Computers in Human Behavior xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study

Stefan Stieger [a,*], Ulf-Dietrich Reips [b,c]

[a] Department of Basic Psychological Research, School of Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria
[b] Departamento de Psicología, Universidad de Deusto, Apartado 1, 48080 Bilbao, Spain
[c] IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Response behavior
Response latency
Response visualization
Non-reactive data
Online questionnaire
Paradata

ABSTRACT

The use of online questionnaires is rapidly increasing. Contrary to manifold advantages, not much is known about user behavior that can be measured outside the boundaries set by standard web technologies like HTML form elements. To show how the lack of knowledge about the user setting in web studies can be accounted for, we present a tool called UserActionTracer, with which it is possible to collect more behavior information than with any other paradata gathering tool, in order to (1) gather additional data unobtrusively from the process of answering questions and (2) to visualize individual user behavior on web pages. In an empirical study on a large web sample ($N = 1046$) we observed and analysed online behaviors (e.g., clicking through). We found that only 10.5% of participants showed more than five single behaviors with highly negative influence on data quality in the whole online questionnaire (out of 132 possible single behavior judgments). Furthermore, results were validated by comparison with data from online address books. With the UserActionTracer it is possible to gain further insight into the process of answering online questionnaires.

1. Introduction

Since 1995 the use of online questionnaires in the social and behavioral sciences has rapidly increased (Birnbaum, 2004; Reips, 2001, 2007). This new way of data collection resulted in a more differentiated perception of advantages and disadvantages of data gathered via the Internet (Reips, 2000). Numerous advantages like asynchronism, alocality, flexibility, and automation are well documented. In addition to self-selection, representativeness, and uncontrolled circulation (Batinic & Bosnjak, 2000), the unknown setting is one of the main disadvantages in web-based studies and is a potential threat to internal validity (for a review of this matter see Reips, 2000). The unknown setting can influence the quality of data negatively.

So far, data quality has mostly been assessed by the face validity of the gathered data, by controlling for multiple submissions and checking for incomplete data sets. Some studies could validate data through external information from online address books and data from registrations (Stieger & Göritz, 2006; Stieger & Reips, 2008; Voracek, Stieger, & Gindl, 2001). In this paper, we will present and evaluate a tool that collects much more behavioral data than usual and integrates the information visually. With the help of technologies that are part of most web browsers (e.g., JavaScript), it is possible to gather additional data from the answering process, allowing us to look into the black box of this process. In face-to-face studies, in which paper-and-pencil questionnaires are used, usually no data are available about the filling-in process. In studies, however, in which browsers are used, paradata (auxiliary data describing the process, e.g., answering times, clicks, scrolls, typing) as well as metadata (e.g., used web browser, used operating system) can easily be collected (Couper, 2000; Heerwegh, 2003; Reips, 1997, 2009). So far, only few studies have used paradata from online behavior in order to judge data quality (Bassili & Fletcher, 1991; Converse, 1970; Couper, 2000; Couper, Traugott, & Lamias, 2001; Jeavons, 1999; Kaufmann & Reips, 2008; Nichols & Sedivi, 1998; Wittchen, Schlereth, & Hertel, 2007). Heerwegh (2003) emphasizes the potential of paradata that are collected client-side:

> "Client-side paradata enable web survey researchers to obtain detailed information on response behavior. Despite some of the problems related to client-side paradata (e.g., the very large data files and the problems associated with extracting useful information from these data), these data do offer researchers the possibility to perform in-depth (methodological) research."

There are also several publications from the area of Human–Computer Interaction (HCI) and Web Usage Mining that deal with the recording of online behavior. Web Usage Mining basically looks

[*] Corresponding author. Tel.: +43 1 4277 47847.
E-mail address: stefan.stieger@univie.ac.at (S. Stieger).

0747-5632/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.chb.2010.05.013

at how frequently web pages, portions of web pages, or services on websites are accessed and how these frequencies or changes in frequencies may be used as indicators. The primary domain, however, is usability tests (Hilbert & Redmiles, 2000; Ivory & Hearst, 2001) to better understand the interaction between users and web pages. The main goal is the improvement of website design, but methodological advancement of online surveying is also an important motivation (Buchanan & Reips, 2001; Reips, 2010).

Basically, there are several ways to collect paradata. Some tools use server log files, which are produced by almost any web server by default (WebQuilt: Hong, Heer, Waterson, & Landay, 2001; OpenWebSurvey: Baravalle & Lanfranchi, 2003; Scientific LogAnalyzer: Reips & Stieger, 2004). The main problem of this approach was described by Cugini and Laskowski (2001) in a somewhat exaggerated yet pointed way: "…a server log tracks the activity of a server, not a subject." Because of this and other problems, such as losing track of "back" button use and caching of websites by proxy servers (Birnbaum & Reips, 2005; Reips, 1997), it is sometimes difficult to interpret results. Furthermore, with standard web questionnaires, conclusions can only be drawn from log file analyses of paths through several web pages, but not about the behavior on a certain web page itself.

Due to these restrictions, browser-based data collection methods are more promising. With these methods, data are not collected server-side but client-side, directly on the user's computer. Here two approaches are apparent: (1) installation of a program or a specific browser on the client's computer (WebTracker: Turnbull, 1998; Uzilla: Edmonds, 2003; WebLogger: Reeder, Pirolli, & Card, 2001; ErgoBrowser: Adams & Kleiss, 2003; ObSys: Gellner & Forbrig, 2003) or (2) using the script languages (e.g., JavaScript) that are part of standard browsers like Internet Explorer, Firefox, Safari, or Opera (WebVIP: NIST, 1999; WET: Etgen & Cantor, 1999; Lucidity: Edmonds, 2001). The big advantage of the latter method is its independence from other tools (users do not have to install new tools or plugins). This also diminishes problems like self-selection, motivational confounding (only highly motivated potential participants are willing to spend this extra burden; Reips, 1997, 2000), technical dropouts (due to a failed installation or performance errors of the tool), and non-response error (only technically experienced participants are able to install programs). Furthermore, using script languages for data collection that are already built into web browsers is unobtrusive. Most users are not aware that paradata are being collected. Thus, participants are not motivated to monitor and adapt their behavior accordingly. This is especially interesting in behavioral and social sciences, where social desirability tendencies are a problem (Kaufmann & Reips, 2008).

A tool called UserActionTracer (UAT) will be introduced in this study, with which it is possible to collect paradata about the answering process. Only a small portion of JavaScript is added to the online questionnaire. The following aspects were of importance for the development of such a tool (1) to observe the answering process in a user's natural environment (e.g., at home, at work, or at the university) and not in a laboratory setting, and (2) to program the online questionnaire server-side in order to reduce technology induced dropout (i.e., dropout caused by client-side technologies like Java that do not work in every Internet browser by default; see Buchanan & Reips, 2001; Schmidt, 2007; Schwarz & Reips, 2001).

The tool itself, as well as its design, will be described based on an empirical study answering the following research questions: (1) Which behaviors can be observed during the process of answering the questions online? How frequently do they occur? (2) Is it possible to validate the used procedure of judging data quality using paradata by comparing demographic data from the online questionnaire with demographic data from online address books?

2. Method

2.1. Procedure

The online questionnaire was uploaded to the web server at the University of Vienna. Recruiting took place via several channels: (1) Portals that collect links to online questionnaires and web experiments (Web Experimental Psychology Lab[1]: Reips, 2001; web experiment list[2]: Reips & Lengler, 2005; Psychological Research on the Net[3]; Social Psychology Network[4]); (2) Newsgroups; (3) Instant Messaging – so-called chat requests were sent to interested ICQ users via a self-programmed interview software (Dynamic Interviewing Program: Stieger & Reips, 2008) that is based on ICQ (acronym for "I seek you", see <http://www.icq.com/>); (4) search engines – the link to the online questionnaire was submitted to the most frequently used search engines (e.g., Google, Yahoo). After 219 days in the field, data collection was closed.

2.2. Participants

Data from 1046 participants were collected. Participants were on average 24.4 years old (*SD* = 9.6; range 12–88 years; 58.5% women) and almost two thirds (62.6%) reported to be students. Most participants reported to come from the USA (60.9%), 7.5% from the United Kingdom, 4.7% from Germany, 3.7% from Canada, 2.2% from China, 1.8% from the Philippines, and 17.9% indicated to come from 59 other countries (1.3% gave an invalid answer). The frequency of participants from English speaking countries is probably due to aspects of Internet penetration and the fact that the online questionnaire was in English.

2.3. Material

The current study is based on an online questionnaire about Instant Messaging (IM: Stieger & Göritz, 2006). Questions about habits in using IM and personal opinions about the use of IM for online interviews were asked. Small groups of questions were asked on separate HTML pages. In total there were 23 questions on 11 web pages. Two questions were semantic differentials with 12 attitude dimensions each (7-point Likert scales). The online questionnaire was programmed with a server-side Perl script that also stored the data. On the last page, participants were offered the possibility to leave further comments.

2.4. JavaScript – UserActionTracer

In order to keep a record of users' actions, a script was programmed and implemented on each HTML page called the UserActionTracer (UAT: the tool can be obtained from the first author on request). The tool has the following specifications. To ensure highest compatibility with different types of browsers, a browser switch is used in order to provide the most appropriate JavaScript code. The primary task of the script is to store the users' actions with mouse and keyboard while filling in the online questionnaire. These actions contain a timestamp, so answering times can be measured on an item level. The following actions are recorded with their exact position (*x* and *y* coordinates): clicks with the mouse (including all mouse buttons); double-clicks with the mouse; clicks on checkboxes, radio buttons, and list boxes; choices in drop-down menus; inserted text in text boxes; clicks on submit buttons; keys pressed on the keyboard; and the position of the mouse pointer

[1] <http://wexlab.eu>.
[2] <http://wexlist.net/>.
[3] <http://psych.hanover.edu/Research/exponnet.html>.
[4] <http://www.socialpsychology.org/expts.htm>.

Please cite this article in press as: Stieger, S., & Reips, U.-D. What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior* (2010), doi:10.1016/j.chb.2010.05.013

every half a second. The resulting data string (for an example see Fig. 1) is stored in a hidden text field on the HTML page. These data are then sent from the client (i.e., browser) to the server and are stored in a different location from that of the questionnaire data.

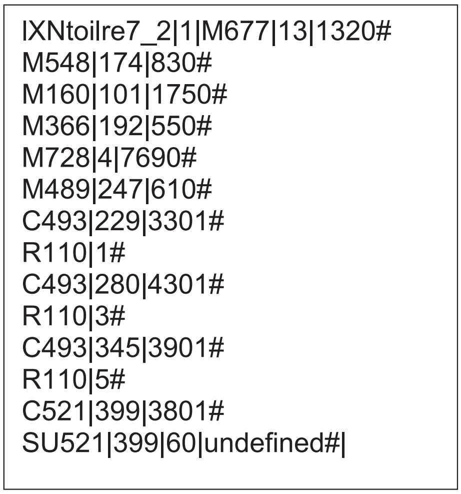
The data string on Fig. 1 can be interpreted in the following way: the pound key symbol (#) is the delimiter between different actions; the pipe-symbol "|" is the delimiter between different elements of an action. Each new participant was assigned an anonymous key (in this case "lXNtoilre7_2") that was used throughout the entire online questionnaire. In the example, the mouse pointer was moved from position $X = 677$, $Y = 13$ to position $X = 548$, $Y = 174$ in 1320 ms after loading the questionnaire (see first two lines in Fig. 1). After another 830 ms it was moved to $X = 160$, $Y = 101$. More movements followed. A single click with the mouse was recorded at the position $X = 493$, $Y = 229$ ("C" stands for "click"; the "1" at the last digit means that the left mouse key was pressed; the click lasted 330 ms). After 110 ms this click resulted in the activation of the radio button ("*R*" stands for "radio button") with the number "1" on the online questionnaire. Further clicks on radio buttons followed ending with the click on the submit button ("SU" stands for "submit").

In addition to the data string described above, metadata about the browser used (brand, version, and language), operating system, and monitor screen (dimensions, resolution) were collected. Data about the monitor screen were important in order to generate pictures of navigation (see next section).

Within this context, it is worth mentioning that the measurement of reaction times in milliseconds via JavaScript is hampered by measurement errors (Chambers & Brown, 2003; Schmidt, 2001; Schwarz & Reips, 2001). These errors depend on the client's computer and its specifications (operating system, browser, central processing unit, used input devices). Yet, it is in a range that is negligible for the current purposes of the study (for example see Reips, submitted for publication; and Schmidt, 2001: mean duration for JavaScript/Jscript animations updating as fast as possible was 130 ms tested for different operating systems and different browsers).

2.5. Preparation of the data string: Database and navigation pictures

Once the study was finished, each data string was reformatted with a Perl script (first aggregation of data) and stored in a database. This was necessary to convert the 336262 single actions all users produced into an analyzable form. Furthermore, the Perl script produced navigation pictures that showed the whole process of a user filling in a page of the online questionnaire (for an example see Figs. 2 and 3; Fig. 3 shows a questionnaire page and the matching navigation picture overlapped). User actions were displayed in graphical form. In order to display the process as well, each action was numbered in ascending order. Longer inactivities were displayed with their respective time. These pictures were very useful in producing a quick overview of the answering process and were also used as a basis for the following analyses.



```
lXNtoilre7_2|1|M677|13|1320#
M548|174|830#
M160|101|1750#
M366|192|550#
M728|4|7690#
M489|247|610#
C493|229|3301#
R110|1#
C493|280|4301#
R110|3#
C493|345|3901#
R110|5#
C521|399|3801#
SU521|399|60|undefined#|
```

Fig. 1. Example of a navigation string produced by the UserActionTracer. *Note:* Line breaks after each action were added for readability.

3. Results

3.1. Research question 1: Which behaviors can be observed during the online answering process and how frequently may they reduce data quality?

A coding scheme was applied to the data about the observed behaviors: low vs. high vs. no negative influence on data quality. We are fully aware that this coding scheme is subjective, and the outcome of this categorization depends on the online questionnaire used. Low influence cases were coded as "1", cases with highly negative influence on data quality were coded as "2" and cases without negative influence were coded as "0" (for examples see the next subsections). In order to demonstrate what can be achieved with the help of the UAT, we divide the following subsections into behaviors that can be observed without the UAT (item non-response, wrong data entry) and behaviors that can only be observed using the UAT. The aforementioned categorization was performed for the following user behaviors.

3.1.1. Behaviors observable without the UAT: Face validity of users' entries

Eight text fields were controlled for plausibility of the given answers. Every single entry was coded as mentioned above. For example, the entry "asdas" for country is invalid and was therefore coded with 2. On the other hand, "pak" could have been invalid, but it could have also stood for "Pakistan", just as "my" could have been meant as an abbreviation for "Malaysia". Therefore, such cases were coded with 1. Tendencies such as giving the same answer to every single question of the semantic differential or different extreme answers within one questionnaire (i.e., only one and seven) were marked, too. Since these are not definite indicators of high negative influence on data quality, these cases were coded with 1.

All in all, it was possible to judge face validity for 938 questionnaires (89.7% of all questionnaires). For the remaining 108 datasets, the JavaScript produced no data traces (i.e., either JavaScript was actively disabled by the participant, see Buchanan & Reips, 2001, or unconventional Internet browsers were used that were not able to interpret JavaScript code). Only 1.9% of the questionnaires showed high negative influence on data quality (see Table 1).

3.1.2. Behaviors observable without the UAT: Item non-response

If participants repeatedly do not fill in items, they show low motivation to participate and thus are likely to produce low data quality. First of all, the quantity of non-response was investigated for each questionnaire page. The sum of all non-response was then divided by the number of displayed items (in case of dropout depending on the page where the participant stopped). This resulted in a ratio (number of non-responses/number of seen items) with "0" meaning all questions were filled in and "1" meaning that no question was filled in. After a visual inspection of the ratios' distribution, questionnaires with more than 30% of displayed items not answered were coded as 2. Between 30% and 10% of displayed items not answered a questionnaire was coded as 1, below 10% it

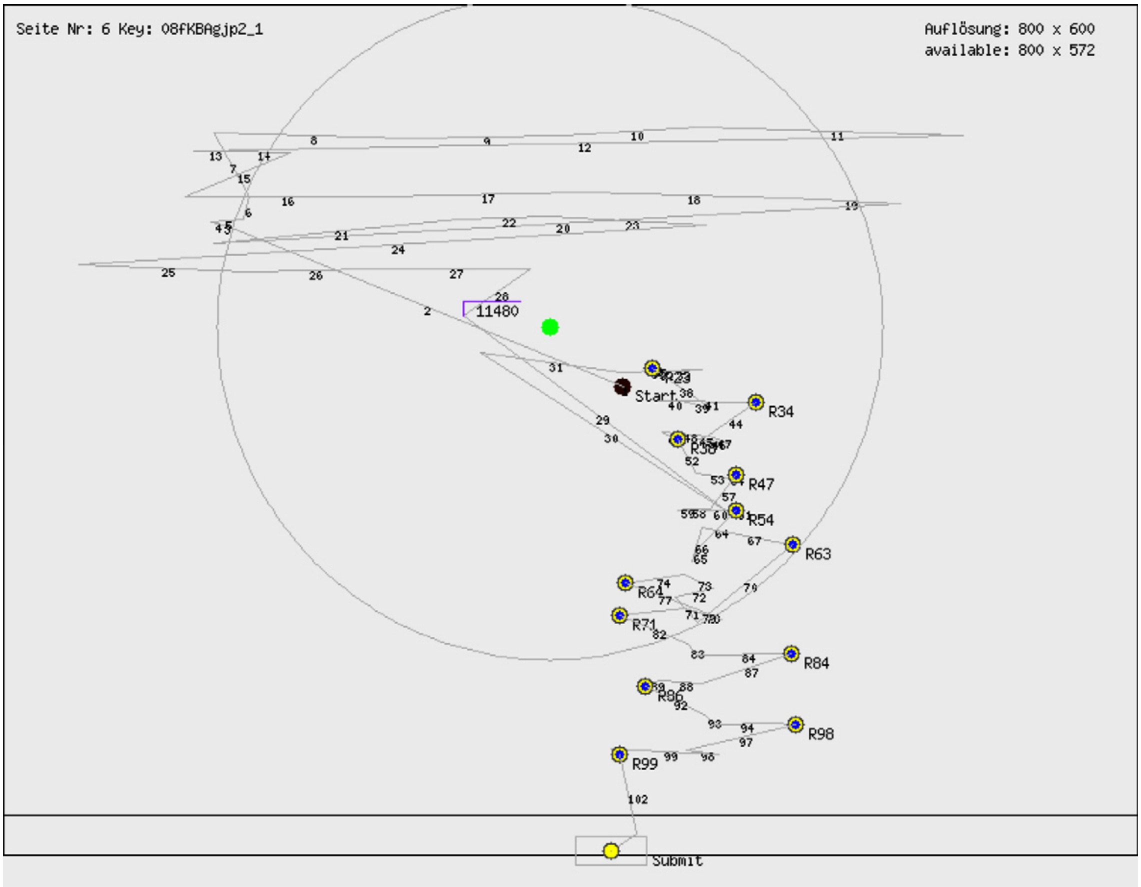


Fig. 2. Navigation picture of a questionnaire page containing introductory text (in the upper third of the page) and a semantic differential with twelve opinion dimensions (each 7-point Likert scale). *Note:* The small full circle in the middle of the screen above the printed number 31 indicates the line median point and the surrounding circle the standard deviation of all mouse movements by that participant on that page. Small line circles indicate clicks with the mouse (e.g., circle at the submit button). Smaller full circles within small line circles indicate activated radio buttons (followed by an "*R*" and a number). The two lines building a corner at "11480" indicate an inactivity of 11.48 s at this place. The full circle in the middle of the screen labeled "Start" indicates the first position of the mouse when the online questionnaire was loaded. On the top left corner is the number of the questionnaire page (German: "Seite") and the participant's individual anonymous identification key. On the top right corner of the screen is the resolution of the user's screen as well as the available resolution (some participants did not maximize the browser window to full screen).

was coded as 0. Considering item non-response only, for 6.6% of questionnaires a highly negative influence on data quality was observed (see Table 1).

3.1.3. Behaviors observable with the UAT: Changes in text fields

Because each key input on the computers' keyboard was recorded, changes of text input were observable. For example, 31 participants changed their reported age. This could have happened on purpose or due to a typing error (in 15 cases the age was changed only by 1 year). In seven cases the country of origin was changed. All these cases were coded with 1 (low negative influence).

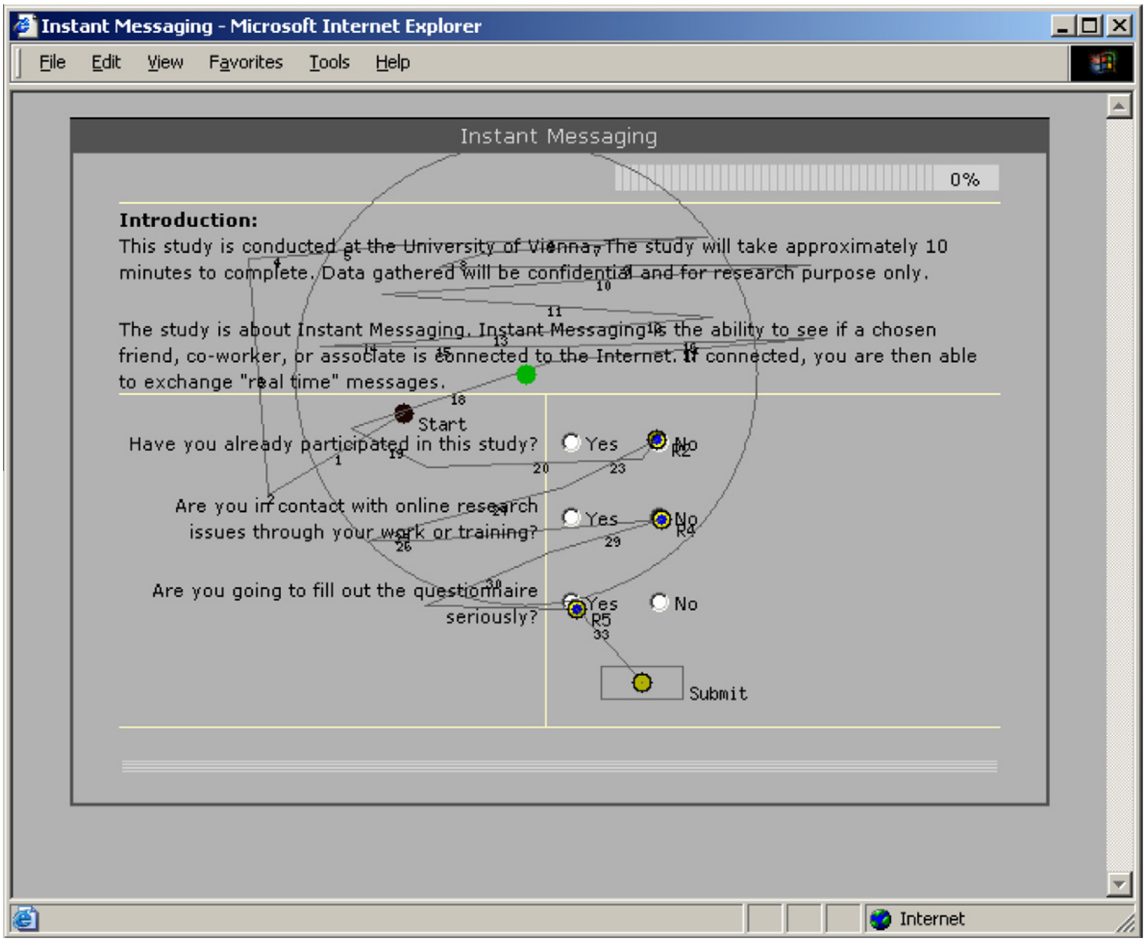
For 928 questionnaires (88.7% of all questionnaires) it was possible to observe if there were changes on text fields. For the remaining 11.3%, a judgment was not possible due to several reasons, e.g., text fields were not filled in or participants dropped out. Only 4.1% of the questionnaires showed some kind of change (see Table 1).

3.1.4. Behaviors observable with the UAT: Changes on radio buttons, checkboxes, and drop-down menus

With radio buttons and drop-down menus it is possible to change already marked options just by clicking on another option and with checkboxes by clicking on the same option again. Such behaviors, called "changing" throughout this study, would be obvious in paper-and-pencil questionnaires, but usually they are not traceable in online questionnaires. The definition of "changing" was "alteration of an option after having already answered one of the following items". All radio button groups, checkboxes and drop-down menus were analysed throughout the questionnaire. Furthermore, we differentiated between factual questions (e.g., questions about sex – either male or female) and opinion questions. This differentiation was necessary because changing in the context of opinion questions is more probable and comprehensible. Factual questions (see also Bassili & Fletcher, 1991) require less thought and, therefore, take less time than opinion questions. We were able to replicate this finding in the present study (detailed results omitted). Changing on opinion questions was more frequent (5.4%) than on factual questions (1.5%). In general, changing on factual questions has a more negative influence on data quality than changing on opinion questions. Therefore we coded changes on these questions with 2 and changes on opinion questions with 1. More than two changes on a particular opinion question were coded with 2 as well.

All in all for 1022 questionnaires (97.7% of all questionnaires) it was possible to judge changes. Only 5.9% of the questionnaires showed highly negative influence on data quality, mostly due to the question "Are you using IM?" (3.6%). This was the first question about the central topic (Instant Messaging). It turned out later, that many people did not read the introduction text (see next section), therefore they did not know what exactly the study was about. This could be the reason for the frequent changes on this particular question. Another peculiarity was the high number of changes on

Please cite this article in press as: Stieger, S., & Reips, U.-D. What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior* (2010), doi:10.1016/j.chb.2010.05.013



Instant Messaging - Microsoft Internet Explorer

Instant Messaging

0%

Introduction:
This study is conducted at the University of Vienna. The study will take approximately 10 minutes to complete. Data gathered will be confidential and for research purpose only.

The study is about Instant Messaging. Instant Messaging is the ability to see if a chosen friend, co-worker, or associate is connected to the Internet. If connected, you are then able to exchange "real time" messages.

Have you already participated in this study? ○ Yes ● No

Are you in contact with online research issues through your work or training? ○ Yes ● No

Are you going to fill out the questionnaire seriously? ● Yes ○ No

Submit

Fig. 3. Navigation picture of the first study page, overlapped with the picture of the questionnaire page as it was displayed in a web browser.

Table 1
Amount of behaviors with low negative and high negative influence on data quality for each analysed category and each question/page.

	Low negative influence	High negative influence	Number of questionnaires where information was available (=100.0%).
Without UAT			
Face validity of user entries	122 (13.0%)	18 (1.9%)	938
Item non-response	16 (1.7%)	61 (6.6%)	919
With UAT			
Changes in text fields	38 (4.1%)	NA	928
Changes on opinion and factual questions	529 (51.8%)	60 (5.9%)	1022
Clicking through	319 (34.3%)	427 (46.0%)	929
Longer inactivities	36 (3.6%)	NA	989
Excessive clicking	NA	62 (6.3%)	978
Excessive mouse movements	NA	115 (11.0%)	991

Note: NA = not applicable, UAT = UserActionTracer.

both semantic differentials. This is a clear indicator that matrix question formats are problematic (also see Reips, 2010).

3.1.5. Behaviors observable with the UAT: Clicking through

We strove to separate those who engaged in click-through behaviors (i.e., answering without really reading the questions) from those who did not hurry through the questionnaire, i.e., answered the questions in adequate time. We therefore set a lower bound threshold level for each item. These lower bound threshold levels were established empirically by using the average reading time of each question (including introduction texts). Reaction time is a compound of reading time and response time. We believe that using only the reading time as the lower bound threshold is appropriate, because if someone answered a question quicker than the threshold this individual cannot even have read the question's text. With the Javascript used in this study, it was possible to capture the reaction time not only for the whole questionnaire, but also for each item, allowing detailed analyses (Reips, submitted for publication). If the reaction time was below the threshold, the participant was considered as having clicked through. The threshold levels for simple questions were between 1 and 3 s, depending on the length of a question's text and on whether a switch between keyboard and mouse was necessary. For longer introduction texts, threshold levels were adjusted accordingly (e.g., reading the introduction text took 10 s; reading the introduction text for the first semantic differential took 16 s on average). Only those questionnaires with a straight sequence of answering (no moving back and forth) were analysed. The decision "clicked through" vs. "answered in adequate time" was made based on the thresholds for every answer. As clicking through definitely has a highly negative influence on data quality, all these cases were coded with 2. We also coded reading times for introduction texts. If someone was below the threshold then these cases were coded with 1.

For 929 questionnaires (88.8% of all questionnaires) it was possible to analyse clicking through. 34.3% of questionnaires showed at least once low negative influence behavior (see Table 1). In further analyses of the reaction times, we were able to determine that almost all of those who clicked through had not read the introduction text on page one in the first place (288 of 319). An astonishing 46.0% ($n = 427$) of participants showed high negative influence behavior, mostly on the semantic differentials ($n = 279$).

3.1.6. Behaviors observable with the UAT: Longer inactivities

Longer inactivities within an online questionnaire may have a number of reasons, such as participants disturbed by someone entering the room or a technical problem. Inactivity was defined as no action of any kind for at least 5 min. Because inactivity does not necessarily influence data quality in a negative fashion at least with the current questionnaire[5], these cases were coded with 1. All in all, for 989 questionnaires (94.6% of all questionnaires) it was possible to judge longer inactivity. Only 3.6% of participants showed behaviors with low negative influence on data quality (see Table 1).

3.1.7. Behaviors observable with the UAT: Excessive clicking

Another observable behavior that can be traced through paradata is excessive clicking. Excessive clicking was defined as twice as many or more clicks than necessary used for the task at hand. For this analysis the number of items and altered options of items were considered. The analysis was conducted for each questionnaire page. If a score of twice as many clicks as necessary plus two for scrolling was exceeded, the page was coded with 2. Cases with more clicks than necessary, but with fewer clicks than those coded with 2, were not treated as influential on data quality in order not to overlap with the categorization of ‘‘changing answers”, i.e., changing answers also influences the number of clicks, but not to the extent used for categorization of ‘‘excessive clicking”. In order to interpret the results unambiguously, only fully completed questionnaire pages were analysed. Among 978 questionnaires (93.5% of all questionnaires), for which it was possible to judge excessive clicking, only 6.3% of questionnaires showed high negative influence behavior on data quality (see Table 1).

3.1.8. Behaviors observable with the UAT: Excessive mouse movements

Among several available measures of mouse movements (e.g., line median point and standard deviation of mouse track; mouse pointer left the browser window), we used the overall length of the mouse track (i.e., mouse movements during the answering process) from each questionnaire page. We did not use empirically established threshold levels as in the analysis before, because the overall length is highly dependent on a number of influences (e.g., where did the mouse pointer start when the page was first loaded; does the mouse pointer follow eye movements – a behavior sometimes observed (for an example see Fig. 3)). How to deal with outliers is a difficult task (see Ratcliff, 1993), therefore we used a frequently used outlier criterion, namely excluding cases with values ±2 standard deviations around the mean (e.g., Heerwegh, 2003). Mouse movements above and below this wide range of ±2 standard deviations can safely be judged as indicating potential for low data quality and were therefore coded with 2. Among 991 questionnaires (94.7% of all questionnaires), for which it was possible to judge excessive mouse movements, only 11.0% showed high negative influence behavior on data quality (see Table 1).

3.1.9. Overall results of observed behaviors

In order to show the advantage of the UAT, we divided the table into analyses that can be performed without using the UAT and analyses that can only be performed on the basis of the data produced by the UAT. As can be seen from Table 2, more negative influence behaviors were found with the UAT (low negative influence: Wilcoxon test: $z = -20.96$, $p < .001$; high negative influence: $z = -18.30$, $p < .001$). The UAT found seven times more participants showing behaviors with high negative influence on data quality than could be found only looking at the face validity and item non-response (the standard procedure in many analyses of data from online questionnaires).

In general, combining behavioral measures with and without the UAT, 10.5% of participants showed more than five single high negative influence behaviors (see Table 2). This must be judged by taking into consideration 132 possible single behavior judgments. The participant with the highest number of high negative influence behaviors got only 20 out of possible 132 negative judgments. This participant clicked through both semantic differentials without considering any semantic dimension. His/her performance on all other questions showed no negative influence behaviors.

3.2. Research question 2: Is it possible to validate the used procedure by comparing the results with data from online address books?

Judging data quality on the basis of paradata is subjective, because we often don’t know the reasons for the observed behavior. Therefore, it is important to validate the paradata approach. In the present article we achieved this goal by calculating internal consistencies and verifying user entries with external sources. Those participants who showed more behaviors with low negative influence on data quality also showed more behaviors with high negative influence on data quality ($r = .24$, $p < .001$). Part of the online questionnaire was an item about personal user identifications that participants use in IM programs: nicknames or identification numbers. This information was also used to give participants feedback about the results of the study (Stieger & Göritz, 2006). With nicknames or identification numbers it was possible to compare demographic data entered in the questionnaire with the data stated in online address books of each IM program, if available. This procedure was used to validate participants’ sex, age, and country (for reviews of this validation procedure see Stieger & Göritz, 2006; Stieger & Reips, 2008).

It was possible to retrieve data from the online address book of 81 participants and to compare these with the data stated in the online questionnaire. It turned out that participants with more highly negative influence behaviors also more often showed suspicious address book entries for age ($r = .29$, $p = .013$) and country ($\alpha = 10\%$; $r = .19$, $p = .085$). With low negative influence behaviors no significant correlations could be found (all $ps > .11$). As there were only four cases with highly suspicious entries regarding sex in the address book among the 81 cases, the hypothesis could not be tested for sex.

4. Discussion

Our study shows that the UAT was successful in collecting highly detailed information about individual answering processes in online questionnaires. Categorizing behaviors by their possible negative influence on data quality and by five observable main behaviors (changing, clicking through, longer inactivities, excessive clicking, and excessive mouse movements) proved to be a useful strategy. Its validation was successful, as far as information given by the participants could be externally validated. However, not all participants stated the appropriate information (i.e., nicknames) or the information was not included in the online address books.

Although results may depend on the type of questionnaire (i.e., number and grouping of questions; kind of questions, e.g., matrix, forced-choice, open text-based, closed checkbox vs. radio button,

[5] In reaction time experiments or studies in which memory effects are investigated, this could be a severe problem.

Please cite this article in press as: Stieger, S., & Reips, U.-D. What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior* (2010), doi:10.1016/j.chb.2010.05.013

Table 2
Amount of behaviors with low and high negative influence on data quality for all participants.

	Low negative influence		High negative influence	
	Without UAT	With UAT	Without UAT	With UAT
Number of observed behaviors with low and high negative influence				
0	914 (87.4%)	367 (35.1%)	968 (92.5%)	498 (47.6%)
1	91 (8.7%)	273 (26.1%)	71 (6.8%)	237 (22.7%)
2	40 (3.8%)	181 (17.3%)	2 (0.2%)	109 (10.4%)
3	1 (0.1%)	91 (8.7%)	2 (0.2%)	50 (4.8%)
4	0 (0.0%)	60 (5.7%)	2 (0.2%)	42 (4.0%)
5+	0 (0.0%)	74 (7.1%)[a]	1 (0.1%)	110 (10.5%)[b]
Sum	1046 (100.0%)	1046 (100.0%)	1046 (100.0%)	1046 (100.0%)

Note: UAT = UserActionTracer.
[a] Highest count was 26 i.e., the participant with the highest number of low negative influence behaviors on data quality had 26 (out of 132 possible judgments).
[b] Highest count was 20 i.e., the participant with the highest number of high negative influence behaviors on data quality had 20 (out of 132 possible judgments).

and so on), the type of coding (low vs. high negative influence on data quality), and treatment of reaction times (threshold levels), it turned out that in general clicking through was the most frequent problematic behavior observed. Introduction texts were rarely read thoroughly and semantic differentials showed higher levels of clicking through than other questions. Furthermore, comparing the first and second semantic differentials revealed an increase in clicking through. Semantic differentials were presented in random order, meaning that the content of the questions cannot account for this effect.

As one practical application the UAT can be used to detect usability problems with questionnaire items during a pretest phase. For example one question in our study asked about “How long have you been using IM? (months)” and “On average, how long do you chat per session? (minutes)”. It turned out that the answering format was problematic. Participants had to calculate the use of IM programs in months and the average session length in minutes. This brought about a high number of changes, likely due to (1) high cognitive burden – participants had to think about the answer thoroughly to calculate a mean score and (2) the two time formats (months and minutes) may have added to the difficulty of calculating the results appropriately. The example illustrates how UAT can be used to detect “problematic items”.

Although the results may depend on idiosyncrasies of the particular online questionnaire as well as the subjectivity of the applied coding scheme, the following recommendations can safely be given. (1) Keep introduction texts as short as possible – avoid unnecessary information. Participants won’t read them if they are too long. (2) Only use matrix questions (e.g., semantic differentials) if absolutely necessary. (3) Avoid questions with high cognitive load (e.g., questions requiring calculations). (4) Don’t put all your questions on one page – rather, follow a one-page-one-item design (Reips, 2007, 2010). If these recommendations are followed, online questionnaires likely improve data quality, even when the actual setting is unknown.

The current empirical study provides an example of how data collected via the UAT can be used to judge data quality. For the current purpose we did not use all available measures that can be collected via the UAT. Further research might include for example the coordinates of clicked radio buttons to visualize the standard deviation of a radio button’s position on a computer screen across different participants (for usability reasons). Furthermore, detailed analyses of the mouse movements (e.g., speed, duration) might reveal new insights about the online behavior of users but many other research questions are potentially conceivable that can be answered using data collected by the UAT. For example, research on visual analog scales vs. other scales (e.g., Funke & Reips, 2007; Reips & Funke, 2008) can be brought to a much more detailed analysis of participant behavior, supporting the development of better models of answering behavior and measurement. To sum up, we think the UAT is an easy to implement tool which can be useful for many different purposes.

References

Adams, C., & Kleiss, J. (2003). *An overview of Ergosoft’s usability testing method*. <http://www.ergolabs.com/Use_overview.htm> Retrieved 18.3.2010.

Baravalle, A., & Lanfranchi, V. (2003). Remote Web usability testing. *Behavior Research Methods, Instruments, and Computers, 35*, 364–368.

Bassili, J. N., & Fletcher, J. F. (1991). Response-time measurement in survey research: A method for CATI and a new look at nonattitudes. *Public Opinion Quarterly, 55*, 331–346.

Batinic, B., & Bosnjak, M. (2000). Fragebogenuntersuchungen im Internet [Questionnaire studies on the Internet]. In B. Batinic (Ed.), *Internet für Psychologen* (pp. 287–317). Göttingen: Hogrefe.

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology, 55*, 803–832.

Birnbaum, M. H., & Reips, U.-D. (2005). Behavioral research and data collection via the Internet. In R. W. Proctor & K.-P. L. Vu (Eds.), *The handbook of human factors in Web design* (pp. 471–492). Mahwah, New Jersey: Erlbaum.

Buchanan, T., & Reips, U.-D. (2001). Platform-dependent biases in online research: Do Mac users really think different? In: K. J. Jonas, P. Breuer, B. Schauenburg, & M. Boos (Eds.), *Perspectives on Internet Research: Concepts and Methods*. <http://www.psych.uni-goettingen.de/congress/gor-2001/contrib/contrib/articles.html> Retrieved 18.3.2010.

Chambers, C. D., & Brown, M. (2003). Timing accuracy under Microsoft Windows revealed through external chronometry. *Behavior Research Methods, Instruments, and Computers, 35*, 96–108.

Converse, P. E. (1970). Attitudes and non-attitudes: Continuation of a dialogue. In E. R. Tufte (Ed.), *The quantitative analysis of social problems* (pp. 168–189). Reading, MA: Addison-Wesley.

Couper, M. P. (2000). Usability evaluation of computer-assisted survey instruments. *Social Science Computer Review, 18*, 384–396.

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly, 65*, 230–253.

Cugini, J., & Laskowski, S. 2001. Design of a file format for logging website interaction. In: *NIST Special Publications 500-248*. <http://www.itl.nist.gov/iad/vug/cugini/webmet/flud/design-paper.html> Retrieved 18.3.2010.

Edmonds, A. 2001. *Lucidity*. <http://sourceforge.net/projects/lucidity/> Retrieved 18.3.2010.

Edmonds, A. (2003). Uzilla: A new tool for Web usability testing. *Behavior Research Methods, Instruments, and Computers, 35*, 194–201.

Etgen, M., & Cantor, J. (1999). *What does getting WET (Web Event-logging Tool): Mean for Web usability? In: Proceedings of the Fifth Conference on Human Factors and the Web*. MD: Gaithersburg.

Funke, F., & Reips, U.-D. (2007). Messinstrumente und Skalen [Measuring devices and scales]. In M. Welker & O. Wenzel (Eds.), *Online-Forschung 2007: Grundlagen und Fallstudien* (pp. 52–76). Köln: Herbert von Halem.

Gellner, M., & Forbrig, F. (2003). ObSys – A tool for visualizing usability evaluation patterns with mousemaps. In J. Jacko & C. Stephanidis (Eds.), *Human–computer interaction: Vol. 1. Theory and practice, proceedings of the 10th international conference on HCI* (pp. 469–473). Mahwah, NJ: Erlbaum.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review, 21*, 360–373.

Hilbert, D. M., & Redmiles, D. F. (2000). Extracting usability information from user interface events. *ACM Computing Surveys, 32*, 384–421.

Hong, J. I., Heer, J., Waterson, S., & Landay, J. A. (2001). WebQuilt: A proxy-based approach to remote web usability testing. *ACM Transactions on Informations Systems, 19*, 263–285.

Ivory, M. Y., & Hearst, M. A. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys, 33*, 470–516.

Jeavons, A. (1999). Ethology and the Web. Observing respondent behavior in web surveys. *Marketing and Research Today, 28*, 69–76.

Please cite this article in press as: Stieger, S., & Reips, U.-D. What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior* (2010), doi:10.1016/j.chb.2010.05.013

Kaufmann, E., & Reips, U.-D. (2008). *Internet-basierte Messung sozialer Erwünschtheit [Internet-based measurement of social desirability]*. Saarbrücken: VDM Verlag Dr. Müller.

Nichols, E., & Sedivi, B. (1998). *Economic data collection via the Web: A census bureau case study. In: Proceedings of the survey research methods section (pp. 366–371)*. Alexandria, VA: American Statistical Association.

NIST (1999). *WebVIP: Overview*. <http://zing.ncsl.nist.gov/cifter/TheCD/WebTools/WebVIP/Readme.html> Retrieved 18.3.2010.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*, 510–532.

Reeder, R. W., Pirolli, P., & Card, S. K. (2001). WebLogger: A data collection tool for Web-use studies. *UIR technical reports* (UIR-R-2000-6). Xerox PARC.

Reips, U.-D. (1997). Psychological experimenting on the Internet. In B. Batinic (Ed.), *Internet für Psychologen* (pp. 245–265). Göttingen: Hogrefe.

Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89–117). San Diego: Academic Press.

Reips, U.-D. (2001). The Web Experimental Psychology Lab: Five years of data collection on the Internet. *Behavior Research Methods, Instruments, and Computers, 33*, 201–211.

Reips, U.-D. (2007). The methodology of Internet-based experiments. In A. Joinson, K. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford handbook of Internet psychology* (pp. 373–390). Oxford, UK: Oxford University Press.

Reips, U.-D. (2009). Schöne neue Forschungswelt: Zukunftstrends [Beautiful new world of research: Future trends]. *Nicht-reaktive Erhebungsverfahren* (pp. 129–138). Bonn, Germany: GESIS Schriftenreihe, Band 1.

Reips, U.-D. (2010). Design and formatting in Internet-based research. In S. Gosling & J. Johnson (Eds.), *Advanced internet methods in the behavioral sciences* (pp. 29–43). Washington, DC: American Psychological Association.

Reips, U.-D. (submitted for publication). Reaction times in internet-based versus laboratory research: Potential problems and a solution.

Reips, U.-D., & Funke, F. (2008). Interval level measurement with visual analogue scales in Internet-based research: VAS generator. *Behavior Research Methods, 40*, 699–704.

Reips, U.-D., & Lengler, R. (2005). The Web experiment list: A Web service for the recruitment of participants and archiving of Internet-based experiments. *Behavior Research Methods, 37*, 287–292.

Reips, U.-D., & Stieger, S. (2004). Scientific LogAnalyzer: A web-based tool for analyses of server log files in psychological research. *Behavior Research Methods, Instruments, and Computers, 36*, 304–311.

Schmidt, W. C. (2001). Presentation accuracy of web animation methods. *Behavior Research Methods, Instruments, and Computers, 33*, 187–200.

Schmidt, W. C. (2007). Technical considerations when implementing online research. In A. N. Joinson, K. Y. A. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford handbook of Internet psychology* (pp. 461–472). Oxford, UK: Oxford University Press.

Schwarz, S., & Reips, U.-D. (2001). CGI versus JavaScript: A web experiment on the reversed hindsight bias. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of internet science* (pp. 75–90). Lengerich: Pabst.

Stieger, S., & Göritz, A. (2006). Using instant messaging for Internet-based interviews. *CyberPsychology and Behavior, 9*, 552–559.

Stieger, S., & Reips, U.-D. (2008). Dynamic interviewing program (DIP): Automatic online interviews via the instant messenger ICQ. *CyberPsychology and Behavior, 11*, 201–207.

Turnbull, D. (1998). *WebTracker: A tool for understanding Web use*. <http://www.ischool.utexas.edu/~donturn/research/webtracker/> Retrieved 18.3.2010.

Voracek, M., Stieger, S., & Gindl, A. (2001). Online replication of evolutionary psychological evidence. Sex differences in sexual jealousy in imagined scenarios of mate's sexual versus emotional infidelity. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of internet science* (pp. 91–112). Lengerich: Pabst.

Wittchen, M., Schlereth, D., & Hertel, G. (2007). Indispensability effects in spite of temporal and spatial separation: Motivation gains in a sequential task during anonymous cooperation on the Internet. *International Journal of Internet Science, 2*, 12–27.