

# The privacy economics of voluntary over-disclosure in Web forms

Sören Preibusch<sup>1\*</sup>, Kat Krol<sup>2</sup>, and Alastair R. Beresford<sup>1</sup>

<sup>1</sup> University of Cambridge · Computer Laboratory · \*sdp36@cl.cam.ac.uk

<sup>2</sup> University College London · Department of Computer Science and Security Science Doctoral Research Training Centre (SECRiT)

**Abstract** The Web form is the primary mechanism to collect personal data from individuals on the Web. Privacy concerns, time spent, and typing effort act as a major deterrent to completing Web forms. Yet consumers regularly provide more data than required. In a field experiment, we recruited 1500 Web users to complete a form asking for ten items of identity and profile information of varying levels of sensitivity. We manipulated the number of mandatory fields (none vs. two) and the compensation for participation (\$0.25 vs. \$0.50) to quantify the extent of over-disclosure, the motives behind it, and the resulting costs and privacy invasion. We benchmarked the efficiency of compulsion and incentives in soliciting data against voluntary disclosure alone.

We observed a high prevalence of deliberate and unpaid over-disclosure of data. Participants regularly completed more form fields than required, or provided more details than requested. Through careful experimental design, we verified that participants understood that additional data disclosure was voluntary, and the information provided was considered sensitive. In our experiment, we found that making some fields mandatory jeopardised voluntary disclosure for the remaining optional fields. Conversely, monetary incentives for disclosing those same fields yielded positive spillover by increasing revelation ratios for other optional fields. We discuss the implications for commercial Website operators, regulators, privacy-enhancing browser standards, and further experimental research in privacy economics.

## 1 Forms on the Web

Web users have been typing data into Web forms since they were added to the HTML standard [1] in 1995. Web forms allow interactive search and retrieval requests to servers, and to submit data to the Web server for further processing and storage. The HTML5 working draft proposes richer semantics for Web forms, including typed input fields which will only accept data of a specific type, such as valid telephone numbers or email addresses. The new draft standard also recognises that Web forms may be used “for purposes other than submitting data to a server” [2]. Indeed, the use of form elements to collect data and process it locally inside the Web browser itself—typically using JavaScript—is already common on the Web today. In a few cases, data processed in this way remains

within the browser; in most cases however, processed data is uploaded to a remote server at a later point in time.

Consequently, the Web form is the primary mechanism by which companies and governments collect personal data from individuals on the Internet today. Yet relatively little is known about the behaviour and privacy attitudes of individuals when completing Web forms as part of completing a purchase or other commercial transaction on the Internet. Previous work has shown that consumers do show some reactance to data collection via forms: 25% of Web users state that they have entered false data into forms [3], and the most frequently entered incorrect data item is their name [4].

Privacy aspects of Web forms have been studied more extensively in survey research. For example, in research into *answer behaviour* to sensitive questions, it has been shown that completion rates and data quality do not differ between on-line and paper forms [5]; however open-ended questions have higher response rates when delivered on-line rather than via paper forms [6]. There remains disagreement on where to place sensitive questions in a questionnaire [7]. Both practice and academia variously advise placement at the beginning [8,9], in the middle [10], or at the end [11].

By analogy, research into user behaviour and recommendations concerning online questionnaires would also apply to transactional Web forms. However, Web forms and online surveys exhibit a number of systematic differences: first is the motivation for completing them. Whereas soliciting participation in surveys largely relies on social exchange theory, the Web form is often a means to an end (e.g., getting a Web order shipped, setting up an account), which should be motivation enough. Consequently, incentives such as money, sweets, or a lottery—common for completing questionnaires—are rarely found for transactional Web forms. Second, the visual appeal is different: the stand-alone Web form typically spans a single screen page and is as condensed as possible; a survey often continues over multiple pages and features elements such as instructions, and a progress indicator. A questionnaire is the Web page, whereas a standalone Web form is embedded into a Web page. Third, transactional Web forms often feature text input fields which prompt users for data through field labels; a questionnaire asks questions, often closed. For many psychometric instruments, a battery of items measured on a Likert-scale results in the visual appeal of a matrix of tick-boxes. Different activities are required from the user (ticking vs. typing; making a judgement vs. mentally looking up some data). The data items requested are also normally quite different, with the possible exception that surveys ask for contact details for a follow-up or a prize draw.

Completing Web forms is a time-consuming business, and therefore anything which can be done to ease the completion of a form has traditionally been considered sensible. Consequently, user experience practitioners and browser vendors have developed techniques to ease the completion of Web forms. This strand of research focuses on lowering the cognitive and mechanical effort of completing forms: label positioning (above the text field, on the left, below) and formatting (with or without trailing colon), the mechanism for indicating mandatory

fields (it is rare [12], but commendable, namely with a red asterisk), and unified text field to reduce tabbing and mouse-keyboard switching [13]. Although some practitioners have strong opinions—for instance on label formatting—it does not seem to matter as long as the user experience is consistent [14].

Autocompletion of Web forms debuted in IE4 in 1997 and was initially called ‘Form AutoFill’ [15,16]. The browser uses a combination of cues, including commonly used field names and names of previously completed fields, to match form fields across Web sites. The browser can then suggest values for form fields it has seen before, reducing the need to type the same information a second time. Today, autocompletion is limited to values typed into text fields. For one-time or sensitive entries, such as payment authorisation codes, Web form authors can prevent automatic form filling for an entire form (or parts of it), thereby mandating interactive completion of fields. Web users can configure their browsers not to store and suggest form values and delete individual values from their autocomplete suggestions.

Web site authors could also attach semantics to form fields regardless of naming: each field may have a ‘VCARD\_NAME’ attribute to draw information from the ‘Profile Assistant’, a local repository of identities [17]. For example, if a field is marked as ‘vCard.Email’, Internet Explorer suggests email addresses previously entered or stored in the Profile Assistant. Whilst the 29 names in the vCard schema cover all practically relevant contact details, the ECML ‘Field Names for E-Commerce’ defined in RFC 2706 and its successors expands on the idea of semantic annotation by introducing further field names for billing addresses and payment details [18]. Thirteen years later, in 2012, the Chrome browser was criticised for proposing yet another naming scheme to mark up semantically equivalent fields [19].

In the early 2000s, despite built-in browser support for the autocomplete feature, there was also demand for third-party tools, such as FormWhiz [20] and Gator eWallet. Gator eWallet called itself “the smart online companion” [21] and was marketed to “fill in FORMS with no typing”. In addition to helping users complete forms, it also transmitted first name, zip code and country to the GAIN Publishing advertising network and served adverts back to the user [21] until it was shut down in July 2006 [22].

The emergence of the autocomplete feature led consumers to question whether their privacy might be violated. For example, PC magazine in 1999 asked “What’s to stop a hacker from stealing your personal data [that] is popping up in the AutoComplete window”? [23, p. 108]. Despite previously entered form data being stored locally in an encrypted file, researchers were able to read out the AutoComplete suggestions by all major browser vendors [24]. In summary, Web developers were encouraged to use autocomplete because it can “collect demographic data more easily” and faster [25], the common assumption being that the form itself is a nuisance or a “pain” [26, p. 19]. Yet, we are unable to find any thorough study or analysis which demonstrates the extent to which the use of autocomplete encourages data entry on Web forms.

Seemingly obvious assumptions to do with Web form completion have shown surprising results before. In the economics of privacy, it is regularly assumed that monetary or other incentives would encourage Web users to ignore their privacy concerns when filling out Web forms [27]. Yet, recent research has found that consumers show no preference for merchants with less privacy-invasive Web order forms even when all other parameters (such as product and price) are equal [28].

*Contribution.* In summary, the behaviour of individuals, and their motivations, when providing personal data via a Web form has not been studied rigorously before. Practitioners' literature and blogs are abound with design guidelines for online forms on how to ease completion, but the advice is given without reference to any study or data. Evidence from survey research is related but not readily applicable. In this paper, we deliver what we believe is the first experimental study into Web users' behaviour when providing personal information via a form. We quantify the amount of data provided, the costs of collecting it, and the motives for voluntary over-disclosure of data. Finally, we also benchmark the efficiency of incentivised data collection against voluntary and mandatory data disclosure.

*Outline.* We briefly revisit motives for voluntary data disclosure (Section 2) before outlining our research hypotheses (Section 3) and study methodology and design (Section 4). We give some descriptive statistics on the observed form-filling behaviour (Section 5) before turning to the analysis (Section 6). Managerial implications and pathways for regulation are discussed before concluding (Section 7).

## 2 Potential motives for over-disclosure of personal information

Based on the existing literature and common sense, we briefly review potential explanations for why Web users provide more information on forms than necessary. We focus on the initial act of over-disclosure and not subsequent failure to limit access to the information after it has been disclosed, for instance because of unusable privacy controls.

*Over-disclosure by accident.* The Web user may reveal more information than requested by accident or out of negligence. The user may ignore the optional status of some of the form fields, perhaps due to the lack of visual cues, deliberately misleading cues, or because she did not read the instructions carefully. This also includes the case of not reading the field labels and typing in more data or more detailed data than is strictly required.

*Over-disclosure by proxy.* A technical mechanism, such as the autocomplete feature described earlier, or a third-party filling out a form on someone else's behalf, may result in more completed form fields than the data subject intended.

*Limit disclosure is costly.* The user may know that some form fields are not mandatory, but is unable to identify them without incurring a high cost. For example, if optional fields are not explicitly marked on the page, the user may need to submit the form with increasing amounts of personal data multiple times to determine which subset of the data is truly mandatory before the form is submitted successfully. A risk-averse user may be afraid of losing her entire form submission if she omitted a field. Reading instructions may also be viewed as prohibitively costly in terms of time or cognitive effort, and users may believe it is easier to complete all the fields.

*Building social capital.* Additional data may be provided in order to “look good” or otherwise stimulate a desired effect. It is a major driver for data disclosure on social networks while mating [29] or job hunting. The analogy to the job market is particularly pertinent in our case, as our experiment is deployed on a crowd-sourcing platform. There could also be a social norm to over-disclosure, the violation of which may hurt one’s social capital.

*Expecting a monetary return.* The Web user may expect (to qualify for) a monetary return now or in the future, whether directly in the form of a discounted price, or by receiving promotional offers. To some extent, this explains voluntary data disclosure on online social lending platforms [30].

*Expecting a non-monetary return.* In the context of online shopping, disclosing (behavioural, preference and profile) information unlocks personalisation, this in turn makes it easier to find products of interest [31]. Further, in the context of online surveys, the respondent may anticipate that answering questions—despite them not being mandatory—will shape public opinion in a manner favourable to her.

*Expecting infrastructure improvements.* “Voluntary information spillovers” [32] promise economic profits if the information is picked up by companies to innovate products that meet a yet unsatisfied demand; this might be particularly pertinent, if exploiting the information oneself is infeasible or would yield inferior results. The theory of free revealing was originally developed for intellectual property, but we see it could apply to personal information, such as health information, as well.

*Acting reciprocally.* Reciprocity is a personality trait that facilitates voluntary disclosure of personal information in the context of social exchange. In surveys, social exchange is a strong driver towards participation [33], and incentives significantly increase response rates [34], until saturation is reached [35]. The user communicates gratitude, and returns a favour by revealing non-mandatory data items.

*Acting benevolently or altruistically.* Data might be provided out of kindness for the person, or organisation behind the Web form. In this scenario, the user fills out optional fields of the form even in the absence of personal benefit. There may also be the desire to help altruistically, and this leads to the assumption that filling out the form will help.

*Personality.* The Web user’s personality may be such that she enjoys disclosing information about herself. Filling out forms is subjectively rewarding.

### 3 Research hypotheses

Our analysis is guided by seven research hypotheses which are designed to tease apart the motivations for over-disclosure as discussed in Section 2. Our hypotheses also quantify the conditions when over-disclosure will (not) occur. The seven hypotheses are:

- H1:** Web users provide more personal information than requested by a form, even though they realise there is no prospect of monetary reward for doing so.
- H2:** The base utility the Web user reaches by submitting the form does not determine the extent of over-disclosure.
- H3:** Over-disclosure of personal data is not an accident.
- H4:** Over-disclosure is costly to the user.
- H5:** Over-disclosure is not seen negatively.
- H6:** Users have good reasons to over-disclose personal information.
- H7:** Making some form fields mandatory reduces disclosure for the remaining optional fields.
- H8:** A reward for some form fields reduces disclosure for the remaining optional fields.

We motivate H2 as follows. For example, if a Website pays users \$2 for the form, the users will not disclose any more optional information on the form than if they were paid \$1. Analogously, volunteering information to a Web shop during checkout will be independent of the value of the product purchased. Experiments have also shown that the expected non-monetary benefits (e.g., personalisation) do not determine the extent of over-disclosure [36]. With regard to the exact differences in monetary incentives, we feature two different base rewards in our study (\$0.25 vs. \$0.50, Section 4.5); previous survey research found no effect on response rate [34].

## 4 Experiment methodology

### 4.1 Not a survey

In privacy economics, surveys are known to yield results with low predictive value for real-world encounters (e.g., [36]). Laboratory and field experiments produce

## About yourself

Please provide some information about yourself. Questions 5 and 6 are mandatory. All other fields are optional. There is no bonus for this HIT.

1. What is your first name?
2. Which city are you in now?
3. What is your favorite color?
4. Do you have any siblings?
5. Which of these questions are mandatory?
6. Do you expect a bonus for this HIT?
7. Is it sunny outside?
8. When did you last spend more than \$100?
9. Which browser are you using?
10. Are you in good health?
11. What is your date of birth?
12. Are you a good person?

finish and submit HIT

**Figure 1.** Web form used for the experiment; shown are the instructions used in treatment  $T_{50}$ . HIT denotes a task on the crowdsourcing platform we used. This form was embedded in a frame on the crowdsourcing platform, but no other elements were shown.

observations with better ecological (external) validity, although there may be trust biases from the ‘secure’ environment of a university laboratory. The single most important problem with survey-style methodologies is the lack of incentive compatibility when actions or preferences are stated rather than performed or expressed.

Although our design may look like a survey to the casual observer, we stress it is actually a field experiment: instead of asking whether respondents would reveal some personal information, we actually asked for those very data items. Participants did not state a willingness to disclose, but provided personal details at their discretion.

## 4.2 Form design and instructions

Figure 1 shows a screenshot of the form we asked participants to complete. We did not provide a cover story for data collection nor did we offer any explicit indication of the purpose for data collection. The form was headed “About yourself”. The instructions written above the form, detailing which questions were mandatory, differed slightly depending on the treatment administered to the participant (Table 2). We deliberately ensured that all the data collection took place on a single page and ensured that it was clear that submitting the form also finished the task for the participant. The form itself did not mention the University of Cambridge, in words or pictures.

All information was collected using text fields. We did not use drop-down lists, radio buttons, or tick boxes, even for questions soliciting a yes or no answer. All text fields had the same visual dimensions and we did not perform any input validation. Participants could enter data in any format they wished. For instance the field asking for date of birth did not require the participant to enter data in a specific way or even require that a day, month and year was present. All field labels were phrased as questions, numbered, and were edited by a native English speaker. We took care to keep questions short, and make sure they were of comparable length. In particular, we made sure that the check questions (5 and 6, Figure 1) did not stand out visually. There was no visual mark for those questions, such as an asterisk, to indicate they were mandatory.

## 4.3 Question selection

We asked participants twelve questions as shown in Figure 1. There was no randomisation in the order of the questions. Questions covered a variety of phrasing, including Wh-questions (When, Which, What...), inversions (Are you, Is it...), and with an auxiliary (Do you...).

The questions include identity-related and profile information, previously identified as sensitive personal information in another study [36]. We did not ask for data items that could directly identify an individual, such as email address or full name details.

Some questions were worded to encourage a yes or no answer (e.g., 4 and 7) or a more elaborate response (e.g., 1 and 9). Every question could be answered with



a single word or date. However, we deliberately gave respondents the opportunity to be more talkative: we expected at least some of the participants to elaborate on their yes/no answers. The questions selected fall in multiple, overlapping categories:

**identity/family** Questions 1, 2, 4, and 11 ask for information typically found on an identity card (name, city, date of birth) or relate to the family (siblings and again name, date of birth).

**profile** Questions 3, 7, 8, 9, 10, and 12 ask for profile information in a broad sense, including the user’s current spatio-temporal context. Some of these questions require a judgement (3, 10, 12), others are factual (7, 8, 9). Orthogonally, some questions relate to the respondent’s personality (namely, 3, 12), whilst others explore the technological context (9).

**check questions** Questions 5 and 6 were the only questions, which were always mandatory. These were included to check whether respondents had read and understood the instructions. Depending on the treatment, different answers were correct. Note that for the bonus check question, we intentionally asked about the participant’s expectation (“Do you expect...” rather than “Is there...”): the belief into a payment determines behaviour.

**easily verifiable** Answers to questions 2, 5, 6, and 9 are easily verifiable as they are factually right or wrong (5, 6) or can be checked by referring to meta-data (IP geo-location, HTTP request header).

**potentially verifiable** Some answers could be verified by requesting the participant submit a photo of an ID card (1, 4, 11). This is not feasible however, as such verification is against the terms and conditions of the crowd-sourcing platform.

**non-verifiable** Even with offline contact, some data items remain unverifiable (3, 7, 8, 10) by lack of omniscience.

**sensitive information** Health information (10) is considered particularly sensitive (for example, it is listed amongst the “special categories of data” in the EU Data Protection directive); Web users themselves are regularly reluctant to share financial information (8), although there could be reputation gains from disclosing spending. Besides special data, date of birth (11) is one of the data items that consumers are least willing to provide online [28].

#### 4.4 Sampling and deployment

We used Amazon’s Mechanical Turk (mTurk), a crowd-sourcing platform, to conduct our field experiment. On the mTurk, requesters like us create and publish tasks, called HITs. HITs are typically short and pay a few US cents. Workers (participants) choose the tasks they want to work on and submit their work, which is then accepted or rejected by the requester. Requesters can require qualifications for their tasks, such as location, experience or past performance of the worker, but the workers are unaware of such restrictions.

Our choice of deployment had some implications for question selection: To comply with the terms and conditions of the mTurk platform, we did not ask for

data items that could directly identify an individual, such as address details, full name, phone number or email address. This is particularly important as workers on the mTurk platform avoid tasks which contravene the terms and conditions since they might not get paid for completing these tasks.

The mTurk platform allows participants to preview the form before deciding to work on it. Consequently, we expect a few cases of non-response from participants. The ability to abandon the form after previewing it and not entering any data at all, mirrors Web users’ ability to navigate to an alternative Website if dissatisfied with the data collection practices of an operator [12].

On the mTurk, the experiment was advertised as “Short survey – fast approval” with the description “Five-minute survey with fast approval”. We decided the description should not to reveal any more detail than the title of the task. The advertisement included our requester name, “University of Cambridge”. Amongst the mTurk worker population, “survey” is the term commonly used to describe all tasks that are published by research organisations. We decided that pretending not to be a research organisation would have been deception and harmed the internal validity of our study. Potential trust biases are discussed in Section 7.

By default, mTurk lists available tasks in the order in which they are advertised. Accordingly, our task moved down the list of available tasks as time progressed and therefore became less prominent. We chose not to re-advertise the task since this would allow a worker to participate in our experiment a second time. Consequently, we cannot use participation frequency as a sensible metric for unit non-response.

Before beginning any data collection, we obtained approval for our study from the Ethics Committee at the University of Cambridge Computer Laboratory.

#### 4.5 Treatments

We piloted the form on 40 participants. No changes were necessary to the form itself, but we did learn that our initial estimate of the payment required to encourage participation (\$0.65) was overly generous, and we decreased the payments for the main study to \$0.50 and \$0.25 depending on treatment type. Note that this payment acts as a show-up fee and is unaffected by actual data disclosure.

We varied treatments by task compensation and the amount of mandatory data. The low data requirement means that only the two check questions were mandatory. In the high data requirement, weather and favourite colour were mandatory answers in addition. A  $2 \times 2$  full experimental design was used (Table 1). The instructions given on the form were amended to reflect the number of mandatory answers (Table 2). In a fifth treatment,  $T_{B25}$ , we only mandated the two check questions, but awarded an extra payment (‘bonus’) of \$0.25 to those participants who voluntarily answered the questions regarding weather and favourite colour. Each treatment was administered to 250 participants, with the exception of  $T_{50}$ , which was administered to 500 participants.

data requirement		compensation		
minimum	extra	\$0.25	\$0.50	\$0.65
high	—	$T_{\underline{25}}$ : 209	$T_{\underline{50}}$ : 445	
low	—	$T_{\underline{25}}$ : 202	$T_{\underline{50}}$ : 216	$T_{\underline{p65}}$ : 38
low	bonus for high	$T_{\underline{B25}}$ : 181		

**Table 1.** Treatments with number of valid observations. ‘p’ indicates the pilot session.

treatment	instructions
<i>all</i>	Please provide some information about yourself.
$T_{\underline{50}}$ , $T_{\underline{25}}$	Questions 3, 5, 6 and 7 are mandatory. All other fields are optional. There is no bonus for this HIT.
$T_{\underline{50}}$ , $T_{\underline{25}}$ , $T_{\underline{p65}}$	Questions 5 and 6 are mandatory. All other fields are optional. There is no bonus for this HIT.
$T_{\underline{B25}}$	Questions 5 and 6 are mandatory. All other fields are optional. You will receive a \$0.25 bonus when completing fields 3 and 7.

**Table 2.** Instructions by treatment

The treatments were deployed as batches on Thursdays and Saturdays in January and February 2012. A two-tailed Mann-Whitney U test on the two batches forming treatment  $T_{\underline{50}}$  indicates that the weekday of deployment does not affect response behaviour for sensitive data items (first name, date of birth:  $p = 0.94$ ), or non-sensitive data items (weather, favourite colour:  $p = 1.000$ ). We did not advertise or promote our experiment other than list it as a possible task or HIT on the mTurk platform. Consequently, participants self-selected to take part. We required workers to be based in the United States but placed no further restrictions on participation. Due to the location of the study, American English spelling (e.g., “favorite color”) and currency (e.g., “\$100”) were used throughout. Repeated participation was prevented.

#### 4.6 Follow-up questionnaire

We actively followed up with participants after they had submitted the form. At least one day later, they received an invitation to complete a feedback questionnaire for an additional payment of \$0.65. Response rate on the follow-up was 74%.

We reminded the participant of the original form they completed with a small screenshot, and then asked a series of twelve questions regarding their motives for participating, time spent, enjoyment, and willingness to participate

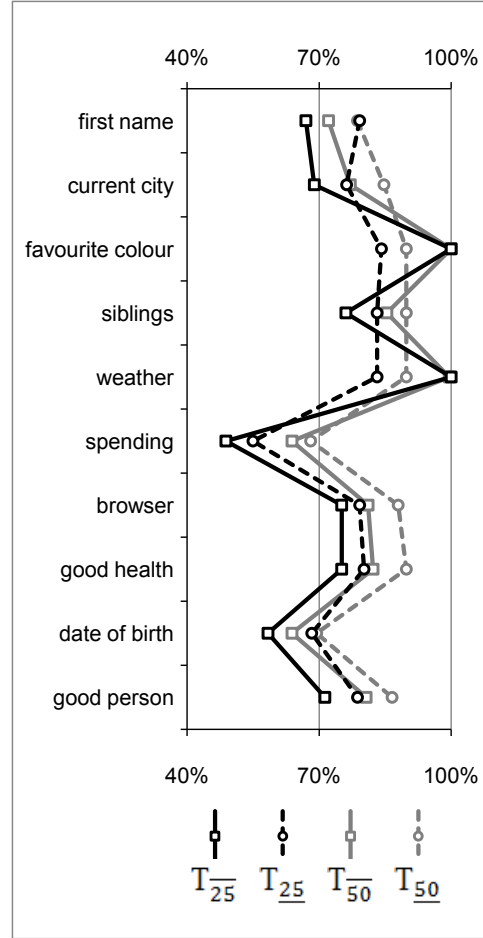
in similar study; finally, we asked them whether they revealed any personal or sensitive information, and if so, which data items were considered as such.

Depending on their original submission, we also asked the participants for their motives for (not) telling us their date of birth. We asked for expected data use, and whether they had any objections against us sharing their data with an online shop. Reciprocity as a personality trait was measured with six-item battery of pre-established reliability [37, qu. 126].

#### 4.7 Data processing and coding

All responses were manually coded by a single skilled analyst prior to analysis. We excluded all participants who had not provided correct answers to both check questions. A two-tailed Mann-Whitney U tests over all pairwise treatment combinations indicates that there the proportion of incorrect answers does not vary systematically between treatments. We checked the respondents' answers for plausibility. Because of the limitations associated with the mTurk platform, we could not fully verify their submissions. The percentage of obvious fake answers was very low, probably because response to most fields on the form was optional.

Taking the example of first name for illustration purposes, we have 1110 correct submissions across all treatments except for  $T_{B25}$ , whose 181 correct submissions we considered separately. 824 (74%) respondents provided



**Figure 2.** Proportion of respondents who provided each data item, broken down by treatment. The data items are ordered by their order of appearance in the form. Solid lines represent treatments with a high data requirement ( $T_{\bar{x}}$ ) for which favourite colour and current weather were mandatory. Black lines correspond to treatments with a low base reward ( $T_{25}$ ).

their full first name, 1% initials only, and 25% did not provide an answer. A single one respondent submitted a first name which is very likely to be fake.

For analysis purposes, we group  $T_{50}$  and  $T_{\overline{50}}$  as the high-paying (hereafter denoted collectively as  $T_{50}$ ) and  $T_{\overline{25}}$  and  $T_{25}$  as the low-paying treatments (denoted  $T_{25}$ ) respectively.  $T_{50}$  and  $T_{25}$  are grouped as the treatments with low data requirements (denoted  $T_{\star}$ ) and  $T_{\overline{50}}$  and  $T_{\overline{25}}$  as those with the high data requirements (denoted  $T_{\overline{\star}}$ ).

## 5 Descriptive statistics

Figure 2 shows the proportion of respondents who provided each data item, broken down by treatment. Across our sample, date of birth was the data item omitted most often. 57% of all submissions included full details, that is day, month, year; 68% provided parts of their date of birth. Date of birth was submitted significantly less often than the second-most often omitted data item which was first name (two-tailed paired t-test:  $p < 0.0001$ ). We will thus consider date of birth *the* sensitive item. Answers concerning weather and favourite colour were included most frequently.

99% of all respondents had JavaScript enabled in their browser; 2% were participating through a mobile device. 66% (78%) of all respondents were running the most (or second-most) recent version of their browser, as far as we could tell from the HTTP request headers (Table 3). Our sample was using more recent browsers than the general online population [38]. The four most prevalent browsers, Firefox, Chrome, Internet Explorer and Safari accounted for 97% of all observed browsers.

As part of the form, we asked participants for the browser they are using. 19% left this field blank. Amongst those who provided an answer, 96% correctly named their browser identified from the HTTP request headers. Less than two per cent provided an incorrect answer, such as “Google?”, “Windows 7” or “Word”. The remaining 2% indicated a browser that did not match the HTTP headers; from mTurk Web forums, we know that some workers use several browsers simultaneously.

Such high levels of awareness of browser type, and the use of such modern browsers, suggests that our sample may have a higher-than-usual level of computer literacy.

## 6 Analysis

### 6.1 Multivariate analysis into disclosure behaviour

In addition to analysing systematic associations between the participants psychometrics, attitudes and their disclosing behaviour (Sections 6.2 and following), we performed multivariate ordinal logistic regressions into the number of data items disclosed and the disclosure of date of birth in particular. The  $-2$ -Log-Likelihood model fitting criterion exhibited a very good significance in both cases

browser	total	current version	count by version lag				
			-0	-1	-2	-3	older
Firefox	422	9	242	27	11	6	67
Chrome	365	16	321	2	3	1	7
IE	221	9	80	105	35	1	0
Safari	68	5	63	4	1	0	0
iPad	11						
Opera	9						
Android	5						
Chromium	2						

**Table 3.** Prevalence of the eight most common browsers used by our participants (IE: Internet Explorer). 50 of the 67 participants using a Firefox older than version 6 were users of version 3.6.

( $p < 0.0001$ , Chi-square test). Treatment parameters (base reward  $T_{50}$  vs.  $T_{25}$ , data requirement  $T_{\underline{x}}$  vs.  $T_{\bar{x}}$ , presence of a bonus), response to the check questions, the browser used by the respondent, enjoyment, motives for participating, and perceiving data as sensitive or personal were used as categorical factors, plus reciprocity (negative and positive) as a metric covariate.

If the correct completion of the check questions are taken as an indicator of having read and understood the instructions, then date of birth is disclosed significantly more often when the instructions are not understood ( $p < 0.0001$ ). Disclosure decreases when the data provided is perceived as personal ( $p = 0.003$ ). Enjoying the form significantly increases disclosure ( $p = 0.002$ ). Neither reciprocity, nor any of the different motivations for participating and submitting the form are systematically associated with disclosing behaviour for date of birth.

The total number of data items provided above and beyond the check questions is also not systematically influenced by reciprocity as a personality trait. Amongst all coded motivations, only enjoyment increases disclosure weakly significantly ( $p = 0.09$ ). Again, not passing the check questions results in more data items provided ( $p = 0.02$ ). Differences between the treatments are without systematic influence. As an aside, users of Internet Explorer are more likely to fill in more data fields ( $p = 0.004$ ). Anecdotally, using an old version of a browser instead of a current one is associated with fewer data items being provided ( $p$  not significant). Higher computer literacy decreases privacy concerns [39], which could facilitate over-disclosure.

Results from regression analysis suggest that differences in payment, as the independent variable manipulated across treatments, does not impact on disclosing behaviour. There is however strong evidence for over-disclosure by accident due to not reading the instructions.

## 6.2 Hypothesis 1

In all but two treatments in our experiment, questions 5 and 6 were mandatory and all other questions were optional. In  $T_{\bar{x}}$ , questions 3, 5, 6 and 7 were mandatory and all other questions were optional. A data item is revealed significantly less often when it is optional instead of mandatory (Fisher’s exact test on favourite colour and sunny weather in treatments  $T_{\underline{x}}$  vs.  $T_{\bar{x}}$ :  $p < 0.0001$ ). Across all treatments, optional questions were answered by a significant proportion of the participants (Fisher’s exact test on full date of birth in  $T_{25}$  as the limit case of lowest disclosure:  $p < 0.0001$ ).

Hypothesis 1 is therefore *supported*. We also found strong and significant evidence for overly detailed disclosure, as discussed below (Section 6.5).

## 6.3 Hypothesis 2

For analysis, we combine all treatments with low base reward,  $T_{25}$ , and all with high base reward,  $T_{50}$ , respectively. For low sensitivity data items, the disclosure ratio<sup>3</sup> increases significantly with the base reward (weather:  $p = 0.001$ , favourite colour:  $p = 0.001$ ; G-test); it also increases in the base reward for medium sensitivity data items (good person:  $p = 0.003$ ). For date of birth, however, as the high sensitivity data item, there is no significant association of the reward level and the disclosure behaviour.

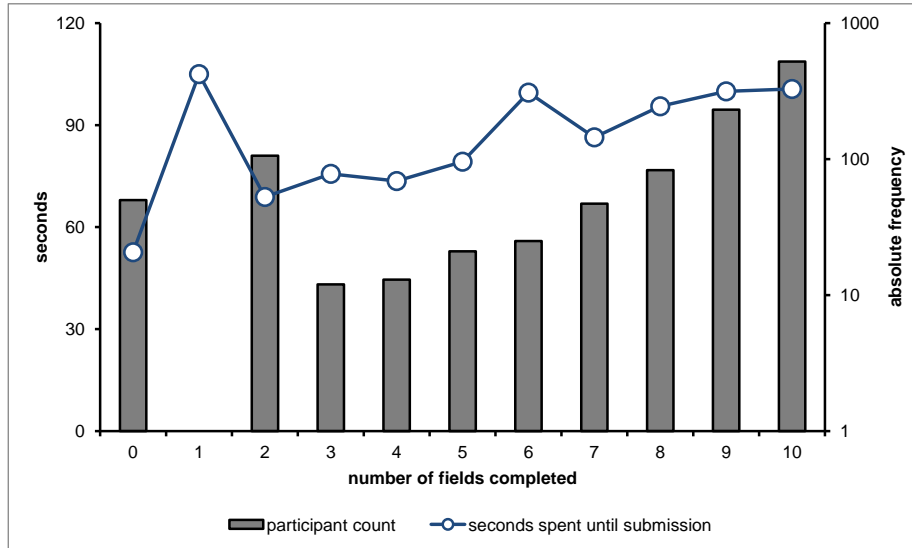
Hypothesis 2 is therefore *partially supported* for high sensitivity data items, but otherwise *rejected*. The results do not differ by high or low reciprocity as a participant’s personality trait. Note that in this experiment, with the exception of the bonus treatment  $T_{B25}$ , the base reward was independent of participants’ actual disclosure.

## 6.4 Hypothesis 3

Two check questions were built into the form, the answers to which revealed whether or not the participants had read and understood the instructions. 93% of all participants correctly identified that none of the answers (except the check questions themselves) were mandatory. We observe that over-disclosure is significantly more prevalent amongst those who did not read the instructions (date of birth:  $p < 0.0001$ , 67% vs. 87%; good person:  $p = 0.0001$ , 81% vs. 90%; Fisher’s exact test). Nevertheless, the majority of participants who understood the instructions over-disclosed (Section 6.2).

The participants knew they had disclosed personal information. In the follow-up questionnaire, 62% of all participants indicated their submission contained personal data, and 8% felt this personal data was sensitive. Without being prompted to list specific data items, 2% of all respondents to the follow-up named date of birth as a sensitive data item.

<sup>3</sup> Revelation ratio or *disclosure ratio* is the proportion of times that a given input field on a form was completed versus the total number of times this form was submitted.



**Figure 3.** Time spent on completing the form and number of participants, by number of fields completed on top of the check questions.

There is a significant positive association between completing a field in the form and indicating, in the follow-up questionnaire, that the participant felt they had revealed ‘personal’ information (date of birth:  $p < 0.0001$ , good person:  $p < 0.0001$ , weather:  $p = 0.003$ , favourite colour:  $p = 0.001$ ; G-test). Only the provision of date of birth made participants describe their submission as containing ‘sensitive’ information ( $p < 0.05$ ).

We conclude that information was, first, provided knowingly voluntarily, and second, perceived as personal. Hypothesis 3 is therefore *supported*.

### 6.5 Hypothesis 4

The user has to spend more effort completing optional fields. This effort includes both the time taken and the physical typing activity. We observe that participants who complete all instead of none of the optional fields take significantly longer ( $p < 0.0001$ , t-test). A regression analysis reveals that participants spend around 57 seconds reading the form plus additional 3.5 seconds per field completed ( $p < 0.0001$ , t-test on the regression coefficients; outlier detection based on inter-quartile range). Completion times are depicted graphically in Figure 3. Interestingly, most participants largely over-estimate the time spent on the form. 86% of respondents in the follow-up questionnaire had an estimate of the time spent that exceeded the actual time. For 13% of the participants, their estimate was more than 10 times larger than the actual time.

Participants who over-disclose also have to type more. The minimum number of characters typed in total increases linearly in the number of fields completed at



a rate of 4.0 characters per field ( $p < 0.0001$ , t-test on the regression coefficients; outlier detection based on inter-quartile range; 98% variance explained). However, the median number of characters typed in total increases quadratically(!) in the number of fields completed (99% variance explained).

As a special case of over-disclosure, we consider overly verbose answers to simple questions, taking the example of weather and spending. For this analysis, we distinguish between three levels of disclosure, again, only considering correct submissions: field left blank, simple answer provided, simple answer given with details that were not asked for. Examples for responses of the latter type include: “No. It’s currently cloudy and rainy” or “no, its cloudy and snowing” when asked for sunny weather, and “last week on textbooks” or “4 days ago getting groceries” when asked when they last spent over \$100. 6% of all participants answering the weather questions provided details that were not asked for; 14% of those indicating the time of their \$100+ purchase also indicated the purpose of spending. In both cases, the prevalence of overly detailed answers is significant ( $p < 0.0001$ , Fisher’s exact test).

From the consistent evidence, we conclude: Hypothesis 4 is *supported*.

## 6.6 Hypothesis 5

We were concerned that participants might not have over-disclosed personal information voluntarily but because they somehow felt compelled to do so. The evidence from the follow-up questionnaire indicates otherwise: From the follow-up questionnaire, we know that 98% of all participants enjoyed completing the form. Acknowledging that taking the follow-up may have introduced a sampling bias, we observe that this share corresponds to 74% of all of the original participants. 99% of all follow-up respondents (74% of the original sample) said they wanted more of these form-filling tasks in the future.

Given the high prevalence of enjoyment, more fine-grained analysis is limited by the low number of those who did not enjoy the task. We combine treatments by their data requirement and take the example of date of birth. Regardless of the data requirements, there is no significant association between enjoying the form and disclosing ( $T_{\star}: p = 0.22$ ,  $T_{\bar{\star}}: p = 0.57$ , Fisher’s exact test). Also, enjoyment as a motivation to participate is not systematically associated with over-disclosure ( $T_{\star}: p = 0.17$ ,  $T_{\bar{\star}}: p = 0.90$ , Fisher’s exact test). The non-significant trend shows that enjoyment was more prevalent amongst those who provided their date of birth.

Hypothesis 5 is therefore *supported*.

## 6.7 Hypothesis 6

Using an open-ended question, we asked participants in the follow-up questionnaire why they had completed the form in the first place. The free-text answers were coded into three reasons for each respondent. 54% had participated for the money. This comes as no surprise since we recruited our participants on a crowd-sourcing platform. 30% said the form looked easy, so they filled it in.

15% indicated they had participated out of joy; 25% because it was interesting. Original responses include: “I enjoy filling out surveys”, “I enjoy doing surveys as a way to destress [sic]” or “It looked interesting, fun and easy to do”. This is opposed to the received wisdom that form-filling is a nuisance. To have their opinion heard or to help research was named by 3% and 8% of the follow-up respondents respectively. Exemplary answers include: “I think it’s really cool to be part of a statistic analysis, to contribute my thoughts and experiences to a collective body of information”, “my information goes towards creating a change in something”, “the opportunity to present an underrepresented demographic (conservatives, mothers) in surveys” or “I like taking surveys to get my opinions heard”. This is in line with the motives for over-disclosure identified in Section 2. Those being motivated by helping research named both the researcher and research per se, such as: “I enjoy helping researchers”, “Any help I can be for research, I am glad to do” or “I appreciate helping (even if only a little) with research”.

These motives work towards completing specific fields in the form (e.g., for opinion shaping) or completing more fields on the form (e.g., when wanting to help research). They can be understood as antecedents for over-disclosure. The high proportion of participants motivated by the base reward or the form itself, combined with the low proportion of those who used the form to express opinions (2%), or were motivated by trust in the university (4%), is promising as it hints that our results might have more general validity. A G-test for each treatment individually and for all treatments together indicates that over-disclosing date of birth, being a good person and favourite colour, does not depend on whether participants were motivated by money or not.

Respondents who provided more data items (nine or ten versus eight and below completed fields) also enjoyed the form more, although the direction of the relationship remains to be determined (approaching significance:  $p = 0.06$ , G-test).

We also saw evidence that participants were motivated by the other reasons named under Section 2, to disclose more personal data than required. We report anecdotally some of the reasons given for disclosing date of birth. Respondents abided by social or self-imposed norms to submit a full form: “I feel a certain obligation to completely fill out surveys”, “A completionist [sic] instinct”, “because it was asked”, “Completeness”, “i felt that i should complete all aspects [sic]”, “I like to fully comply with requests”. The last example in particular indicates that the socially desirable behaviour for an optional field may be to complete it rather than skip it. The desire to be helpful (in an altruistic or reciprocal sense) was also frequently reported.

Entering the data also happened habitually (“Force of habit”, “Habbit [sic]”, “habit”, “I probably automatically put it down without thinking”). Many respondents reported they did not know, why they had provided their date of birth, which indicates the lack of a conscious decision (or deliberation)—a strong sign of a habit. We even observed cases of extroversion (“I have a unique birthday, it being on christmas [sic], so i just wanted to share”).

Some participants anticipated future payoffs and intended to improve their social capital on the platform (“to boost my mturk hit approval rate”, “Even though it was optional, I thought that if I did not disclose the information, you would be unable to classify me for future HITs and I would miss out on the opportunities”).

From the combined evidence, we conclude: Hypothesis 6 is *supported*.

## 6.8 Hypothesis 7

We compare data disclosure in treatments  $T_{\star}$  and  $T_{\bar{\star}}$  to see if the mandatory revelation level makes a difference for the disclosure of the remaining, optional fields. In  $T_{\bar{\star}}$ , weather and favourite colour were mandatory; they are two data items revealed voluntarily most often in  $T_{\star}$ .

Making two low-sensitivity fields mandatory decreases the revelation ratio for the high sensitivity item date of birth ( $p < 0.02$ , G-test) as well as for the medium sensitivity item of being a good person ( $p < 0.04$ , G-test).

We now only consider participants who provided answers to questions 3 and 7, regardless of whether they were in a treatment where these questions were mandatory or optional. Amongst this cohort, disclosure behaviour for the remaining fields depended on whether questions 3 and 7 were marked as mandatory or optional. The average number of fields completed when questions 3 and 7 were marked as mandatory is reduced by about 1.3 fields in  $T_{\bar{\star}}$  compared to  $T_{\star}$  ( $p < 0.0001$ , two-tailed t-test). The data therefore suggests that as the number of mandatory fields in a form is increased, the the total number of completed fields reduces. Also negatively affected is the revelation ratio for date of birth: fewer participants are willing to disclose in  $T_{\bar{\star}}$  than in  $T_{\star}$  ( $p < 0.0001$ , G-test).

Hypothesis 7 is therefore *supported*.

## 6.9 Hypothesis 8

We benchmark incentivised disclosure against voluntary and mandatory disclosure by comparing  $T_{B25}$  with treatments  $T_{25}$ ,  $T_{\bar{25}}$  and  $T_{\bar{50}}$ .  $T_{25}$  is the limit case for  $T_{B25}$  respondents who do not accept the incentive;  $T_{\bar{50}}$  is the limit case for those who choose to collect the incentive through extra disclosure.

Incentives yield the same disclosure ratio as mandatoriness ( $T_{B25}$  vs.  $T_{\bar{\star}}$ :  $p = 0.55$ , Fisher’s exact test). They improve disclosure for fields that are optional ( $T_{B25}$  vs.  $T_{\star}$ :  $p < 0.0001$ , Fisher’s exact test).

To our surprise, we do not see evidence for crowding-out of incentives, whereby a monetary incentive would replace the intrinsic motivation and result in an overall lower inclination to cooperate. On the contrary, there is strong evidence for crowding-in. When comparing  $T_{B25}$  and  $T_{\bar{50}}$ , we find that incentives for disclosing low sensitivity data also increase disclosure for the remaining, optional, medium, and high sensitivity fields on the same form (good person:  $p = 0.002$ , date of birth:  $p < 0.001$ ; Fisher’s exact test).

Hypothesis 8 is therefore *rejected* and we find significant evidence for the opposite relationship.

## 7 Summary and discussion

Forms are ubiquitous on the Web. They are the primary mechanism used to collect personal information relating to one's identity or profile. They are the metaphor for the explicit invasion of privacy.

The received wisdom is that completing Web forms is a nuisance. User experience practitioners have argued for various styles to make the form filling exercise more comfortable. However, their design recommendations do not appear to have been backed by empirical evidence. At the same time, browser vendors and add-on programmers have eased the mechanics of form filling, in particular through the autocomplete feature.

In privacy economics, we are interested in two aspects of filling in forms: the time spent (including mechanical effort) and the invasion of privacy. The traditional assumption is that Web users complete as few fields as possible on a Web form to reduce their privacy exposure and save on typing. We challenge this assumption with the first field experiment into the prevalence and extent of voluntary over-disclosure on Web forms. The empirical evidence gives a consistent picture.

Firstly, over-disclosure is a common occurrence and this is no accident. Across all levels of sensitivity, Web users provide data items for which they know disclosure is optional and not rewarded. In doing so, they reveal information subjectively considered as personal and they incur significantly costs in terms of typing effort and time spent on the form.

Secondly, Web users have good reasons for disclosing personal data despite the negative side-effects. These motives include well-being by abiding to social norms and one's personality, reciprocity, shaping public opinion, and also the build-up of social capital. A base reward, independent of and not systematically associated with disclosure, remains the strongest driver for submitting the form at all.

Thirdly, for Website operators, optional fields deliver a good data return, even for sensitive data items, which may explain why we still find them on the Web. Operators should be cautious, however, that increasing revelation ratios by making fields mandatory can backfire, because it jeopardises voluntary disclosure for the remaining fields on the form. A better approach are incentives for voluntarily provided optional data. Rewards for extra disclosure have the added benefit of crowding-in, stimulating further disclosure on the form beyond the incentive.

### 7.1 Recommendations and managerial implications

The implications for industry are quite profound. The single most important message is to mandate fewer fields. More mandatory fields mean less voluntary data disclosure whatever the sensitivity of a data item. Optional fields yield a good data return. A company that wants to extract many details should use optional fields but not highlight them. When users are unsure whether a field is optional or mandatory, users would rather fill in a field than skip it.

Next, Web sites should capitalise on Web users' motives for voluntary disclosure. It can be helpful to frame data collection as a social exchange rather than an economic exchange. Our results (H2) also suggest that the privacy-friendly opt-in into personalisation features is viable. The increased base reward from a personalised service can stimulate further disclosure on low and medium sensitivity data items. We also recommend the use of free text input fields even for yes/no answers: they allow fine-tuned hiding and over-disclosure alike, thereby appealing to both privacy concerned and unconcerned users. Free text fields give opportunity to talk so that customers service can learn issues and needs.

We add to the policy debate with the following ideas for regulation and for browser behaviour as de-facto standards. In line with current data protection legislation, Web site operators should make the purpose of data collection explicit. Otherwise, Web users come up with their own good reasons for providing personal information, typically resulting in over-disclosure. Regulators assessing the privacy invasion of a Web form should be aware that an optional field is often as privacy-invasive as a mandatory field.

Browsers can help users to limit their flow of personal information. With HTML5, there is now an attribute to distinguish optional fields from mandatory fields. The browser could blur optional fields, delay or disable autocomplete for optional fields, and warn the user if a form submission contains optional fields.

In the meantime, educating Web users to identify form fields as optional is crucial so they can spot opportunities for data hiding. One of our participants reported in the follow-up: "I will be much more careful in the future about giving out my personal information. Thank you for this very important lesson that I have learned."

To academics across disciplines using Web forms, for instance in a surveys or exit-questionnaires, we also recommend re-assessing the privacy invasion of optional fields. Further, privacy was salient in our setup: two third of our respondents were aware they had submitted personal information. We also take the opportunity to reiterate good practice in field experimentation: test thoroughly, pilot, monitor and be open to receive feedback from your participants.

We also caution researchers in privacy economics to prepare their control treatments carefully. The voluntary over-disclosure of personal information warrants further research into privacy-friendliness as a desirable property! Web users' preference to reveal personal data could be so strong that they prefer privacy-invasive alternatives over privacy-friendly alternatives—a serious threat to the validity of control treatments. Given the ambiguous role of optional fields, we also recommend experimenters consider making all fields mandatory.

## 7.2 Limitations and future work

We are aware of the limitations of our findings. They are of three kinds.

Firstly, our results may exhibit a *trust bias* that originates from our university status rather than as an unknown commercial entity; having said this, only a minority of respondents (4%) named trust in the university as a driver for participation. Our results may not generalise to other transactional Web

forms found in electronic commerce or online social networking. We knowingly incurred this limitation in external validity for the sake of internal validity. We also emphasise that well-known retailers or social networks may benefit from similar trust biases resulting from brand effects. Indeed, a trust bias in favour of our experiment may have originated in good ratings on relevant mTurk forums rather than in our status as a university.

Secondly, owing to the deployment specifics of our field experiment, we were limited in our ability to perform *data verification*. We checked users' responses regarding their Web browsers and found truthful reporting for the overwhelming majority of participants. We planned to verify participants' answers for their current city with the location returned by geo-IP. However, in the end we did not do so, as we were unable to obtain detailed knowledge of nested geographical areas. We inspected all other data items for syntax correctness and plausibility. We note that a commercial Website is similarly handicapped in its ability to test the accuracy of users' personal information, but the motives for voluntary over-disclosure work against lying. Also, misreporting only partially affects the economics of over-disclosure: the typing effort for an answer is independent of its truthfulness. Still, as future work, we are currently considering mechanisms to enforce truthful reporting or at least assess the prevalence of misreporting.

We performed analyses only on participants who had successfully passed both check questions that tested for understanding the instructions and the optionality of the data items explained therein. We acknowledge that this results in underestimating the extent of over-disclosure: the 14% of participants who did not answer the check questions correctly, and supposedly ignored the voluntariness of disclosure, were strongly significantly more likely to over-disclose.

Thirdly, we acknowledge a potential *sampling bias*. By deploying on mTurk, it is possible that we only recruited form-lovers: the skills and the mindset of mTurk workers may be such, and they are trained to complete forms quickly. When optimising for speed, uniformity and thereby over-disclosure may be more desirable than time-consuming, selective disclosure. However, we carefully checked compliance with instructions and removed participants from our analysis who had not passed the check questions. The mTurk platform does not provide access to participant statistics, such as the number of previously completed tasks for each worker, which could have been moderating variables. We considered requesting those details in the follow-up questionnaire, however, in the end, we decided other questions were more important given our budget constraints and the need for a high response rate (and therefore short) follow-up questionnaire.

Workers on the mTurk platform may deliberately over-disclose to increase their chance of future working opportunities. In our case, the participant might have believed the Web form was the first one in a series. In the exit questionnaire, we did ask participants for their motivation and found only limited evidence for signalling behaviour. We further notice that the quest for future tasks cannot explain over-disclosure at the level of detail, and that most mTurk workers are trained for compliance rather than volunteering personal data. Reputation build-

ing on mTurk works mainly for and against the requester whose tasks are chosen (or ignored) by the worker population.

Our participants' submissions indicated privacy concerns. We also consider our sample more representative of the Western online population than a convenience sample of computer science or psychology undergraduates. All in all, we are therefore confident that our findings generalise beyond our sample. In particular, the following findings should hold beyond the mTurk environment: the relative magnitudes of revelation ratios; the moderating factors (or their lack of influence) for base reward, personality and signals such as the browser used; the effects from incentives and mandatory fields; and our estimates of typing effort and time spent.

Exploring the effects of aforementioned limitations are one possible strand for future research. Other promising avenues for future work include: How is over-disclosure affected by the number of fields on the form? Will over-disclosure persist when the penalties are increased beyond time, effort and sensitivity? As a priming/salience effect, does the explicit mention of data collection purposes impact disclosure ratios? Is form filling affected by the number of pages on which form fields are spread out? How do past experiences with forms, including misuse of data entered into them, affect disclosing behaviour in the long run? Privacy economics meet usability research is the ultimate conundrum for privacy advocates: why is it so easy to collect Web users' personal information?

## Acknowledgements

The authors wish to thank Google for providing financial support for this work through a Focused Research Award. Sören Preibusch and Kat Krol were supported by a Volkswagen Foundation travel grant to present this work at WEIS 2012.

hyp.	support	operationalisations	signif./prop.
H1	supp.	optional revealed less often than mandatory	$p < 0.0001$
		optional revealed more often than necessary	$p < 0.0001$
H2	partial	discl. ratio increases in base reward (low sensit.)	$p = 0.001$
		discl. ratio increases in base reward (medium sensit.)	$p = 0.003$
		discl. ratio increases in base reward (high sensit.)	n.s.
H3	supp.	not reading instructions and over-disclosure	$p = 0.002$
		over-disclosed subjectively personal information	$p = 0.003$
H4	supp.	over-disclosure is time-consuming	$p < 0.0001$
		overly detailed disclosure is prevalent	$p < 0.0001$
H5	supp.	enjoying the form	n/a: 98%
		wanting more of such forms	n/a: 99%
		over-disclosure and enjoying less	$p = 0.19$
		over-disclosure and less motivated by joy	$p = 0.32$
H6	supp.	motivated by joy	n/a: 15%
		motivated by interest	n/a: 25%
		motivated by ease	n/a: 30%
		motivated by opinion shaping opportunity	n/a: 2%
		motivated by contributing to science	n/a: 8%
		motivated by monetary prospects	n/a: 54%
H7	supp.	motivated by trust in university	n/a: 4%
		medium sensit. revelation ratio	$p < 0.04$
		high sensit. revelation ratio	$p < 0.02$
		average number of fields completed	$p < 0.0001$
H8	rej./opp.	high sensit. revelation ratio given compliance	$p < 0.0001$
		incentives increases disclosure for trigger fields	$p < 0.0001$
		incentives increases disclosure for remaining fields	$p = 0.002$

**Table 4.** Summary of findings from the field experiment: overview of supported and rejected hypotheses and operationalisations used. The worst significance level is reported if the same operationalisation applied to several data items of different sensitivity (sensit.); no significance levels are reported for operationalisations based on pure occurrence.



## References

1. Berners-Lee, T., Connolly, D.: Hypertext Markup Language - 2.0. <http://tools.ietf.org/html/rfc1866> (1995)
2. W3C, Ian Hickson (ed.): HTML5. A vocabulary and associated APIs for HTML and XHTML, Section 4.10 Forms. <http://www.w3.org/TR/html5/forms.html> W3C Working Draft 25 May 2011.
3. BITKOM: Jedes vierte Mitglied flunkert in sozialen Netzwerken. [http://www.bitkom.org/de/presse/70864\\_67989.aspx](http://www.bitkom.org/de/presse/70864_67989.aspx) (2011)
4. BITKOM: 12 Millionen Deutsche machen Falschangaben im Web. [http://www.bitkom.org/62107\\_62102.aspx](http://www.bitkom.org/62107_62102.aspx) (2010)
5. McCabe, S.E., Boyd, C.J., Young, A., Crawford, S., Pope, D.: Mode effects for collecting alcohol and tobacco data among 3rd and 4th grade students: A randomized pilot study of web-form versus paper-form surveys. *Addictive Behaviors* **30**(4) (2005) 663–671
6. Denscombe, M.: Item non-response rates: a comparison of online and paper questionnaires. *International Journal of Social Research Methodology* **12**(4) (2009) 281–291
7. Tourangeau, R., Yan, T.: Sensitive questions in surveys. *Psychological Bulletin* **133**(5) (2007) 859–883
8. Frick, A., Bächtiger, M.T., Reips, U.D.: Financial incentives, personal information and drop-out rate in online studies. In: *Current Internet science. Trends, techniques, results.* (1999)
9. Acquisti, A., John, L.K., Loewenstein, G.: The impact of relative standards on the propensity to disclose. *Journal of Marketing Research (JMR)* **49**(2) (2012) 160–174
10. European Commission: EUROPA - EuropeAid - Evaluation - Guidelines: How is a questionnaire developed? [http://ec.europa.eu/europeaid/evaluation/methodology/egeval/tools/too\\_qst\\_how\\_qst\\_en.htm](http://ec.europa.eu/europeaid/evaluation/methodology/egeval/tools/too_qst_how_qst_en.htm) (2005)
11. NHS Wirral: Questionnaire design tips. [http://www.wirral.nhs.uk/document\\_uploads/Governance/QuestionnaireDesignTips.pdf](http://www.wirral.nhs.uk/document_uploads/Governance/QuestionnaireDesignTips.pdf) (2010)
12. Preibusch, S., Bonneau, J.: The privacy landscape: product differentiation on data collection. In: *The Tenth Workshop on the Economics of Information Security (WEIS).* (2011)
13. Anthony T.: Why users fill out forms faster with unified text fields. <http://uxmovement.com/forms/why-users-fill-out-forms-faster-with-unified-text-fields/> (2011)
14. Jarrett, C., Gaffney, G.: *Forms that work: designing web forms for usability.* Morgan Kaufmann (2008)
15. Microsoft: How to Use the AutoComplete Feature in Internet Explorer 4. <http://support.microsoft.com/kb/171230> (2007)
16. Microsoft: How to use the AutoComplete feature in Internet Explorer 5 and 6. <http://support.microsoft.com/kb/217148> (2007)
17. Microsoft: Using AutoComplete in HTML Forms. <http://msdn.microsoft.com/en-us/library/ms533032.aspx> (2008)
18. Eastlake, D., Goldstein, T.: ECML v1: Field Names for E-Commerce. <http://tools.ietf.org/html/rfc2706> (1999)
19. Braun, H.: Chrome hilft beim Formular-Ausfüllen. <http://heise.de/-1422502> (2012)

20. Philippot, P., Canter, S.: Fill in web forms automatically. *PC Magazine* **18**(16) (1999) 205
21. GAIN Publishing: Gator.com - Home. <http://web.archive.org/web/20031031020937/http://www.gator.com/home2.html> (via Internet Archive) (2003,2012)
22. GAIN Publishing: Gator.com - Home (Important information about GAIN software). <http://web.archive.org/web/20060630073649/http://www.gator.com/home2.html> (via Internet Archive) (2006,2012)
23. Rubenking, N.J.: Autocomplete for web forms—is it safe? *PC Magazine* **19**(3) (2000) 105–108
24. van Eitzen, C.: Auto-complete: browsers disclose private data - update. <http://h-online.com/-1043122> (2010)
25. Microsoft: Collect Demographic Data More Easily with Internet Explorer 5. <http://msdn.microsoft.com/en-us/library/bb250414.aspx> (1999)
26. Wroblewski, L.: Web form design: filling in the blanks. Rosenfeld Media (2008)
27. Taylor, D., Davis, D., Jillapalli, R.: Privacy concern and online personalization: The moderating effects of information control and compensation. *Electronic Commerce Research* **9** (2009) 203–223
28. Beresford, A., Preibusch, S., Kübler, D.: Unwillingness to pay for privacy: A field experiment. IZA Discussion Papers 5017, Institute for the Study of Labor (IZA) (June 2010)
29. Ellison, N.B., Steinfield, C., Lampe, C.: The benefits of Facebook “friends:” social capital and college students’ use of online social network sites. *Journal of Computer-Mediated Communication* **12**(4) (2007) 1143–1168
30. Böhme, R., Pötzsch, S.: Privacy in online social lending. In: Proc. of AAAI Spring Symposium on Intelligent Information Privacy Management. (March 2010)
31. Personalization Consortium: Personalization & privacy survey. <http://personalization.org/SurveyResults.pdf> (2000, 2005) via Internet Archive.
32. Harhoff, D., Henkel, J., von Hippel, E.: Profiting from voluntary information spillovers: how users benefit by freely revealing their innovations. *Research Policy* **32**(10) (2003) 1753–1769
33. Helgeson, J.G., Voss, K.E., Terpening, W.D.: Determinants of mail-survey response: Survey design factors and respondent factors. *Psychology and Marketing* **19**(3) (2002) 303–328
34. James, J.M., Bolstein, R.: The effect of monetary incentives and follow-up mailings on the response rate and response quality in mail surveys. *Public Opinion Quarterly* **54**(3) (1990) 346–361
35. Schupp, J., Kroh, M.: Incentives and response rates – experience from the SOEP-innovation-sample 2009. In: 4th Conference of the European Survey Research Association. (2011)
36. Spiekermann, S., Grossklags, J., Berendt, B.: E-privacy in 2nd generation e-commerce: privacy preferences versus actual behavior. In: EC '01: Proceedings of the 3rd ACM conference on Electronic Commerce, New York, NY, USA, ACM (2001) 38–47
37. TNS Infratest Sozialforschung: Living in Germany: Survey 2005 on the social situation of households (individual question form). Technical report, SOEP, DIW Berlin (2005)
38. Net Applications.com: Desktop browser version market share (January, 2012). <http://marketshare.hitslink.com/browser-market-share.aspx?qprid=2&qpcustomd=0&qptimeframe=M&qpsp=156> (2012)

39. Dinev, T., Hart, P.: Internet privacy, social awareness, and internet technical literacy—an exploratory investigation. In: Proceedings of the 17th Bled eCommerce Conference. (2004)