 0.05, and nation, F (1, 91) = 8.95, ε^2 = 0.04, p < 0.01, were significant.

4.2.4. Secondary-task performance

The correlation between hits per false alarm for the disorientation scale with whole-form layout (performance on the secondary task) and GEFT scores was significant, r = -0.30, p < 0.05. Thus, those who possessed greater field dependence may have been more distracted by the secondary task with whole-form layout because the full screen was filled with information.

With whole-form layout, descriptives indicated there were more false alarms (untransformed mean [*SD*] = 0.22 [0.54]) per secondary-task stimulus) than with single-item layout (mean [*SD*] = 0.14 [0.27]) for the disorientation scale. Mixed ANOVA with independent variables questionnaire layout and nation confirmed this result for questionnaire layout, F (1, 93) = 5.60, ε^2 = 0.01, p < 0.05.[14]

The correlations between time to complete and secondary-task performance (false alarms) were 0.33, p < 0.05, for the disorientation scale with single-item layout, and 0.36, p < 0.05, for the flow scale with whole-form layout. Therefore, more distraction was associated with longer time-on-task.

4.3. Summary of results

The main findings were as follows. First, the psychometric properties of the scales were good overall and for each of the two questionnaire layouts. Second, in support of *Hypothesis 1*, time to complete was faster under single tasking. Third, single-item layout produced fewer false alarms than whole-form layout for the disorientation scale, providing evidence for *Hypothesis 2*. Fourth, in support of *Hypothesis 3*, the

[14] The distribution of false alarms was positively skewed and therefore for this analysis the distribution was improved by a square-root transformation.

correlation of hits per false alarm with field dependence was significant, but only for the disorientation scale with whole-form layout. Fifth, the results do not allow *Hypothesis 4*, as stated, to be tested because this hypothesis was made conditional on a significant correlation between field dependence and national culture, but this was not significant. In any case, the effect of national culture on responding to the secondary task was not significant. To further test the effect of questionnaire layout and the sensitivity of psychometric measurement, Experiment 3 used different types of task and website.

5. **Experiment 3**

5.1. Method

5.1.1. Design

A 2×2×(2) mixed experimental design was used. The independent variables were task complexity (high or low), rating-scale task condition (single or dual; see Figure 3) (both used with independent measures) and questionnaire layout (within subjects; single item and whole form; see Figure 3), in order to allow sensitivity of psychometric measurement to be tested. In a low-complexity task the answer to an information-retrieval task was available on a page two links from the homepage, but in a high-complexity task the answer was available on a page four links from the homepage. Dependent variables were time to complete rating-scale items and secondary-task performance.

5.1.2. Participants

Students from Teesside University, England, (*N* = 113; 95 female) took part. Mean years of experience with using Internet sites was 10.49 (*SD* = 4.50). Mean frequency per week of using Internet sites was 19.15 (*SD* = 10.78).

5.1.3. Materials and equipment

The experiment was programmed as a Visual Basic form application, created with Visual Studio 2008, and ran on personal computers (Intel Pentium, 1.86 GHz, 2 GB RAM, Microsoft Windows XP operating system, 17-inch monitors). In the whole-form conditions, for the flow scale 9 items were presented on the first page, 10 on the second and 11 on the third (see Figure 3). A website was modelled as a typical psychology site for university students, and specially designed for the experiment. In the whole-form conditions, all seven disorientation items were presented on one page. The screen dimensions were 1280×1024. Contrast and brightness were set to optimal levels. Four items from Guay et al.'s (2000) 16-item Situational Motivation Scale (SIMS) measured intrinsic motivation (see Appendix), with a 7-point Likert scale. The same scales were used as in Experiment 1 to measure disorientation and flow.

5.1.4. Procedure

The experiment ran in a computer laboratory with groups of 15-20 participants who worked independently. In *Phase 1*, they completed the SIMS. In *Phase 2*, they viewed screenshots of five web pages of the intranet site. In *Phase 3*, an interactive information retrieval task followed which included typical tasks that users perform with educational intranet sites. After three practice questions, the main set of information retrieval tasks followed, with a duration of 20 minutes. The mean number of web pages visited was 142.69 (*SD* = 42.15). In *Phase 4*, participants completed both first the disorientation scale and then the flow scale in each questionnaire layout, the order of which was counterbalanced. Finally, participants answered questions requesting demographic details. The experiment took about 40 minutes to complete.

5.2. Results

5.2.1. Psychometric analysis

The psychometric analysis of scales is presented in Online Appendix A3. The results showed that psychometric properties were good for each of the two questionnaire layouts. There were no statistically significant differences on AVE between the two layouts. However, there were statistically significant differences on the loadings of some flow subscale items between the two layouts (on Item TS2, in favour of single item, and Item CT4, in favour of whole form).

5.2.2. Manipulation check of task complexity

According to descriptives, task performance was more accurate (in terms of proportion of correct answers over completed tasks) with simple tasks (untransformed mean = 0.84; SD = 0.12) than with complex tasks (mean = 0.62; SD = 0.25). A between-subjects *t* test confirmed this result, t (111) = 6.11, r = 0.50, p < 0.001.[15]

5.2.3. Time to complete rating-scale items

Descriptives indicated that time was shorter under single-item layout than under whole-form layout and under single-tasking conditions than under dual-tasking conditions (see Table 4). Mixed ANOVA confirmed that for disorientation the effects of questionnaire layout, F (1, 109) = 20.97, ε^2 = 0.06, p < 0.001, and rating-scale task condition, F (1, 109) = 14.07, ε^2 = 0.07, p < 0.001, were significant.[16] Further mixed ANOVA confirmed that for flow experience the effects of questionnaire layout, F (1,

---

[15] Correctness of answers was negatively skewed and therefore for this analysis the distribution was improved by reflection followed by an inverse transformation.

[16] Time to complete was positively skewed and therefore this and the following analysis the distribution was improved by logarithmic transformation.

109) = 40.71, ε^2 = 0.11, p < 0.001, and rating-scale task condition, F (1, 109) = 12.17, ε^2 = 0.05, p < 0.001, were significant. The effect of task complexity and the interaction effects were not significant for either scale.

5.2.4. Secondary-task performance

Mixed 2×(2) ANOVA showed that the effect of questionnaire layout on misses for the flow scale was significant, F (1, 49) = 4.73, ε^2 = 0.03, p < 0.05.[17] There were fewer misses in the single-item conditions (untransformed mean = 0.09, SD = 0.14) than in the whole-form conditions (mean = 0.11, SD = 0.15). The effect of task complexity, F (1, 49) = 1.96, p > 0.05, and the interaction effect, F < 1, were not significant. The correlation between time to complete and secondary-task performance (misses) was significant, r = 0.31, p < 0.05, for the disorientation scale with single-item layout, but not for the flow scale, r = 0.31, p > 0.05.

5.2.5. Criterion validity

Criterion validity of psychometric scales can be established by calculating correlations between scales (see results in Online Appendix A3), but also by calculating correlations between scales and other measures of the quality of interaction, such as measures of participants' task performance and navigation behaviour in the information retrieval task before they completed the rating scales. Task performance was measured in terms of speed (time-on-task over completed web navigation tasks) and accuracy (percentage of correct answers over completed web navigation tasks), and web navigation behaviour as number of pages visited

[17] The number of misses per secondary-task stimulus was positively skewed and therefore for this analysis the distribution was improved by square-root transformation.

(over completed tasks) (van Schaik & Ling, 2007).[18] The following results (see Table 5) provide further evidence for criterion validity. The correlations of page visits and time-on-task with disorientation and feedback were significant for both questionnaire layouts. The correlation of situational motivation with autotelic experience and transformation of time was significant for both questionnaire layouts. The correlation of accuracy with all (sub)scales, except transcendence of self, was significant for both questionnaire layouts.

5.2.6. Sensitivity of psychometric scales

Descriptives indicated that, when tasks were more complex, disorientation increased, and levels of flow dimensions decreased (see Table 6). The results of 2×2×(2) mixed ANOVA, with independent variables task complexity, questionnaire layout and rating-scale task condition, confirmed these results. The effect of task complexity was significant on disorientation, F (1, 109) = 27.60, $\varepsilon^2 = 0.18$, $p < 0.001$; balance of challenge and skills, F (1, 109) = 15.21, $\varepsilon^2 = 0.10$, $p < 0.001$; control, F (1, 109) = 3.92, $\varepsilon^2 = 0.02$, $p = 0.05$; feedback, F (1, 109) = 7.81, $\varepsilon^2 = 0.05$, $p < 0.01$; goal clarity, F (1, 109) = 7.55, $\varepsilon^2 = 0.05$, $p < 0.01$; and mergence of action and awareness, F (1, 109) = 13.41, $\varepsilon^2 = 0.08$, $p < 0.001$. Furthermore, the effect of questionnaire layout on feedback was significant, F (1, 109) = 5.91, $\varepsilon^2 = 0.002$, $p < 0.05$; with whole-form layout, feedback was higher (mean = 4.13 and SD = 1.45 for single item, and mean = 4.27 and SD = 1.39 for whole form).

[18] Time-on-information-retrieval-task and page visits were positively skewed and therefore for statistical inferential analysis the distributions were improved by a logarithmic transformation.

5.3. Summary of results

First, the psychometric properties of the scales were good overall and for each of the two questionnaire layouts. Second, time to complete was faster with single tasking supporting *Hypothesis 1*. Third, single-item layout produced fewer misses than whole-form layout for the flow scale, providing partial evidence for *Hypothesis 2*. Fourth, there was evidence for criterion validity: the same specific pattern of correlations of disorientation and flow subscales with web navigation behaviour, accuracy and speed of web navigation, and situational motivation occurred for both questionnaire layouts. Fifth, disorientation and five flow subscales were sensitive to the manipulation of task complexity, irrespective of questionnaire layout.

6. General discussion

This section first discusses the results from the experiments in relation to each of the hypotheses. Directions for future research are then presented.

6.1. Exploration of main findings

The psychometric properties of the disorientation scale (Ahuja & Webster, 2001) and the flow scale (Guo & Poole, 2009) for web navigation were good overall, for different subgroups and for both questionnaire layouts. The finding of the robustness of established scales to measure disorientation and flow in web navigation across experimental manipulations is consistent with the findings of Davis and Venkatesh's landmark (1996) study that their technology-acceptance scales were robust in the face of experimental manipulations, with psychometric quality unaffected. Psychometrics have also been found robust to variations in control over the test environment (Templer & Lange, 2008). Nevertheless, in Experiment 1, using an independent measures design, significantly less variance was extracted by the latent variable for the Flow Subscale Concentration with whole-form layout than with

single-item layout, indicating an advantage for single item. Furthermore, significantly less variance was extracted by the latent variables for the Flow Subscales Concentration and Mergence of Action and Awareness in Malaysians than in English participants.

Hypothesis 1: time to complete psychometric items is longer under whole-form layout than under single-item layout. Three of the four experiments found converging evidence for *Hypothesis 1*, when questionnaire layout was a within-subject factor, with medium to large effect size (in the pilot experiment only under dual tasking). These results indicate that respondents' attention may indeed inadvertently switch back and forth between the next question to be answered and simultaneously presented previous and following items. The findings in the pilot experiment indicate that this disadvantage of whole form could be stronger when more attention is needed for (the primary task of) responding to online psychometrics. The results of Experiment 1 do not support *Hypothesis 1*. This is likely due to the use of an independent measures design, presumably creating too much variation between participants to detect an effect of questionnaire layout. Presumably the other experiments increased their sensitivity to the manipulation of questionnaire layout because they employed this independent variable within subjects. The results from the three mixed-measures experiments reported here together with previous results (van Schaik & Ling, 2007), now provide evidence that, across user-experience ratings of different websites, different scales to measure different aspects of the quality of web navigation and different types of task (exploration and information retrieval), single item is advantageous in terms of faster completion of online psychometrics.

Significant positive correlations between secondary-task performance and time to complete questionnaire found in Experiments 2 and 3 indicate a speed-attention trade-off, as with faster completion of questions performance was worse. Therefore, more attention on the secondary required more time to complete the primary task of questionnaire completion. However, these results do not provide a consistent account for the advantage of single-item layout in terms of time to complete. This is because the positive correlation between completion time and secondary-task performance did not occur consistently across experiments and performance measures.[19]

Hypothesis 2: secondary-task performance is better with single-item layout than with whole-form layout. Using questionnaire layout as a within-subject factor, three of the four experiments found partial evidence for *Hypothesis 2*, with small to large effect sizes. The results indicate that whole-form layout divides attention and thereby requires more attention to answer psychometric items. These results were not uniform, but depended on measures of secondary task performance (false alarms in the pilot experiment and Experiment 2; misses in Experiment 3) and scales (flow scale in the pilot experiment and Experiment 3; disorientation scale in Experiment 2) that showed an effect. The results of Experiment 1 do not support *Hypothesis 2*. Again, this could be due to a lack of sensitivity because an independent-measures design was used. With the current questionnaires and layouts, we found partial evidence for the idea that whole-form layout requires more attention to respond to psychometric items. Another way of thinking about these results is that although secondary-task performance was not uniformly degraded by layout, the addition of

[19] If the correlation was consistent then mediation analysis would establish if secondary-task performance was a mediator of the effect of questionnaire layout on questionnaire completion time.

the secondary-task consistently increased completion time in three out of four experiments: with secondary task added, time with whole form exceeded that with single item. As participants were not forced to keep any particular pace on the primary task of questionnaire completion, they could perhaps at least partially compensate for the additional task load by slowing down. Moreover, positive correlations between time to complete and secondary-task performance (Experiments 1, 2 and 3) indicate that attention is potentially an important factor in questionnaire completion. In particular, with whole-form layout less variance was extracted for the Flow Subscale Concentration in Experiment 1 and in Malaysian participants less variance was extracted for the Flow Subscales Concentration and Mergence of Action and Awareness.

Hypothesis 3: field dependence reduces susceptibility to distraction by secondary-task performance. Two out of two experiments (Experiments 1 and 2) found evidence for *Hypothesis 3*, with medium to medium-to-large effect size. The results indicate that, because questions are processed together holistically rather than in isolation analytically, field dependence divides attention between the next question to be answered and other items on the screen in online psychometric items. In Experiment 1, evidence for the hypothesis was found for both scales (disorientation and flow), but in Experiment 2 only for the disorientation scale with whole-form layout. Given the lack of uniformity of results in Experiment 2, the positive evidence for *Hypothesis 3* should be treated with caution at this stage.

Hypothesis 4: national culture is a determinant of secondary-task performance. As stated, the hypothesis was conditional on a correlation between field dependence and national culture (or difference on field dependence between cultures). Although previous research has demonstrated a significant difference (Nisbett et al., 2001), it

has been noted that in people who have a bi-cultural background this difference may be reduced (Nisbett & Miyamoto, 2005) or perhaps even eliminated. As a result of experiencing higher education and being exposed to Western culture, our highly educated Malaysian sample may have become less 'mono-cultural' in terms of information-processsing style, resulting in a non-significant difference. Moreover, another reason for a lack of difference may be the measurement of field dependence. Not all instruments measuring field dependence are equally sensitive; in particular, instruments other than the GEFT may be more sensitive (Meys & Sanderson, 2013; see also Section 6.2).

The striking difference in results between Experiment 1 (using a fully independent-measures design) and the other experiments and previous research (van Schaik & Ling, 2007) (manipulating independent variables with repeated measures) may have wider implications. Because of individual differences, perhaps design parameters, such as questionnaire layout, may be best manipulated within subjects to demonstrate the effects of parameters.

6.2. Future research

6.2.1. Questionnaire layout

Given the robust effect of questionnaire layout on time to complete in online psychometrics, the question arises to what extent these results apply the psychometrics more generally, irrespective of visual presentation medium (pencil-and-paper, desktop, laptop, tablet, smartphone). Because of the increased cost of printing items in single-item layout in the use of paper questionnaires, this layout has normally not been considered, irrespective of its advantages. In fact, given this reluctance, the advantages of single-item presentation (mainly reduced time) could be another reason for moving from pencil-and-paper- to online questionnaires.

However, many psychometric questionnaires are still being presented on computer screens with whole-form layout, although the current studies and previous research (van Schaik & Ling, 2007) indicate that single-item layout is advantageous. Across presentation media, the smaller the device, the smaller the difference in presentation between single item and whole form. This is because, even though smaller font sizes are used on smartphones, with a small screen still only few lines could be displayed previously (see also Peytchev & Hill, 2010). Therefore, the difference between single item and whole form might be reduced in terms of attention inadvertently switching back and forth between the next question to be answered and simultaneously presented previous and following items. Indeed, Peytchev and Hill (2010) found no effect of number of items per screen page (1 or 10) in responses to a factual survey rather than a psychometric questionnaire. However, in their design of the 10 items on a page only 1 was visible at any one time, so participants had to scroll through the items. Therefore, effectively it was navigation mode rather than questionnaire layout that was manipulated. Nevertheless, smartphones with larger screens have recently been introduced, allowing for a substantial difference between single item and whole form in terms of the number of items presented per screen and consequently speed in questionnaire completion. A progress indicator may then help respondents in particular on smartphones (but also on other devices) realise how fast they are progressing through an online psychometric questionnaire, thereby addressing the reported disadvantage of uncertainty about questionnaire length, but not necessarily completion rate (Conrad, Couper, Tourangeau & Peytchev, 2010). More generally, designs for different platforms (personal computers, tablet, smartphone) require different user-interface designs (Nielsen, 2012) and this would apply equally well to the design of online psychometrics.

Another consideration is that smaller devices (tablets and smartphones) use touch for interaction rather than keyboard and mouse used by larger devices (desktop- and laptop computers). One issue with the use of touch screens for online psychometrics could be a lack of precision and more time required in responding (and difficulty with correcting responses if this is allowed), in particular when small response alternatives are placed close together. Although there is some work on factual online surveys with small devices (Peytchev & Hill, 2010), there appears to be a lack of solid empirical research investigating online psychometrics on devices smaller then desktops to produce empirical results and design implications based on these. In addition, as the size of visual displays and their resolution of desktop computers are increasing, research should also address online psychometrics on such displays. The difference in completion time between single item and whole form that we found here would be expected to be even bigger with large screens. Moreover, line length would be an even more important issue than on smaller screens (Ling & van Schaik, 2006).

The distinction between single item and whole form disappears when single-item scales are used to measure experience. These scales can be particularly useful when momentary experience (affect during human-computer interaction) is to be measured (e.g. Broekens & Brinkman, 2013). How momentary experience influences remembered experience (after interaction) is an important research question, as research in other domains (e.g. Wirtz et al., 2003) has demonstrated that remembered experience rather than momentary experience drive subsequent behaviour.

6.2.2. Field dependence and national culture

Irrespective of a link between field dependence and secondary-task performance in responding to online psychometrics, Experiment 1 found evidence for an association between field dependence and disorientation. With increasing field dependence, participants in the experiments were more disoriented in navigation a website. Thus, more generally field-dependent Internet- and intranet users may be more likely to experience disorientation in navigating sites. This makes it even more important to create websites with good information architectures that reduce disorientation (Muzahir, 2013). A question for future research is then whether field-dependent and field-independent users benefit equally from improved information architectures.

No evidence was found for a difference in field dependence between English and East-asian participants. In future research, a more sensitive – but also bulky and costly – test of field dependence such as the Rod-and-Frame Test (Nisbett et al., 2001), may uncover such a difference, and a more consistent link of field dependence with secondary-task performance, psychometric quality and disorientation in web navigation. Therefore, such research into online psychometrics and web navigation may measure field dependence with a less bulky and more convenient virtual-reality version of the Rod-and-Frame Test (Reger et al., 2003) instead of the GEFT. Moreover, samples with different demographics (e.g. not highly educated) which differ more strongly on national culture (Western versus East-asian) may not only provide evidence for a difference in field dependence, but also for national culture as a mediator of the effect of field dependence on secondary-task performance.

7. Conclusion

We found that the flow- and the disorientation scales for web navigation used here have good psychometric quality overall, for different subgroups and for two questionnaire layouts (single item and whole form). The findings should give researchers and practitioners increased confidence in using these scales. As questionnaire completion was faster with single-item layout, this should be used to administer psychometric questionnaire for the purpose of reducing completion time and facilitating attention in questionnaire completion.

Acknowledgement

The authors are grateful to the British Academy (award number 30105028) for funding the research that is reported in this publication, and to Gabor Aranyi and Raza Habib Muzahir for assistance with data collection.

References

Ahuja, J.S., Webster, J., 2001. Perceived disorientation: An examination of a new measure to assess web design effectiveness. Interact Comput 14, 15-29.

Broekens, J., Brinkman, W.-P., 2013. AffectButton: A method for reliable and valid affective self-report. Int J Hum Comput Stud 71, 641-667.

Card, S., Moran, T., Newell, A., 1983. The Psychology of Human-Computer Interactin. Erlbaum, Hillsdale, NJ.

Conrad, F.G., Couper, M.P., Tourangeau, R., Peytchev, A., 2010. The impact of progress indicators on task completion. Interact Comput 22, 417-427.

Davis, F.D., Venkatesh, V., 1996. A critical assessment of potential measurement biases in the technology acceptance model: Three experiments. International Journal of Human Computer Studies 45, 19-45.

Evans, R.R., Burnett, D.O., Kendrick, O.W., Macrina, D.M., Snyder, S.W., Roy, J.P.L., Stephens, B.C., 2009. Developing valid and reliable online survey instruments using commercial software programs. J. Consum. Health Internet 13, 42-52.

Guay, F., Vallerand, R.J., Blanchard, C., 2000. On the assessment of situational intrinsic and extrinsic motivation: The Situational Motivation Scale (SIMS). Motiv. Emotion 24, 175-213.

Guo, Y.M., Poole, M.S., 2009. Antecedents of flow in online shopping: a test of alternative models. Information Systems Journal 19, 369-390.

Jaccard, J., 1998. Interaction Effects in Factorial Analysis of Variance. Sage, Thousand Oaks.

Ling, J., Schaik, P. van, 2006. The influence of font type and line length on visual search and information retrieval in web pages. Int J Hum Comput Stud 64, 395-404.

Liu, X., 2011. Assessing Measurement Invariance of the Teachers' Perceptions of Grading Practices Scale across Cultures. International Journal of Humanities and Social Science 1, 70-80.

Marí-Beffa, P., Houghton, G., Estévez, A.F., Fuentes, L.J., 2000. Word-based grouping affects the prime-task effect on semantic priming. Journal of Experimental Psychology: Human Perception and Performance 26, 469-479.

Meys, H.L., Sanderson, P.M., 2013. The effect of individual differences on how people handle interruptions. Proc Hum Factors Ergon Soc , 868-872.

Miyake, A., Witzki, A.H., Emerson, M.J., 2001. Field dependence-independence from a working memory perspective: A dual-task investigation of the Hidden Figures Test. Memory 9, 445-457.

Moore, D.W., 2002. Measuring new types of question-order effects: Additive and subtractive. Public Opin. Q. 66, 80-91.

Muzahir, R.H., 2013. Computational Cognitive Modelling of Information Architecture and Web Navigation. Unpublished PhD Thesis. Teesside University, Middlesbrough.

Ndosi, M., Tennant, A., Bergsten, U., Kukkurainen, M.L., Machado, P., De La Torre-Aboki, J., Vliet Vlieland, T.P.M., Zangi, H.A., Hill, J., 2011. Cross-cultural validation of the Educational Needs Assessment Tool in RA in 7 European countries. BMC Musculoskelet. Disord. 12.

Nielsen, J., 2012. Transmedia design for the 3 screens (make that 5).

Nisbett, R., Miyamoto, Y., 2005. The influence of culture: holistic versus analytic perception. Trends in cognitive sciences 9, 467-473.

Nisbett, R.E., Choi, I., Peng, K., Norenzayan, A., 2001. Culture and systems of thought: Holistic versus analytic cognition. Psychol. Rev. 108, 291-310.

Norman, K.L., Friedman, Z., Norman, K., Stevenson, R., 2001. Navigational issues in the design of online self-administered questionnaires. Behav. Inf. Technol. 20, 37-45.

Peytchev, A., Hill, C.A., 2010. Experiments in mobile web survey design: Similarities to other modes and unique considerations. Soc. Sci. Comput. Rev. 28, 319-335.Ray, C., 1974. Some components of colour-word interference and their relationship with field dependence. Acta Psychol. 38, 323-330.

Reger, G.M., McGee, J.S., Van der Zaag, C., Thiebaux, M., Buckwalter, J.G., Rizzo, A.A., 2003. A 3D Virtual Environment Rod and Frame Test: The Reliability and Validity of Four Traditional Scoring Methods for Older Adults. J. Clin. Exp. Neuropsychol. 25, 1169-1177.

Schaik, P. van, Ling, J., 2003. Using online surveys to measure three key constructs of the quality of human-computer interaction in web sites: psychometric properties and implications. International Journal of Human-Computer Studies 59, 545-567.

Schaik, P. van, Ling, J., 2005a. Five psychometric scales for online measurement of the quality of human-computer interaction in Web Sites. Int. J. Hum. -Comput. Interact. 18, 309-322.

Schaik, P. van, Ling, J., 2005b. The psychometric evaluation of educational intranets. Journal of Educational Computing Research 33, 81-100.

Schaik, P. van, Ling, J., 2007. Design parameters of rating scales for web sites. ACM Transactions on Computer-Human Interaction 14, 4.

Schaik, P. van, Ling, J., 2012a. An experimental analysis of experiential and cognitive variables in web navigation. Hum Comput Interact 27, 199-234.

Schaik, P. van, Ling, J., 2012b. A cognitive-experiential approach to modelling web navigation. Int J Hum Comput Stud 70, 630-651.

Schwarz, N., 1999. How the questions shape the answers. Am. Psychol. 54, 93-105.

Schwarz, N., Strack, F., Mai, H.-P., 1991. Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. Public Opin. Q. 55, 3-23.Spenkelink, G., Besuijen, K., 1999. Visual display quality , in Kent, A., Hall, C. (Eds.), Encyclopedia of Library and Information Science, Volume 64. New York: Dekker, pp. 330-357.

Templer, K.J., Lange, S.R., 2008. Internet testing: Equivalence between proctored lab and unproctored field conditions. Comput. Hum. Behav. 24, 1216-1228.

Vlachopoulos, S.P., Asci, F.H., Cid, L., Ersoz, G., González-Cutre, D., Moreno-Murcia, J.A., Moutão, J., 2013. Cross-cultural invariance of the basic psychological needs in exercise scale and need satisfaction latent mean differences among Greek, Spanish, Portuguese and Turkish samples. Psychol. Sport Exerc. 14, 622-631.

Wickens, C., McCarley, J., 2008. Applied Attention Theory. Taylor & Francis, Boca Raton.

Wirtz, D., Kruger, J., Scollon, C.N., Diener, E., 2003. What to do on Spring Break? The Role of Predicted, on-Line, and Remembered Experience in Future Choice. Psychol. Sci. 14, 520-524.

Witkin, H., Oltman, P., Rasking, E., Karp, S., 1971/2002. Group embedded figures test.

Wulf, G., Lewthwaite, R., 2010. Effortless motor learning? An external focus of attention enhances movement effectiveness and efficiency, in Bruya, B. (Ed.), Effortless Attention. MIT Press, Cambridge, MA, pp. 75-101.

Appendix – psychometric questionnaires

Intrinsic motivation (Guay et al., 2000)

Why are you currently engaged in this activity?

IM1	Because I think that this activity is interesting.
IM2	Because I think that this activity is pleasant.
IM3	Because this activity is fun.
IM4	Because I feel good when doing this activity.

Response format: 7-point Likert scale with endpoints ‘Corresponds not all’ and ‘Corresponds exactly’.

Disorientation (Ahuja & Webster, 2001)

DIS1	I felt lost.
DIS2	I felt like I was going around in circles.
DIS3	It was difficult to find a page that I had previously viewed.
DIS4	Navigating between pages was a problem.
DIS5	I didn't know how to get to my desired location.
DIS6	I felt disoriented.
DIS7	After browsing for a while I had no idea where to go next.

Response format: 7-point Likert scale with endpoints ‘never’ and ‘always’.

Flow (Guo & Poole, 2009)

Balance of challenge and skill

CS1	My abilities matched the challenge of the situation.
CS2	I felt I was competent enough to meet the demands of the situation.

Clarity of goals

CG1	I knew clearly what I wanted to do.
CG2	I had a strong sense of what I wanted to do.

CG3	I knew what I wanted to achieve.
CG4	My goals were clearly defined.
Feedback	
F1	It was really clear to me that I was doing well.
F2	I was aware of how well I was performing.
F3	When using the website, I had a good idea about how well I was doing.
F4	I could tell by the way I was using the website how well I was doing.
Concentration	
CN1	My attention was focused entirely on what I was doing.
CN2	It was no effort to keep my mind on what was happening.
CN3	I had total concentration.
CN4	I was completely focused on the task at hand.
Control	
CT1	I felt in total control of what I was doing.
CT2	I felt like I could control what I was doing.
CT3	I had a feeling of total control.
CT4	I felt in total control of my action.
Mergence of action and awareness	
M1	I reacted to the website automatically.
M2	I did things spontaneously and automatically without having to think.
Transformation of time	
TT1	Time appeared to go by very quickly.
TT2	I lost track of time.
TT3	Time flew.

Loss of self-consciousness

TS1	I was not concerned with what others may have been thinking of me.
TS2	I was not concerned with how I was presenting myself.
TS3	I was not worried about what others may have been thinking of me.

Autotelic experience

AE1	I really enjoyed the experience.
AE2	I loved the feeling experienced and I want to capture it again.
AE3	The experience left me feeling great.
AE4	I found the experience extremely rewarding.

Response format: 7-point Likert scale with endpoints ‘Not challenging at all’ and ‘Very challenging’.

Table 1

Descriptives for time to complete (pilot experiment)

		Original (time[a], all items)		Transformed (log(time))		Retransformed (exp[log(time)])	
Type of tasking		Single item	Whole form	Single item	Whole form	Single item	Whole form
Single tasking	Mean	124.09	96.29	11.29	11.16	79.73	69.97
	SD	110.21	75.85	1.07	0.88		
Dual tasking	Mean	137.89	189.36	**11.37**	**11.79**	86.46	131.97
	SD	226.81	254.09	0.82	0.73		

Note. Flow scale. Planned comparisons on transformed scores (single item versus whole form): bold text: p (t, 2-tailed) < 0.01.
[a]Seconds.

Table 2

Descriptives for time to complete (Experiment 1)

		Disorientation					
		Original (time[a]/item)		Transformed (item/time)		Retransformed (1/(item/time))	
Type of tasking		Single item	Whole form	Single item	Whole form	Single item	Whole form
Single tasking	Mean	8.09	7.49	0.15	0.16	6.66	6.26
	SD	6.43	6.25	0.07	0.08		
Dual tasking	Mean	6.76	9.31	0.15	0.13	6.82	7.57
	SD	2.95	7.97	0.05	0.06		
		Flow					
		Original (time/item)		Transformed (item/time)		Retransformed (1/(item/time))	
Type of tasking		Single item	Whole form	Single item	Whole form	Single item	Whole form
Single tasking	Mean	7.83	8.24	0.15	0.14	6.50	7.32
	SD	9.12	6.44	0.06	0.06		
Dual tasking	Mean	7.75	9.33	0.14	0.13	7.06	7.77
	SD	4.88	7.43	0.06	0.06		

Note. Planned comparisons on transformed scores (single item versus whole form): no significant differences..

[a]Seconds.

Table 3

Descriptives for time to complete (Experiment 2)

		Disorientation					
		Original (time[a], all items)		Transformed (log(time))		Retransformed (exp[log(time)])	
Type of tasking		Single item	Whole form	Single item	Whole form	Single item	Whole form
Single tasking	Mean	43.92	50.63	*10.53*	*10.74*	37.30	46.05
	SD	29.96	23.19	0.55	0.44		
Dual tasking	Mean	56.37	71.30	**10.84**	**11.08**	50.82	65.01
	SD	30.78	31.98	0.45	0.43		
		Flow					
		Original (time, all items)		Transformed (log(time))		Retransformed (exp[log(time)])	
Type of tasking		Single item	Whole form	Single item	Whole form	Single item	Whole form
Single tasking	Mean	120.00	152.92	**11.52**	**11.86**	100.78	141.74
	SD	70.43	59.97	0.65	0.40		
Dual tasking	Mean	137.04	171.56	**11.76**	**11.97**	128.16	158.13
	SD	51.56	68.78	0.37	0.43		

Note. Planned comparisons on transformed scores (single item versus whole form): italicized text: p (t, 2-tailed) < 0.05; bold text: p (t, 2-tailed) < 0.01.

[a]Seconds.

Table 4
Descriptives for time to complete (Experiment 3)

		Disorientation					
		Original (time[a], all items)		Transformed (log(time))		Retransformed (exp[log(time)])	
Type of tasking		Single item	Whole form	Single item	Whole form	Single item	Whole form
Single tasking	Mean	29.28	38.99	***3.37***	***3.60***	29.16	36.78
	SD	8.66	16.99	0.28	0.42		
Dual tasking	Mean	38.04	44.65	<u>3.62</u>	<u>3.75</u>	37.49	42.66
	SD	11.24	17.71	0.29	0.37		

		Flow					
		Original (time, all items)		Transformed (log(time))		Retransformed (exp[log(time)])	
Type of tasking		Single item	Whole form	Single item	Whole form	Single item	Whole form
Single tasking	Mean	101.34	122.43	***4.58***	***4.77***	97.14	117.80
	SD	37.39	38.07	0.31	0.31		
Dual tasking	Mean	113.73	149.46	***4.70***	***4.97***	110.36	143.40
	SD	31.65	47.04	0.29	0.32		

Note. Planned comparisons on transformed scores (single item versus whole form): underlined text: p (t, 2-tailed) < 0.10; italicized bold text: p (t, 2-tailed) < 0.001.
[a]Seconds.

Table 5
Correlations between (sub)scales and performance on information retrieval task (Experiment 3)

	Single item				Whole form			
	Page visits	Accuracy	Time-on-task	Situational motivation	Page visits	Accuracy	Time-on-task	Situational motivation
DIS	***0.39	***-0.43	***0.34	-0.09	***0.33	***-0.53	***0.32	-0.07
CS	-0.10	***0.49	-0.10	0.04	-0.13	***0.50	-0.19	-0.09
FB	*-0.24	***0.36	*-0.18	0.05	*-0.22	***0.33	*-0.19	-0.04
CG	0.01	***0.38	-0.01	0.09	-0.17	***0.35	-0.18	0.01
CN	0.05	**0.29	0.13	0.16	0.04	**0.26	0.06	0.06
CT	-0.09	***0.41	-0.10	0.07	-0.08	***0.40	-0.08	0.11
M	-0.16	***0.41	-0.16	0.11	-0.09	***0.42	-0.17	0.14
TS	0.01	0.15	0.06	-0.11	-0.04	0.18	-0.03	-0.03
TT	0.09	**0.30	0.16	**0.28	0.05	**0.25	0.10	**0.31
AE	-0.07	**0.28	0.07	***0.56	-0.04	**0.20	0.11	***0.55

Note. AE: autotelic experience. CS: balance of challenge and skill. CN: concentration. CT: control. FB: feedback. CG: clarity of goals. MA: mergence of action and awareness. TS: transcendence of self. TT: transformation of time. DIS: disorientation.
*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

Table 6
Descriptives for disorientation scale and flow subscales as a function of task complexity (Experiment 3)

(Sub)scale	Task complexity			
	Simple		Complex	
	Mean	*SD*	Mean	*SD*
Disorientation***	2.77	1.11	3.84	0.99
Autotelic experience	2.56	1.22	2.47	1.18
Balance of challenge and skill***	5.57	0.96	4.74	1.25
Concentration	4.76	1.43	4.60	1.33
Control*	5.05	1.38	4.51	1.33
Feedback**	4.55	1.31	3.82	1.37
Clarity of goals**	5.27	1.05	4.68	1.19
Mergence of action and awareness***	5.04	1.09	4.21	1.25
Transcendence of self	5.34	1.42	5.08	1.20
Transformation of time	3.70	1.62	3.54	1.27

Note. The statistical significance of tested differences in task complexity is presented with probability levels.
* $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

Online Appendix A0: psychometric analysis (pilot experiment)

Given the relatively small sample in the pilot experiment, psychometric analysis was restricted to internal-consistency reliability, and convergent and discriminant validity. Cronbach's coefficient alpha ranged for the flow subscales from 0.80 to 0.98 for each of the two questionnaire layouts, exceeding the cut-off points, so the subscales were deemed to be reliable. Correlations between single-item and whole-form layouts were analysed per subscale (see Table A0.1). Correlations between the questionnaire layouts for the same subscale were higher (most equal to or exceeding 0.70) than correlations of a subscale with any other subscale under the other questionnaire layout, with only 3 out of 64 exceptions, therefore providing evidence for convergent (high correlations between formats for the same [sub]scales) and discriminant validity (lower correlations in other pairs of subscale-questionnaire layout combinations).

Table A0.1
Correlations among subscales between questionnaire layouts

	CS-SI	GC-SI	FB-SI	CN-SI	CT-SI	MA-SI	TT-SI	TS-SI	AE-SI
CS-SI	0.88	0.73	-0.01	0.33	0.36	0.01	0.00	0.47	-0.09
GC-SI	0.64	0.90	0.42	0.59	0.65	0.27	0.16	0.27	0.04
FB-SI	0.20	0.44	0.77	0.57	0.24	0.75	0.78	0.16	0.47
CN-SI	0.46	0.66	0.60	0.77	0.45	0.49	0.52	0.43	0.45
CT-SI	0.62	0.75	0.34	0.61	0.70	0.36	0.24	0.32	0.01
MA-SI	-0.02	0.16	0.61	0.43	0.12	0.66	0.81	0.11	0.45
TT-SI	-0.06	0.05	0.59	0.45	-0.02	0.64	0.87	0.02	0.55
TS-SI	0.31	0.37	0.27	0.29	0.16	0.23	0.07	0.79	0.15
AE-SI	-0.23	0.10	0.58	0.45	-0.08	0.58	0.72	0.15	0.90

Note. AE: autotelic experience. CS: balance of challenge and skill. CN: concentration. CT: control. FB: feedback. CG: clarity of goals. MA: mergence of action and awareness. TS: transcendence of self. TT: transformation of time. SI: single item. WF: whole form.

Online Appendix A1: psychometric analysis (Experiment 1)

Partial-least-squares path modelling (PLS for short; Vinzi et al., 2010) was used for psychometric data analysis using SmartPLS (Ringle et al., 2005). PLS does not require some of the assumptions imposed by covariance-based structural equation modelling - including those of large sample size, and univariate and multivariate normality. Recent simulation studies have demonstrated that PLS path modelling performs at least as well as and, under various conditions, is superior to covariance-based structural equation modelling in terms of bias, root mean square error and mean absolute deviation (Hulland et al., 2010; Vilares et al., 2010). In the analyses, a standard PLS bootstrapping procedure (N = 5000, recommended by Henseler et al., 2009) was used to test the significance of model parameters.

In testing the measurement model, reliability was analysed (see Table A1.1), and convergent and discriminant validity was assessed (see Table A1.2). The reliability of each individual reflective item is assessed by its loading on the construct of which it is an indicator, which should be 0.7 or higher (Henseler et al., 2009).

Using a bootstrapping procedure, the loadings of all items of the flow scale and the disorientation scale were found to be statistically significant. There were no substantial cross-loadings on any item across the scales.

The loadings of the disorientation- and flow items exceeded the cut-off point, except DIS3, DIS7, CN2 and TT2. However, these items were retained, given that (a) there were no substantial cross-loadings, (b) their loadings were significant, (c) the scales have been previously validated and (d) in PLS including weaker items helps “to extract what useful information is available in the indicator to create a better construct score” (Barroso et al., 2010, p. 433), where weaker items contribute to construct scores with a lower weight. At the construct level, reliability was analysed

using the composite reliability coefficient, which needs to be 0.7 or higher. All values exceeded this cut-off point.

Convergent validity - the extent of consistency among the items measuring a particular construct - was analysed using the average variance extracted (AVE) by a construct from its indicators, which should be 0.7 or higher (Henseler et al., 2009). Most values of the flow scale exceeded this cut-off point, but the coefficients of three flow subscales and the disorientation scale did not. However, as all coefficients for flow subscales and the disorientation scale were greater than 0.5, on average, more variability in the items was accounted for by their respective factor than was not. Therefore, the scales were retained.

Discriminant validity (the extent to which a measure of a particular construct differs from measures of other constructs) was assessed by analyzing the AVE by each construct from its indicators, which – according to the Fornell-Larcker-criterion – should be greater than its squared correlation with the remaining constructs. All values met this condition. The findings of discriminant validity provide further evidence for the distinction between the dimensions of flow and disorientation. A composite score was created for each of the factors of flow and disorientation per participant, using the PLS weighted-average algorithm.

Further psychometric analyses were conducted on pairs subsamples (English and Malaysian; single item and whole form; transformed GEFT [reflection followed by square root] score below or equal to median value or above median; with and without correction facility; with and without secondary task; random and ascending

order).[20] The same pattern of results was found in each analysis. Moreover, the pattern of loadings in each analysis was the same as that of the full sample.

In addition, Henseler et al.'s (2009) permutation test for model parameters in PLS path-modelling was conducted to compare AVE and loadings between single item and whole form, and between England and Malaysia. AVE should be seen as more important because (sub)scale items are used together, rather than in isolation, to measure a particular construct and therefore an evaluation of the combined items through AVE is more useful. AVE for the Flow Subscale Concentration (CT) was significantly higher for single item (SI) (0.83) than for whole form (WF) (0.64), $p < 0.01$. Still AVE values exceeded 0.50 for both questionnaire layouts, so more variance was accounted for than not. This result was consistent with significant differences in loadings on Item CT1 (SI: 0.91; WF: 0.75; $p < 0.05$) and Item CT4 (SI: 0.93; WF: 0.77; $p < 0.001$). In addition, there were significant differences in loadings on the Flow Subscale Feedback (FB), with higher values for whole form on Item FB1 (SI: 0.76; WF: 0.91; $p < 0.05$) and Item FB2 (SI: 0.90; WF: 0.96; $p < 0.05$). Despite the differences in loadings, all loadings exceeded the cut-off point of 0.70.

AVE for the Flow Subscale Concentration (CT) was significantly higher for England (0.83) than for Malaysia (0.63), $p < 0.01$, as was AVE for Mergence of Action (MA)and Awareness (England: 0.75; Malaysia: 0.58), $p < 0.05$. Still AVE values exceeded 0.50 for both nations, so more variance was accounted for than not. These results were consistent with significant differences in loadings on the CT and MA, with higher values for England on Item CT1 (England: 0.91; Malaysia: 0.73; $p < 0.01$), Item CT4 (England: 0.91; Malaysia: 0.80, $p < 0.05$) and Item MA2 (England:

[20] Given insufficient size for covariance-based structural equation modelling and other considerations stated above, covariance-based invariance analysis would not be appropriate.

0.91; Malaysia: 0.73, 0.05 < $p = 0.06$). In addition, loadings were higher for England on the following items: GC4 (England: 0.89; Malaysia: 0.76, $p < 0.05$), CN2 (England: 0.82; Malaysia: 0.50; $p < 0.05$) and TS2 (England: 0.89; Malaysia: 0.67; p < 0.05). Despite the differences in loadings, most loadings exceeded the cut-off point of 0.70.

Table A1.1

Coefficients of reliability and convergent validity

Construct/indicator	Average variance extracted	Composite reliability	Loading	Standard error	t^a
Balance of challenge and skill	0.73	0.84			
- CS1			0.78	0.12	6.68
- CS2			0.92	0.03	34.70
Clarity of goals	0.79	0.94			
- CG1			0.90	0.02	41.54
- CG2			0.92	0.01	61.80
- CG3			0.86	0.03	25.51
- CG4			0.86	0.02	37.20
Feedback	0.80	0.94			
- F1			0.85	0.04	19.43
- F2			0.93	0.02	48.66
- F3			0.91	0.02	40.78
- F4			0.89	0.02	40.26
Concentration	0.61	0.86			
- CN1			0.84	0.03	31.82
- CN2			0.66	0.09	7.68
- CN3			0.81	0.05	15.79
- CN4			0.80	0.06	14.19
Control	0.74	0.92			
- CT1			0.84	0.04	20.53
- CT2			0.86	0.03	29.52
- CT3			0.89	0.02	39.23
- CT4			0.85	0.03	27.69
Mergence of action and awareness	0.67	0.80			
- M1			0.84	0.04	20.87
- M2			0.79	0.07	11.02
Transformation of time	0.68	0.86			
- TT1			0.90	0.02	52.38
- TT2			0.66	0.10	6.48
- TT3			0.89	0.03	27.55
Transcendence of self	0.72	0.88			
- TS1			0.90	0.10	8.79
- TS2			0.81	0.13	6.14
- TS3			0.83	0.13	6.20
Autotelic experience	0.80	0.94			

- AE1			0.87	0.02	42.05
- AE2			0.91	0.02	49.96
- AE3			0.89	0.02	38.57
- AE4			0.90	0.02	50.97
Disorientation	0.51	0.87			
- DIS1			0.70	0.08	9.14
- DIS2			0.74	0.06	13.33
- DIS3			0.65	0.08	7.66
- DIS4			0.76	0.07	11.52
- DIS5			0.82	0.05	16.00
- DIS6			0.83	0.05	16.91
- DIS7			0.38	0.13	2.99

Note. AE: autotelic experience. CS: balance of challenge and skill. CN: concentration. CT: control. FB: feedback. CG: clarity of goals. MA: mergence of action and awareness. TS: transcendence of self. TT: transformation of time. DIS: disorientation.

[a]Bootstrap, $N = 5000$.

Table A1.2
Coefficients of discriminant validity

	sqrt(AVE)	DIS	AE	CS	CN	CT	FB	CG	MA	TS
DIS	0.71									
AE	0.90	0.06								
CS	0.85	-0.30	0.13							
CN	0.78	-0.09	0.47	0.37						
CT	0.86	-0.27	0.25	0.39	0.59					
FB	0.89	-0.13	0.59	0.48	0.57	0.53				
CG	0.89	-0.05	0.60	0.41	0.51	0.50	0.79			
MA	0.82	-0.26	0.22	0.45	0.46	0.50	0.44	0.41		
TS	0.85	-0.28	0.04	0.35	0.15	0.34	0.13	0.08	0.23	
TT	0.83	-0.03	0.45	0.23	0.42	0.31	0.48	0.45	0.27	0.22

Note. sqrt(AVE): square root of average variance extracted. Off-diagonal values are correlations. AE: autotelic experience. CS: balance of challenge and skill. CN: concentration. CT: control. FB: feedback. CG: clarity of goals. MA: mergence of action and awareness. TS: transcendence of self. TT: transformation of time. DIS: disorientation.

Online Appendix A2: psychometric analysis (Experiment 2)

As in Experiment 1, partial-least-squares path modelling (PLS for short; Vinzi et al., 2010) was used for psychometric data analysis using SmartPLS (Ringle et al., 2005). In testing the measurement model, reliability was analysed (see Table A2.1), and convergent and discriminant validity was assessed (see Table A2.2). Using a bootstrapping procedure, the loadings of all items of the flow scale and the disorientation scale were found to be statistically significant. There were no substantial cross-loadings across the scales for either questionnaire layout, except FB4 loading 0.74 with single-item layout and 0.72 with whole-form layout on the factor goal clarity; however, loadings on the factor feedback were substantially higher: 0.93 with single-item layout and 0.92 with whole-form layout.

The loadings of the disorientation- and flow items exceeded the cut-off point of 0.70, except DIS3 with single-item layout, CN2 for both questionnaire layouts, TT2 for single-item layout and TS2 for whole-form layout. However, these items were retained, given that (a) there were no substantial cross-loadings, (b) their loadings were significant, (c) the scales have been previously validated and (d) in PLS including weaker items helps “to extract what useful information is available in the indicator to create a better construct score” (Barroso et al., 2010, p. 433), where weaker items contribute to construct scores with a lower weight. At the construct level, reliability was analysed using the composite reliability coefficient. All coefficients exceeded the cut-off point of 0.70.

Regarding convergent validity, most AVE values of the flow scale exceeded the cut-off point of 0.70, but the values of the disorientation scale did not. However, as all coefficients for flow subscales and the disorientation scale were greater than 0.5, on

average, more variability in the items was accounted for by their respective factor than was not. Therefore, the scales were retained.

Discriminant validity (the extent to which a measure of a particular construct differs from measures of other constructs) was assessed by analyzing the AVE by each construct from its indicators, which – according to the Fornell-Larcker-criterion – should be greater than its squared correlation with the remaining constructs. All values met this condition. The findings of discriminant validity provide further evidence for the distinction between the dimensions flow and disorientation. The same pattern of results was found for single-item layout and whole-form layout. A composite score was created for each of the factors of flow and disorientation per participant, using the PLS weighted-average algorithm.

Correlations between single-item and whole-form layouts were analysed per subscale (see Table A2.3). Correlations for the same subscale were higher than correlations of a (sub)scale with any other (sub)scale under the other questionnaire layout, providing further evidence for discriminant validity.

In addition, Henseler et al.'s (2009) permutation test for model parameters in PLS path-modelling was conducted to compare AVE and loadings between single item and whole form. There were no significant differences on AVE for any of the (sub)scales. There were significant differences in loadings on Item CS2 (SI: 0.84; WF: 0.90; $p < 0.05$), Item CT1 (SI: 0.84; WF: 0.92; $p < 0.05$) and Item CT4 (SI: 0.85; WF: 0.73; $p < 0.05$). Despite the differences in loadings, all loadings exceeded the cut-off point of 0.70. Sample size for the subsamples of England and Malaysia were too small to analyse separately, according to the minimum requirements for sample size (Barclay et al., 1995), and then compare these two nations.

Table A2.1

Coefficients of reliability and convergent validity

Construct/indicator	Average variance extracted		Composite reliability		Loading		Standard error		t^a	
	SI	WF	SI	WF	SI	WF	SI	WF	SI	WF
Balance of challenge and skill	0.75	0.76	0.85	0.86						
- CS1					0.89	0.85	0.03	0.09	34.09	9.82
- CS2					0.84	0.90	0.05	0.02	18.51	40.89
Clarity of goals	0.80	0.82	0.94	0.95						
- CG1					0.93	0.92	0.01	0.02	62.81	52.97
- CG2					0.90	0.91	0.02	0.02	39.77	39.08
- CG3					0.91	0.92	0.03	0.02	32.42	41.36
- CG4					0.85	0.88	0.05	0.04	15.47	23.57
Feedback	0.74	0.78	0.92	0.93						
- F1					0.92	0.89	0.02	0.03	56.43	27.74
- F2					0.89	0.93	0.02	0.01	42.01	64.22
- F3					0.76	0.86	0.08	0.04	9.95	22.85
- F4					0.86	0.85	0.03	0.04	24.73	19.38
Concentration	0.75	0.63	0.85	0.87						
- CN1					0.88	0.90	0.04	0.02	24.46	37.83
- CN2					0.57	0.48	0.11	0.12	5.10	4.11
- CN3					0.86	0.85	0.04	0.05	21.77	15.82
- CN4					0.89	0.89	0.03	0.03	35.34	29.51
Control	0.66	0.72	0.88	0.91						
- CT1					0.84	0.92	0.04	0.02	21.67	60.52
- CT2					0.87	0.86	0.03	0.03	31.46	25.74
- CT3					0.84	0.87	0.03	0.04	25.29	23.70
- CT4					0.85	0.73	0.04	0.07	21.00	9.90

Mergence of action and awareness	0.76	0.68	0.86	0.80						
- M1					0.92	0.93	0.03	0.05	31.81	18.78
- M2					0.82	0.69	0.07	0.12	11.53	5.64
Transformation of time	0.68	0.71	0.86	0.88						
- TT1					0.93	0.89	0.02	0.04	43.21	24.21
- TT2					0.58	0.71	0.16	0.12	3.71	6.10
- TT3					0.91	0.92	0.03	0.04	30.82	25.43
Transcendence of self	0.72	0.70	0.88	0.87						
- TS1					0.89	0.95	0.04	0.03	25.07	28.62
- TS2					0.73	0.61	0.11	0.18	6.90	3.31
- TS3					0.90	0.90	0.03	0.07	26.82	12.77
Autotelic experience	0.82	0.82	0.95	0.95						
- AE1					0.93	0.91	0.02	0.02	46.84	50.34
- AE2					0.88	0.92	0.03	0.02	26.35	45.04
- AE3					0.91	0.95	0.03	0.01	34.45	94.19
- AE4					0.91	0.83	0.02	0.02	40.00	52.97
Disorientation	0.53	0.59	0.89	0.91						
- DIS1					0.79	0.87	0.06	0.03	12.71	24.92
- DIS2					0.80	0.76	0.07	0.07	11.86	10.32
- DIS3					0.57	0.64	0.13	0.11	4.50	5.62
- DIS4					0.74	0.77	0.09	0.07	8.26	10.80
- DIS5					0.68	0.67	0.09	0.10	7.67	6.89
- DIS6					0.75	0.84	0.08	0.04	9.94	21.48
- DIS7					0.74	0.79	0.11	0.06	6.65	13.72

Note. SI: single item. WF: whole form. AE: autotelic experience. CS: balance of challenge and skill. CN: concentration. CT: control. FB: feedback. CG: clarity of goals. MA: mergence of action and awareness. TS: transcendence of self. TT: transformation of time. DIS: disorientation.

[a]Bootstrap, $N = 5000$.

Table A2.2
Coefficients of discriminant validity

Single item

	sqrt(AVE)	DIS	AE	CS	CN	CT	FB	CG	MA	TS
DIS	0.91									
AE	0.81	0.63								
CS	0.86	0.41	0.55							
CN	0.85	0.35	0.53	0.46						
CT	0.73	-0.11	-0.27	-0.36	-0.40					
FB	0.86	0.53	0.43	0.61	0.33	-0.09				
CG	0.90	0.70	0.54	0.47	0.32	-0.14	0.73			
MA	0.87	0.33	0.37	0.38	0.35	-0.13	0.41	0.36		
TS	0.85	0.15	0.38	0.37	0.41	-0.09	0.13	0.12	0.47	
TT	0.82	0.60	0.64	0.36	0.37	-0.09	0.23	0.34	0.36	0.26

Whole form

	sqrt(AVE)	DIS	AE	CS	CN	CT	FB	CG	MA	TS
DIS	0.90									
AE	0.80	0.61								
CS	0.87	0.33	0.52							
CN	0.85	0.41	0.66	0.65						
CT	0.77	-0.13	-0.35	-0.38	-0.44					
FB	0.88	0.66	0.50	0.42	0.47	-0.22				
CG	0.90	0.73	0.54	0.39	0.41	-0.09	0.71			
MA	0.82	0.18	0.50	0.40	0.45	-0.30	0.26	0.16		
TS	0.83	0.00	0.35	0.39	0.37	-0.20	0.06	0.09	0.42	
TT	0.84	0.44	0.57	0.37	0.38	-0.10	0.13	0.22	0.31	0.22

Note. sqrt(AVE): square root of average variance extracted. Off-diagonal values are correlations among (sub)scales. AE: autotelic experience. CS: balance of challenge and skill. CN: concentration. CT: control. FB: feedback. CG: clarity of goals. MA: mergence of action and awareness. TS: transcendence of self. TT: transformation of time. DIS: disorientation.

Table A2.3

Correlations among subscales between questionnaire layouts

	DIS-WF	CS-WF	FB-WF	GC-WF	CN-WF	CT-WF	MA-WF	TS-WF	TT-WF	AE-WF
DIS-SI	0.92	0.60	0.33	0.40	0.57	0.67	0.16	0.02	0.47	-0.14
CS-SI	0.59	0.82	0.40	0.56	0.43	0.52	0.34	0.24	0.51	-0.28
FB-SI	0.36	0.51	0.67	0.58	0.48	0.43	0.34	0.26	0.26	-0.35
GC-SI	0.32	0.47	0.47	0.73	0.35	0.30	0.33	0.32	0.19	-0.42
CN-SI	0.54	0.39	0.50	0.42	0.83	0.70	0.23	0.08	0.11	-0.13
CT-SI	0.70	0.48	0.39	0.36	0.68	0.89	0.15	0.04	0.22	-0.11
MA-SI	0.28	0.44	0.39	0.43	0.38	0.34	0.67	0.38	0.29	-0.19
TS-SI	0.07	0.39	0.51	0.36	0.05	0.11	0.37	0.77	0.31	-0.20
TT-SI	0.48	0.58	0.27	0.40	0.15	0.26	0.34	0.21	0.82	-0.10
AE-SI	-0.02	-0.28	-0.27	-0.39	-0.14	-0.04	-0.12	-0.04	0.06	0.81

Note. AE: autotelic experience. CS: balance of challenge and skill. CN: concentration. CT: control. FB: feedback. CG: clarity of goals. MA: mergence of action and awareness. TS: transcendence of self. TT: transformation of time. DIS: disorientation. SI: single item. WF: whole form.

Online Appendix A3: psychometric analysis (Experiment 3)

As in Experiment 1, partial-least-squares path modelling (PLS for short; Vinzi et al., 2010) was used for psychometric data analysis using SmartPLS (Ringle et al., 2005). In testing the measurement model, reliability was analysed (see Table A3.1), and convergent and discriminant validity was assessed (see Table A3.2). Using a bootstrapping procedure, the loadings of all items of the flow scale and the disorientation scale were found to be statistically significant. There were no substantial cross-loadings on any item across the scales.

The loadings of the flow items exceeded the cut-off point of 0.70. The loadings of the disorientation scale also exceeded the cut-off point, except for DIS3 and DIS4 with both questionnaire layouts. However, these items were retained, given that (a) there were no substantial cross-loadings, (b) their loadings were significant, (c) the scales have been previously validated and (d) in PLS including weaker items helps "to extract what useful information is available in the indicator to create a better construct score" (Barroso et al., 2010, p. 433), where weaker items contribute to construct scores with a lower weight. At the construct level, reliability was analysed using the composite reliability coefficient. All coefficients exceeded the cut-off point of 0.70.

Regarding convergent validity, all AVE values exceeded this cut-off point of 0.7, but the coefficients of the disorientation scale did not. However, as all coefficients for flow subscales and the disorientation scale were greater than 0.5, on average, more variability in the items was accounted for by their respective factor than was not. Therefore, the scales were retained.

Discriminant validity (the extent to which a measure of a particular construct differs from measures of other constructs) was assessed by analyzing the AVE by each

construct from its indicators, which – according to the Fornell-Larcker-criterion – should be greater than its squared correlation with the remaining constructs. All values met this condition. The findings of discriminant validity provide further evidence for the distinction between the dimensions flow and disorientation. The same pattern of results was found for single-item layout and whole-form layout. A composite score was created for each of the factors of flow and disorientation per participant, using the PLS weighted-average algorithm.

Correlations between single-item and whole-form layouts were analysed per subscale (see Table A3.3). Correlations between the questionnaire layouts for the same subscale were higher than correlations of a subscale with any other subscale under the other questionnaire layout, therefore providing further evidence for discriminant validity.

In addition, Henseler et al.'s (2009) permutation test for model parameters in PLS path-modelling was conducted to compare AVE and loadings between single item and whole form. There were no significant differences on AVE for any of the (sub)scales. There were significant differences in loadings on Item CT4 (SI: 0.89; WF: 0.95; $p < 0.01$), and Item TS2 (SI: 0.91; WF: 0.81; $p < 0.01$). Despite the differences in loadings, all loadings exceeded the cut-off point of 0.70.

Table A3.1

Coefficients of reliability and convergent validity

Construct/indicator	Average variance extracted		Composite reliability		Loading		Standard error		t^a	
	SI	WF	SI	WF	SI	WF	SI	WF	SI	WF
Balance of challenge and skill	0.86	0.89	0.93	0.94						
- CS1					0.92	0.94	0.03	0.02	34.36	55.27
- CS2					0.94	0.94	0.01	0.02	85.19	54.81
Clarity of goals	0.70	0.74	0.90	0.92						
- CG1					0.81	0.85	0.06	0.04	12.61	19.63
- CG2					0.87	0.92	0.03	0.02	34.13	49.70
- CG3					0.87	0.90	0.05	0.02	18.16	47.62
- CG4					0.77	0.76	0.06	0.06	13.89	13.36
Feedback	0.78	0.83	0.93	0.95						
- F1					0.84	0.88	0.04	0.02	22.26	35.88
- F2					0.90	0.93	0.02	0.02	41.37	55.58
- F3					0.88	0.92	0.03	0.02	27.16	48.95
- F4					0.92	0.93	0.02	0.02	44.78	53.80
Concentration	0.80	0.77	0.94	0.93						
- CN1					0.92	0.91	0.02	0.02	58.46	52.77
- CN2					0.83	0.79	0.04	0.04	21.49	18.83
- CN3					0.92	0.92	0.02	0.02	52.23	61.58
- CN4					0.91	0.87	0.03	0.03	32.72	27.76
Control	0.83	0.81	0.95	0.94						
- CT1					0.92	0.92	0.01	0.02	64.00	47.71
- CT2					0.90	0.87	0.03	0.05	35.39	18.02
- CT3					0.93	0.86	0.02	0.05	58.49	18.04
- CT4					0.89	0.94	0.02	0.01	35.98	93.95

Mergence of action and awareness	0.83	0.81	0.90	0.90						
- M1					0.91	0.91	0.02	0.02	44.11	46.19
- M2					0.91	0.89	0.03	0.03	30.29	32.10
Transformation of time	0.78	0.80	0.91	0.92						
- TT1					0.94	0.95	0.01	0.02	64.76	62.16
- TT2					0.78	0.78	0.08	0.08	9.33	10.36
- TT3					0.92	0.95	0.02	0.01	52.24	65.67
Transcendence of self	0.74	0.69	0.90	0.87						
- TS1					0.86	0.86	0.05	0.04	18.57	21.03
- TS2					0.91	0.81	0.02	0.06	49.21	13.99
- TS3					0.80	0.81	0.06	0.06	13.39	13.53
Autotelic experience	0.80	0.78	0.94	0.93						
- AE1					0.90	0.88	0.02	0.04	41.94	24.73
- AE2					0.88	0.84	0.03	0.06	28.81	14.93
- AE3					0.90	0.90	0.02	0.03	38.09	32.24
- AE4					0.90	0.91	0.02	0.02	39.55	48.61
Disorientation	0.54	0.57	0.89	0.90						
- DIS1					0.84	0.81	0.03	0.04	29.17	18.89
- DIS2					0.80	0.81	0.05	0.04	17.02	20.61
- DIS3					0.53	0.65	0.08	0.06	6.30	10.29
- DIS4					0.63	0.68	0.07	0.06	9.71	11.47
- DIS5					0.74	0.78	0.05	0.05	14.47	14.32
- DIS6					0.83	0.76	0.03	0.05	24.36	15.38
- DIS7					0.72	0.77	0.05	0.04	14.23	17.40

Note. SI: single item. WF: whole form. AE: autotelic experience. CS: balance of challenge and skill. CN: concentration. CT: control. FB: feedback. CG: clarity of goals. MA: mergence of action and awareness. TS: transcendence of self. TT: transformation of time. DIS: disorientation.

[a]Bootstrap, $N = 5000$.

Table A3.2
Coefficients of discriminant validity

Single item

	sqrt(AVE)	DIS	AE	CS	CN	CT	FB	CG	MA	TS
DIS	0.89									
AE	0.89	0.39								
CS	0.93	0.25	0.50							
CN	0.91	0.31	0.65	0.72						
CT	0.74	-0.35	-0.35	-0.45	-0.52					
FB	0.88	0.33	0.43	0.52	0.59	-0.49				
CG	0.83	0.29	0.58	0.55	0.62	-0.42	0.60			
MA	0.91	0.29	0.38	0.60	0.51	-0.53	0.47	0.45		
TS	0.86	0.05	0.44	0.40	0.36	-0.26	0.34	0.51	0.39	
TT	0.88	0.50	0.53	0.32	0.31	-0.15	0.14	0.35	0.21	0.12

Whole form

	sqrt(AVE)	DIS	AE	CS	CN	CT	FB	CG	MA	TS
DIS	0.88									
AE	0.88	0.35								
CS	0.94	0.12	0.44							
CN	0.90	0.30	0.66	0.64						
CT	0.75	-0.30	-0.31	-0.47	-0.46					
FB	0.91	0.23	0.46	0.47	0.63	-0.44				
CG	0.86	0.22	0.55	0.47	0.70	-0.42	0.69			
MA	0.90	0.31	0.50	0.52	0.52	-0.50	0.44	0.52		
TS	0.83	0.05	0.48	0.19	0.28	-0.29	0.37	0.40	0.34	
TT	0.90	0.53	0.41	0.17	0.22	-0.21	0.06	0.22	0.32	0.20

Note. sqrt(AVE): square root of average variance extracted. Off-diagonal values are correlations among (sub)scales. AE: autotelic experience. CS: balance of challenge and skill. CN: concentration. CT: control. FB: feedback. CG: clarity of goals. MA: mergence of action and awareness. TS: transcendence of self. TT: transformation of time. DIS: disorientation.

Table A3.3

Correlations among subscales between questionnaire layouts

	DIS-WF	CS-WF	FB-WF	GC-WF	CN-WF	CT-WF	MA-WF	TS-WF	TT-WF	AE-WF
DIS-SI	0.90	-0.39	-0.45	-0.43	-0.27	-0.48	-0.46	-0.26	-0.14	-0.32
CS-SI	-0.51	0.75	0.60	0.59	0.56	0.73	0.57	0.39	0.16	0.20
FB-SI	-0.46	0.46	0.91	0.62	0.45	0.63	0.49	0.33	0.01	0.24
GC-SI	-0.43	0.44	0.49	0.74	0.58	0.58	0.52	0.49	0.27	0.20
CN-SI	-0.37	0.39	0.41	0.53	0.87	0.65	0.41	0.43	0.41	0.37
CT-SI	-0.53	0.70	0.55	0.63	0.59	0.88	0.47	0.27	0.16	0.21
MA-SI	-0.55	0.53	0.40	0.38	0.35	0.45	0.65	0.24	0.14	0.22
TS-SI	-0.25	0.23	0.32	0.29	0.44	0.26	0.23	0.73	0.04	-0.02
TT-SI	-0.20	0.24	0.09	0.27	0.43	0.32	0.39	0.26	0.86	0.40
AE-SI	-0.32	0.15	0.26	0.20	0.32	0.30	0.33	0.09	0.54	0.81

Note. AE: autotelic experience. CS: balance of challenge and skill. CN: concentration. CT: control. FB: feedback. CG: clarity of goals. MA: mergence of action and awareness. TS: transcendence of self. TT: transformation of time. DIS: disorientation. SI: single item. WF: whole form.

References for Online Appendixes

Barroso, C., Carrión, G.C., Roldán, J.K., 2010. Applying maximum likelihood and PLS on different sample sizes: Studies on SERVQUAL model and employee behavior model, in Vinzi, V.E., Chin, W.W., Henseler, J., Wang, H. (Eds.), Handbook of Partial Least Squares: Concepts, Methods and Applications in Marketing and Related Fields. Springer, Berlin, pp. 427-447.

Barclay, D., Higgins, C., Thompson, R., 1995. The partial least squares approach to causal modeling: person computer adoption as illustration. Technology Studies 2, 285-309.

Chin, W.W., 2010. How to write up and report PLS analyses, in Vinzi, V.E., Chin, W.W., Henseler, J., Wang, H. (Eds.), Handbook of Partial Least Squares: Concepts, Methods and Applications in Marketing and Related Fields . Springer, Berlin, pp. 655-690.

Henseler, J., Ringle, C., Sinkovics, R., 2009. The use of partial least squares modeling in international marketing, in Cavusgil, T., Sinkovics, R., Ghauri, P. (Eds.), New Challenges in International Marketing (Advances in International Marketing). Emerald, London, pp. 277-319.

Hulland, J., Ryan, M., Rayner, R., 2010. Modeling customer satisfaction: A comparative evaluation of covariance structure analysis versus partial least squares, in Vinzi, V.E., Chin, W.W., Henseler, J., Wang, H. (Eds.), Handbook of Partial Least Squares: Concepts, Methods and Applications in Marketing and Related Fields. Springer, Berlin, pp. 307-325.

Vilares, M., Almeida, M., Coelho, P., 2010. Comparison of likelihood and PLS estimators for structural equation modeling: A simulation with customer satisfaction data, in Vinzi, V.E., Chin, W.W., Henseler, J., Wang, H. (Eds.), Handbook of Partial Least Squares: Concepts, Methods and Applications in Marketing and Related Fields. Springer, Berlin, pp. 289-305.

Vinzi, V.E., Chin, W.W., Henseler, J., Wang, H., 2010. Handbook of partial least squares: Concepts, methods and applications in marketing and related fields.