APPLIED COGNITIVE PSYCHOLOGY
*Appl. Cognit. Psychol.* **22**: 51–68 (2008)
Published online 12 February 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/acp.1331

# Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times

TING YAN[1,2]* and ROGER TOURANGEAU[1,2]

[1] *Joint Program in Survey Methodology, University of Maryland, USA*
[2] *Survey Research Center, University of Michigan, USA*

SUMMARY

This paper examines response times (RT) to survey questions. Cognitive psychologists have long relied on response times to study cognitive processes but response time data have only recently received attention from survey researchers. To date, most of the studies on response times in surveys have treated response times either as a predictor or as a proxy measure for some other variable (e.g. attitude accessibility) of greater interest. As a result, response times have not been the main focus of the research. Focusing on the nature and determinants of response times, this paper examines variables that affect how long it takes respondents to answer questions in web surveys. Using the survey response model proposed by Tourangeau, Rips, and Rasinski (2000), we include both item-level characteristics and respondent-level characteristics thought to affect response times in a two-level cross-classified model. Much of the time spent on processing the questions involves reading and interpreting them. The results from the cross-classified models indicate that response times are affected by question characteristics such as the total number of clauses and the number of words per clause that probably reflect reading times. In addition, response times are also affected by the number and type of answer categories, and the location of the question within the questionnaire, as well as respondent characteristics such as age, education and experience with the Internet and with completing web surveys. Aside from their fixed effects on response times, respondent-level characteristics (such as age) are shown to vary randomly over questions and effects of question-level characteristics (such as types of questions and response scales) vary randomly over respondents. Copyright © 2007 John Wiley & Sons, Ltd.

The collection of response times (RT), or response latencies, is one of the most important means for investigating hypotheses about mental processing (see van Zandt, 2002). Since the middle of the 19th century when the Dutch psychologist F. C. Donders conducted his pioneering work on response times (Donders, 1868), experimental and cognitive psychologists have routinely collected response time data. Even though response times have not been used extensively in survey research, their potential as a measure of 'the amount of information processing necessary to answer a question' (Bassili & Scott, 1996) is increasingly catching the attention of survey researchers, especially since these data are easily captured as one type of paradata in computer-assisted interviews. Paradata are automated data generated directly by the survey data collection process and can be used to

*Correspondence to: Ting Yan, ISR—University of Michigan, 426 Thompson Street, Room 4055, Ann Arbor, MI 48106, USA. E-mail: tingyan@isr.umich.edu describe and evaluate that process (Couper, 2000). Examples of paradata include keystrokes in computer-assisted interviews and time stamps and mouse clicks in web surveys.

Typically, the distribution of response times is not normal, but asymmetrical and positively skewed. In addition, response times from a single individual are generally not independent of each other (Ratcliff, 1993; van Zandt, 2002). Therefore, the analysis of response time is usually carried out at the level of means 'with hypotheses formulated with regard to predicted average increases or decreases in RT' (van Zandt, 2002). Despite this, there are at least two reasons that means are not necessarily the best unit of analysis for response time data. First, because response times are generally skewed, the mean is not necessarily the best estimate of the middle of the distribution. A related problem is that the mean is sensitive to outliers, which are virtually inevitable in response time measures. van Zandt and Ratcliff both suggest alternative measures of central tendency to characterise response time distributions. They both recommend such estimators as trimmed means, harmonic means and medians. The three measures of central tendency are generally highly correlated and they usually yield similar conclusions.[1]

## SURVEY RESPONSE TIMES

Roughly speaking, survey use of response times falls into one of the three broad categories: testing of theories, pretesting and diagnosis of response problems and investigation of web survey methodologies. Most of the work in the first category follows Fazio's 'attitudes as object-evaluation association' model (Dovidio & Fazio, 1992; Fazio, 1990; Fazio, Sanbonmtsu, Powell, & Kardes, 1986). According to this model, the strength of an attitude is represented as the strength of the association between an attitude object and the evaluation of that object. This association strength determines the accessibility of the attitude and the likelihood that it will be activated upon mere observation of that attitude object. Fazio's work (Fazio et al., 1986) has measured attitude accessibility by the response time from stimulus onset (typically, a single word, like 'cockroach') to a response (pressing a key); this is thought to measure only the retrieval of an evaluation from memory.

Building on Fazio's work, Bassili and Fletcher (1991) published the first paper to examine response times in a survey context and related response times to the stability of the answers. They found that 'movers' (respondents who changed their minds when confronted with a counterargument) took longer to respond to questions than respondents who did not change their views when challenged. This finding was replicated by Heerwegh (2003) in a web study, who found that respondents with less stable attitudes needed more time to respond to an attitude question than those with more stable attitudes. Furthermore, respondents who did not know the answer to a knowledge question also took longer to respond. Thus, both unstable attitudes and lack of knowledge tended to result in longer response times (Heerwegh, 2003). These findings supported the distinction first proposed by Converse (1970) between reporting an existing attitude (or answer) versus improvising one on the spot (see also Strack & Martin, 1987).

Also consistent with Fazio's model, Bassili found response times to be a better predictor of discrepancies between voting intentions and voting behaviour than other indices of attitude strength, such as certainty or self-reported attitude strength (Bassili, 1993).

[1] Our initial analysis showed that the correlations among the three central measures are in the 0.90 s for our data.

DOI: 10.1002/acp

Tourangeau, Rasinski, and D'Andrade (1991) timed respondents as they answered attitude questions about abortion and welfare and found that responses to agree/disagree items were faster when an item followed another item on the same topic. In contrast to Fazio's model, Tourangeau and his colleagues argued that answers to attitude questions are based on considerations (beliefs, feelings, values related to the issue) that are retrieved and integrated when the question is administered. Because the considerations related to an issue are organised by topic, answering one attitude question on a specific topic increased the accessibility of considerations related to that topic, speeding up the retrieval process (and overall response times) when respondents answered another question on the same topic. Respondents were also faster when they answered an earlier item that concerned a different topic related to the same issue, but the increase in speed was greater when the prior item was about the same topic.

The second line of research using response times in surveys involves diagnosing response problems with draft survey questions. For instance, Bassili and his colleagues have used response time data to identify bad questions. Their results indicate that poor questions take longer to answer than good ones, demonstrating the feasibility of using average response times as an indicator of question problems (Bassili, 1996; Bassili & Scott, 1996). Draisma and Dijkstra (2004) demonstrated that nonsubstantive answers produce the longest response times, followed by incorrect answers, and correct answers; these results also suggest that response times are an indicator of uncertainty and response error.

The final use of response time data in surveys has been to explore various issues in web survey design. Web surveys provide a rich array of paradata captured from either the server on which the web survey resides or the respondent's computer. Server-side paradata record activities at the server level, such as 'visits' to each web page and time stamps. Client-side paradata record activities on the respondents' computer and records such behaviours as clicking radio buttons or changing answers (see Heerwegh, 2003). Response times can be collected from both the server and the respondent's computer. Server-side response times show the elapsed time from the moment the server delivers a survey question to a respondent's computer to the moment when it receives an answer from the respondent. By comparison, client-side response times include the elapsed time from when a survey question is fully displayed on respondent's computer to when an answer is sent. Therefore, client-side response times do not include the added downloading and uploading time that are included in the server-side response times.

Both Heerwegh and Tourangeau and his colleagues have used response time data to study the effects of different question or response formats in web surveys. For example, in one study, Heerwegh examined response times for two different response formats (radio buttons vs. drop-down boxes) and found faster response times with radio buttons (Heerwegh, 2002). Tourangeau, Couper, and Conrad (2004) showed that respondents answered more quickly when the response options followed a logical order from top to bottom than when the options did not follow a logical order.

## PROBLEMS WITH RESPONSE TIMES

We see two major problems in the current research using response time data. The first has to do with how to measure response times. The second has to do with how to analyse response time data.

Two approaches have been used in measuring response times in the survey literature—active timers and latent timers. The first approach includes interviewer-activated timers and voice-activated timers, both of them used with computer-assisted telephone interviews. With interviewer-activated timers, interviewers are supposed to press a key to start the timer as soon as they finish reading a question and press the key again as soon as respondents start to give an answer. The elapsed time between the two key presses is the response time. This method of measuring response times is not complex, but it does require extra interviewer training and puts more burden on the interviewers, who inevitably introduce errors in starting and stopping the timer (Bassili, 1996; Bassili & Fletcher, 1991; Mulligan, Grant, Mockabee, & Monson, 2003). A voice-activated timer uses a voice-key to stop the timer the moment respondents utter a sound. But it still relies on interviewers to start the timer. Comparatively speaking, voice-activated measures are free from the errors introduced by interviewers in stopping the timer. But, as an empirical matter, voice-activated timers seem to generate more invalid measurements than interviewer-activated timers (Bassili, 1996; Bassili & Fletcher, 1991). The other major approach involves latent timers that measure the entire time required to administer and answer a survey question. In a CATI survey, this involves timing from the moment the question appears on the interviewer's monitor to the interviewer's coding of the answer. It should be noted that, in web surveys, both server-side and client-side response times involved the use of a latent timer.

A key difference between active and latent timers lies in the implicit assumption each makes about when the survey response process starts. Active timers reflect the influence of Fazio's model and assume that the response process begins only after the stimulus has been completely presented to the respondent. Latent timers assume that the response process begins as soon as the question starts being presented to the respondent. Bassili has consistently advocated the use of active timers and urged researchers to avoid latent timers (Bassili, 1996, 2000). Indeed, almost all studies measuring survey response times except the ones investigating web surveys have employed this measurement approach.

However, recent work using response times has challenged the validity of the process assumptions underlying active timers. There is no reason why respondents have to wait until the end of a survey question to start the response process (Draisma & Dijkstra, 2004; Mulligan et al., 2003).[2] It is not uncommon in practice that (at least some) respondents jump the gun and give an answer before interviewers finish reading the question. In addition, empirical evidence has demonstrated that response times obtained via active timers and latent timers are significantly correlated; they produce consistent and comparable model parameter estimates given correct model specification (Grant, Mulligan, Mockabee, & Monson, 2000; Mulligan et al., 2003). Since much of the response formulation process is likely to take place while the interviewer reads the question to the respondent (or while the respondent reads it on the screen), the latent timer approach would seem to be more appropriate.

The second problem has to do with the analysis of response time data. Response times are recorded for each question and for each individual respondent in a survey. As a result, response times to survey questions are cross-classified by the respondents and the questions. In other words, response times are nested within the cell formed by cross-classifying the respondents and the questions. In addition, the response times from a

[2] Bassili (1996) later acknowledged that comprehension of the questions and reporting an answer take time and thus affect overall response latencies as well.

single individual will not be independent; intra-respondent correlations can produce a 'design effect' increasing the variance of parameter estimates. In a similar vein, the response times to a particular question item produce an intra-item correlation as well, reflecting the common impact of properties of that question item on response times. Thus, analyses of response times need to take into account this cross-classified structure and the associated intra-respondent and intra-item correlations. Unfortunately, most of the studies have carried out analyses at either the question level or the respondent level. For example, Bassili recommends examining either questions or respondents:

> When the focus is on properties of questions, interquestion response latency comparisons are most informative. When the focus is on the cognitive properties of attitudes, intersubject latency comparisons are also most informative (1996: 331).

Bassili's recommendation is most appropriate for looking at response times for a single question. However, in analysing response times from multiple questions and multiple respondents, it is important to take the doubly nested structure of the response time data into account.

## RESPONSE TIMES AS DEPENDENT VARIABLE

This paper takes on a different perspective on response times to survey questions. We consider response times to be a dependent measure in their own right and take a broader look at their determinants. We examine response times to questions in four web surveys and attempt to identify the variables affecting response times. (Because collecting client-side response times requires more intensive programming with JavaScript, we collected them only for a small set of questions and we use server-side response times in this analysis. To save space, we omit the word 'server-side' from now on and use the term only when we contrast it to client-side response times. We will talk about the validity of the server-side response times in the discussion section.) We adopt the survey response model outlined in Tourangeau et al. (2000) as our theoretical framework; in addition, we draw on the work of Bassili and his colleagues and on Just and Carpenter's (1992) capacity theory of comprehension to identify potential predictors for the model of response times.

Tourangeau et al. (2000) divide the survey response process into four major components—comprehension of the question, retrieval of relevant information, use of that information to render the judgment or estimate required by the question and the selection and reporting of an answer (Tourangeau et al., 2000, p: 7). Each of these components encompasses one or more processes that take time and affect the overall response latency (Bassili, 1996). The comprehension and reporting components are closely linked to characteristics of the questions and are driven, at least in part, in a bottom–up manner by these question properties (Bassili, 1996). For example, the reporting process is strongly affected by whether the item is open or closed and, if it is closed, by how many and what type of response categories it offers (see Tourangeau et al., 2000, Chapter 8). Retrieval and judgment may be determined by respondent characteristics, being driven, at least in part, in a top–down manner by structural features of the respondent's knowledge or attitudes (Bassili, 1996). In any case, response times are likely to reflect both question characteristics and respondent characteristics.

DOI: 10.1002/acp

This conclusion is also consistent with Just and Carpenter's (1992) capacity theory of comprehension. According to Just and Carpenter, comprehension is constrained by working memory capacity. Whether a particular sentence is misunderstood depends both on features of the sentence (such as its length and syntactic complexity) and the working memory capacity of the person trying to comprehend it. Thus, we examine both item-level and respondent-level characteristics in our models to investigate their effects on response time.

We examined two sets of item characteristics, reflecting features of the question itself and features of the response options. Question complexity is represented by the number of clauses in the question (excluding the response categories), the number of words per clause and the question type (i.e. whether a question is an attitude/opinion question, factual/behavioural question or a demographic question). We also looked separately at the number of response categories, whether they formed a scale or not, and, if so, whether every scale point was labelled or just the scale end points. We also looked at whether the scale was a frequency scale or some other type of rating scale. Finally, we included the position of the question within the questionnaire in the models as well.

Among respondent characteristics, we believed that age, education, experience with the Internet and the number of prior surveys done would all affect response times. Age and education are particularly important factors in determining a respondent's working memory capacity (see Salthouse, 1991). Experience with the Internet and web surveys would reflect practice in responding to questions via the Internet. Table 1 summarises the item and respondent characteristics included in our model.

We will build a two-level cross-classified model to examine the effects of question and respondent characteristics jointly on response times; thus, the emphasis of the paper is not on the relation between response time and data quality *per se*. Even though long response times do not necessarily suggest bad data quality, longer response times than average do signal longer processing in one or more components of the survey response framework and call for attention from survey researchers. With the current two-level analysis, we hope to show average effects on response times of various characteristics in web surveys.[3]

Table 1. Summary of item and respondent characteristics

| | |
|---|---|
| Item characteristics | Complexity of the question<br>Number of clauses<br>Number of words per clause (clause length)<br>Question type |
| | Complexity of the response options<br>Number of response categories<br>Nature of response categories (not a scale, fully labelled scale, scale with only end points labelled, rating scales and frequency scales) |
| | Location (position of the question within the questionnaire) |
| Respondent characteristics | Age<br>Education<br>Experience with the Internet<br>Number of surveys done |

[3] Other methods of analysing the data (such as the method outlined by Clark, 1973 ) make it difficult or impossible to estimate the effects of both types of characteristics at the same time.

## DATASETS

The response times are from four web surveys conducted by MSInteractive. Survey Sampling Inc. (SSI) selected the samples for all the four surveys.

For the first two web surveys, SSI sampled from its Survey Spot and the eLite frames. The Survey Spot frame included more than a million web users who had signed up online to receive survey invitations. The eLite frame consisted of more than seven million E-mail addresses of web users, who had agreed to receive messages on a topic of interest. We found no differences between members of the two frames and combined them for each study. SSI selected 14 264 E-mail addresses for the first survey and 39 217 for the second and sent out E-mail messages inviting the recipients to take part in ‘a study of attitudes and lifestyle’. The E-mail invitations included the URL of the web site with our questionnaire and a unique ID number (which prevented respondents from completing the survey more than once). The first survey ran from 2 April 2002 to 23 April 2002 and the second from 26 March 2003 to 7 April 2003. A total of 2568 respondents completed the first survey for a response rate (AAPOR (2000) RR1) of 18.0%; 2722 completed the second for a response rate (AAPOR (2000) RR1) of 6.9%. These response rates are not high by the standards of high-quality mail surveys but are quite typical for web surveys (see Couper, 2000).

For the third and fourth web studies, SSI again used two different sampling sources. The same sampling procedure and data collection protocol used in the first two studies were implemented with the Survey Spot frame for studies three and four. These studies also sampled from a second source, however. The America Online Opinion Place provides access to approximately 25 million AOL account holders. Opinion Place uses a technique they call ‘river sampling’. Survey invitations are posted on banners throughout the AOL service and related sites. Users willing to click through are screened against the respondent requirements for active surveys, and then passed through to a survey for which they qualify. Respondents who complete surveys accrue miles in the American Airlines AAdvantage Program. This sampling technique makes it impossible to compute a response rate. Study 3 was fielded from 18 December 2003 to 31 December 2003 while Study 4 from 5 November 2004 to 14 November 2004.

The questionnaires for the four web surveys were quite similar and included questions on a range of topics such as health, diet and travel. We did not attempt to model a number of question items that were in a grid or matrix format, since we did not have separate response times for each question in the grid. The analysis included 27 questions from the first survey, 26 items from the second, 59 from the third and 61 from the fourth survey. Table 2 presents general information on the four web studies.

Table 2. Selected characteristics of the four web surveys

| | Study 1 | Study 2 | Study 3 | Study 4 |
|---|---|---|---|---|
| Response rate[a] | 18.0% | 6.9% | 10.8%[b] | 4.6%[b] |
| No of completes | 2568 | 2722 | 2717 | 2587 |
| Field period | 2/4/2002–<br>23/4/2002 | 26/3/2003–<br>7/4/2003 | 18/12/2003–<br>31/12/2003 | 5/11/2004–<br>14/11/2004 |
| No of questions | 27 | 28 | 59 | 61 |

[a] Response rates reported here are AAPOR RR1.
[b] Response rates are reported for the Survey Spot frame only.

DOI: 10.1002/acp

## RESULTS

Given the nature of our datasets (i.e. the response times to question items are cross-classified by survey respondents and survey questions), we used a cross-classified random effects model to analyse the response times (see Raudenbush & Bryk, 2002, for detailed discussion). A cross-classified model is a variant of a multilevel model. As Bryk and Raudenbush (1992) have shown, common problems that can result when multilevel data are analysed with single-level methods include aggregation bias, misestimated standard errors, heterogeneity of regression slopes and systematic misestimation of group effects. We adopted a cross-classified model for our analysis here. The level 1 data consists of response times to all survey questions by all the respondents. Item-level characteristics such as the number of clauses, the number of words per clause, the question type, the number of response categories, the format of the response categories and the location of the question within the questionnaire are column variables whereas respondent-level characteristics—age, education, number of surveys done, experience with the Internet and Internet connection—are specified as row variables. We also include survey indicator variables as row variables to account for any unexplained random effects by surveys.

### Dealing with outliers and skewed distribution

Because of their skewed distribution, a major issue in dealing with response times is how to handle outliers, in particular very long response times. We followed the recommendations by Ratcliff (1993) and replaced observations beyond the upper and lower one percentile with the upper and lower one percentile values separately.[4] These cutoffs were determined separately for each item. To further reduce the skewness, we took a log transformation on the response time variable and used the log transformed response times as the dependent variable in our cross-classified model. Table 3 presents the distribution parameters of response times in its original scale and in the log scale. The correlation between the raw and the transformed response times is 0.83.

### Unconditional model

As a first step, we fitted a fully unconditional model (which is equivalent to a two-way ANOVA model with random respondent and question effects) to partition the total variance of response times into within-cell ($\sigma^2$) and between-cell components.

The level-1 model (that is, the within-cell model) is

$$Y_{ijk} = \pi_{0jk} + e_{ijk}, \quad e_{ijk} \sim N(0,\ \sigma^2)$$

Table 3. Response times in original scale and log scale

| | Mean | Standard deviation | Skewness |
|---|---|---|---|
| Response times in original scale | 11.76 | 11.00 | 5.74 |
| Response times in log scale | 2.23 | 0.64 | 0.50 |

[4] We also dropped the upper and lower one percentile of data (2% of the total data) and reran the analysis with this alternative treatment of outliers. Conclusions regarding fixed and random effects remain the same.

where $Y_{ijk}$ is the $i$th response time for respondent $j$ to question $k$; $\pi_{0jk}$ is the mean (or expected) response time for cell $jk$ cross-classified by respondent $j$ and question $k$ and $e_{ijk}$ is the random effect of individual response time.

At level 2, the between-cell component variance is further partitioned into variance between respondents and variance between questions. That is, we only consider a 'main effects model'. The random effect associated with the respondent-by-question interaction is omitted because the cell sizes ($n = 1$) do not allow us to reliably distinguish this source of variation from the within-cell error. Thus, the level-2 model (the between-cell model) is

$$\pi_{0jk} = \theta_0 + b_{00j} + c_{00k},$$

$$b_{00j} \sim N(0,\ \tau_{b00}),$$

$$c_{00j} \sim N(0,\ \tau_{c00})$$

where $\theta_0$ is the grand mean response time; $b_{00j}$ is the random main effect for respondent $j$, that is the contribution of respondent $j$ averaged over all questions items; $c_{00k}$ is the random main effect for item $k$, that is the contribution of item $k$ averaged over all respondents.

The level-1 and level-2 models give rise to the combined unconditional model:

$$Y_{ijk} = \theta_0 + b_{00j} + c_{00k} + e_{ijk}$$

Table 4 shows the results from this model (Model 1), which will serve as the base to be compared to our final fitted models.

A couple of points are worth mentioning about the results from the unconditional model. First, the overall intercept ($\theta_0$) is a grand mean; its estimated value indicates that respondents took about 10 seconds on an average to answer a question. In addition, about 30% of the total variation in response times is across respondents and about 27% is across survey questions. This is shown in the intra-respondent (or intra-item) correlation coefficients, which are the ratios of the between-respondent (or between-item) variance to the total variance. A major portion of the variability in the response times—43%—is the residual variance of individual response times.

## Conditional models

To account for variations across respondents and questions, we included several respondent-level and item-level predictors in the level-2 model. Our simplest conditional model—Model 2—fixes the effects of question-level and respondent-level predictors; that is, the main effects of the question-level predictors are assumed to be constant over respondents and the main effects of the respondent-level predictors are assumed to be constant over questions.

The respondent-level predictors included in the level-2 model were whether the respondent was over 56 or not (age), whether he or she completed high school or not (edu),[5] whether he or she had done more than 15 surveys before or not (svydone) and whether the respondent rated his or her ability to use the Internet as 'advanced' or higher or not (web).[6]

[5] We ran the same models with different parameterisations of age and education in our prior exploratory analyses. The conclusions hold whether age is treated as a continuous variable or is treated as four age groups; similarly, the main results do not change whether we formed three educational categories or two. For simplicity, we present the results with age and education dichotomised.

[6] The particular cut-off points chosen for variables 'web' (web experience) and 'svydone' (the number of surveys completed) were median values.

DOI: 10.1002/acp

Table 4. Results from the cross-classified models

| | Model 1 (Unconditional) | | Model 2 (Fixed row- and column-effects) | | Model 3 (Random row- and column-effects) | |
|---|---|---|---|---|---|---|
| | Log scale | *SE* | Log scale | *SE* | Log scale | *SE* |
| Fixed effects | | | | | | |
| Intercept (Grand mean $\theta_0$) | 2.32 | 0.02 | 2.00 | 0.08 | 2.02 | 0.08 |
| $\gamma_{01}$ (edu) | | | −0.05 | 0.01 | −0.05 | 0.01 |
| $\gamma_{02}$ (age) | | | 0.14 | 0.01 | 0.15 | 0.01 |
| $\gamma_{03}$ (web) | | | −0.09 | 0.01 | −0.09 | 0.01 |
| $\gamma_{04}$ (svydone) | | | −0.03 | 0.01 | −0.03 | 0.01 |
| $\gamma_{05}$ (modem) | | | 0.28 | 0.01 | 0.26 | 0.01 |
| $\gamma_{06}$ (surv2) | | | −0.03 | 0.06 | −0.03 | 0.06 |
| $\gamma_{07}$ (surv3) | | | −0.22 | 0.05 | −0.21 | 0.05 |
| $\gamma_{08}$ (surv4) | | | −0.37 | 0.05 | −0.39 | −0.39 |
| $\beta_{01}$ (ans_cat) | | | 0.02 | 0.00 | 0.02 | 0.00 |
| $\beta_{02}$ (location) | | | −0.06 | 0.02 | −0.07 | 0.02 |
| $\beta_{03}$ (totclaus) | | | 0.11 | 0.02 | 0.13 | 0.02 |
| $\beta_{04}$ (wdclaus) | | | 0.02 | 0.00 | 0.02 | 0.00 |
| $\beta_{05}$ (fact_opin) | | | −0.02 | 0.02 | −0.02 | 0.02 |
| $\beta_{06}$ (demo_oth) | | | −0.19 | 0.06 | −0.21 | 0.05 |
| $\beta_{07}$ (freq_rate) | | | −0.01 | 0.04 | 0.02 | 0.02 |
| $\beta_{08}$ (freq_label) | | | 0.01 | 0.02 | 0.01 | 0.02 |
| $\beta_{09}$ (rate_label) | | | 0.05 | 0.02 | 0.04 | 0.02 |
| $\beta_{10}$ (scale_no) | | | −0.01 | 0.01 | −0.01 | 0.01 |
| Variance components | | | | | | |
| Respondent variance | | | | | | |
| var($b_{00j}$) | 0.12 | | 0.08 | | 0.09 | |
| Question variance | | | | | | |
| var($c_{00k}$) | 0.1 | | 0.05 | | 0.05 | |
| Residuals | | | | | | |
| var($e_{ijk}$) | 0.17 | | 0.17 | | 0.15 | |
| Deviance (Model DF) | 467594.55 (4) | | 464238.37 (22) | | 452047.93 (44) | |

*Note*: Altogether 10 376 respondents and 173 items are used in this analysis.

These variables are expected to affect survey response times either through their relation to working memory capacity or through expertise.

We also included as respondent-level characteristics whether the respondent used a telephone modem or not to connect to the Internet (modem), and three dummy variables to indicate which of the four web surveys respondents participated in (surv2, surv3 and surv4 with the reference category being participation in the first survey). These respondent-level characteristics are not of substantive interest to us; they were included in the model to account for effects that are out of our control so as to obtain a better estimate of the respondent's main effects of interest.

Question-level predictors included in the between-cell (level-2) model are the number of answer categories (ans_cat), the location of the question (whether it was in the first, second, third or fourth quarter of the questionnaire), the total number of clauses in the question (totclaus), the number of words per clause (wdclaus) and six variables that represent the question type and the formats of the response categories. The first of these variables contrasts factual questions with attitude questions (fact_opi); the second contrasts demographic questions against other questions. The third distinguishes frequency scales from other rating scales (freq_rat). The fourth (freq_lab) and fifth (rate_lab) variables contrast fully labelled scales with end-labelled frequency/rating scales. The last variable (scale_no) compares whether the response options constituted a scale or not.[7]

The level-2 model (between-cell model) is presented below:

$$\begin{aligned}\pi_{0jk} = {} & \theta_0 + \gamma_{01}\mathrm{EDU}_j + \gamma_{02}\mathrm{AGE}_j + \gamma_{03}\mathrm{WEB}_j + \gamma_{04}\mathrm{SVYDONE}_j + \gamma_{05}\mathrm{MODEM}_j \\ & + \gamma_{06}\mathrm{SURV2}_j + \gamma_{07}\mathrm{SURV3}_j + \gamma_{08}\mathrm{SURV4}_j \\ & + \beta_{01}\mathrm{ANS\_CAT}_k + \beta_{02}\mathrm{LOCATION}_k + \beta_{03}\mathrm{TOTCLAUS}_k + \beta_{04}\mathrm{WDCLAUS}_k \\ & + \beta_{05}\mathrm{FACT\_OPI}_k + \beta_{06}\mathrm{DEMO\_OTH}_k + \beta_{07}\mathrm{FREQ\_RAT}_k + \beta_{08}\mathrm{FREQ\_LAB}_k \\ & + \beta_{09}\mathrm{RATE\_LAB}_k + \beta_{10}\mathrm{SCALE\_NO}_k + b_{00j} + c_{00k}\end{aligned}$$

The level-2 model captures the main effects of respondent-level characteristics such as education ($\gamma_{01}$), age ($\gamma_{02}$), experience with doing surveys ($\gamma_{03}$) and web experience ($\gamma_{04}$) as well as the main effects of question-level characteristics such as the number of answer categories ($\beta_{01}$), the location of the question within the questionnaire ($\beta_{02}$) and so on.

In contrast, Model 3 allows the main effects of certain selected respondent-level characteristics to vary randomly across question items and the main effects of certain selected question-level characteristics to vary randomly across survey respondents. The selection of the random effects is driven by both empirical reasons (the random effects are comparatively large) and theoretical reasons (e.g. models of processing capacity theory suggested that age might have differential impact on response times to different survey questions). Specifically, we allowed the effect of age to vary randomly across questions and the effect of question types (FACT_OPI and DEMO_OTH) and the types of response scale (FREQ_RAT, FREQ_LAB and RATE_LAB) to vary randomly over respondents. This model is thus:

$$\begin{aligned}\pi_{0jk} = {} & \theta_0 +_{\gamma01} \mathrm{EDU}_j + (\gamma_{02} + c_{02})\mathrm{AGE}_j +_{\gamma03} \mathrm{WEB}_j +_{\gamma04} \mathrm{SVYDONE}_j +_{\gamma05} \mathrm{MODEM}_j \\ & +_{\gamma06}\mathrm{SURV2}_j +_{\gamma07} \mathrm{SURV3}_j +_{\gamma08} \mathrm{SURV4}_j \\ & + \beta_{01}\mathrm{ANS\_CAT}_k + \beta_{02}\mathrm{LOCATION}_k + \beta_{03}\mathrm{TOTCLAUS}_k + \beta_{04}\mathrm{WDCLAUS}_k \\ & + (\beta_{05} + b_{05})\mathrm{FACT\_OPI}_k + (\beta_{06} + b_{06})\mathrm{DEMO\_OTH}_k + (\beta_{07} + b_{07})\mathrm{FREQ\_RAT}_k \\ & + (\beta_{08} + b_{08})\mathrm{FREQ\_LAB}_k + (\beta_{09} + b_{09})\mathrm{RATE\_LAB}_k \\ & + \beta_{10}\mathrm{SCALE\_NO}_k + b_{00j} + c_{00k}\end{aligned}$$

The results of the two conditional models are displayed in Table 4. The random effects of the selected level-2 predictors are presented in Table 5.

## Random effects

Comparing the estimates of variance components of the conditional models to the corresponding base numbers of the unconditional model (see Table 4), we can compute the percentage of the systematic variance explained at both respondent and question level. Both the fixed effects (Model 2) and random effects (Model 3) conditional models accounted for about 30% of the variation across respondents and about 50% of the variation across items. Specifying certain respondent-level and question-level main effects, randomly reduced the within-cell residual variance from 0.17 (from the unconditional

[7] Given the very large sample sizes for our analyses (close to half a million observations), fitting a model with significant predictors is less of a problem than finding a reasonably parsimonious model. We chose the level-2 predictors based on extensive exploratory analyses of the first two web surveys. We think this final set of predictors struck a good balance between parsimony and model fit.

Table 5. Estimates of random level-2 effects from the random effects conditional model

Respondent-level random effects: Variance-covariance matrix

| | var($b_{00j}$) | Fact_opi ($b_{05}$) | Demo_oth ($b_{06}$) | Freq_rat ($b_{07}$) | Freq_lab ($b_{08}$) | Rate_lab ($b_{09}$) |
|---|---|---|---|---|---|---|
| var($b_{00j}$) | 0.085 | 0.000 | 0.013 | −0.006 | −0.001 | 0.001 |
| Fact_opi ($b_{05}$) | | 0.004 | 0.004 | −0.003 | 0.007 | 0.001 |
| Demo_oth ($b_{06}$) | | | 0.024 | −0.008 | 0.006 | 0.005 |
| Freq_rat ($b_{07}$) | | | | 0.017 | −0.019 | −0.009 |
| Freq_lab ($b_{08}$) | | | | | 0.034 | 0.008 |
| Rate_lab ($b_{09}$) | | | | | | 0.009 |

Question-level random effects: Variance-covariance matrix

| | var ($c_{00k}$) | Age ($c_{02}$) |
|---|---|---|
| var ($c_{00k}$) | 0.053 | 0.004 |
| Age ($c_{02}$) | | 0.002 |

model) to 0.15 (from the random effects conditional model), a reduction of 8%. The introduction of level-2 predictors and the specification of level-2 random effects improved the model fit significantly.

## Impact of respondent characteristics

Here we summarise the main findings from the fixed effects in the cross-classified models, beginning with the respondent characteristics that seem to affect response times. The estimates of fixed effects do not vary much across the two conditional models; thus, the estimates of the random effects conditional model are used for discussion of fixed effects.

Education, as expected, has a significant effect on response times. Holding everything else constant, respondents who completed high school are faster than those who did not, a result consistent with Just and Carpenter's (1992) capacity theory.

If aging reduces working memory capacity (as Salthouse, 1991, and others have argued), older respondents should take longer to respond than younger people do. Our analysis shows the expected age effect, as do earlier findings in the cognitive aging literature (cf. Schwarz, Park, Knauper, & Sudman, 1999). Older respondents on an average are slower than younger respondents. This result is also consistent with findings by Fricker, Galesic, Tourangeau, and Yan (2005), who compared web and telephone versions of the same questions. They showed that the time needed to complete the questions increased with age for both web and telephone respondents, but the relation between age and completion times was much steeper for those who completed the web version of the questions. In addition, the random effects model show that the negative effect of age on response times does vary significantly across question items.

We believed that people who have more experience using the Internet might have an advantage in completing our surveys because they are more familiar with navigating across hyperlinks, reading on screen and completing the other tasks involved in responding to web questions than their less experienced counterparts. For the same reason, we expected people who had already completed at least 15 surveys to be faster in responding to our

DOI: 10.1002/acp

questions than those who did not complete as many questionnaires. Our analysis showed such an expertise effect of web experience and survey experience even after controlling for other respondent-level and question-level characteristics.

### Impact of item characteristics

The model included two simple measures of the structural complexity of the question—the number of clauses in the question (excluding the response options) and the number of words per clause (cf. Just & Carpenter, 1992). Not surprisingly, holding everything else constant, the more clauses in the question, the longer it took for the respondents to answer it. In addition, the more words per clause, the longer the response time. Questions with more and longer clauses seem to increase burden on respondents, and they take longer to process as a result.

The models also distinguished three question types. One contrast indicated that our respondents were in general faster with factual than with attitudinal questions. Even though this difference in response times between factual and attitudinal questions did not reach statistical significance, it significantly varied over survey respondents. The variance of this random effect was 0.004. That is, whether a question is attitudinal or factual affects the response times by respondents differentially.

A second set of contrasts showed that respondents were much faster with demographic questions than with the other two types of questions (factual and attitudinal questions). Our findings are consistent with Bassili and Fletcher's finding that factual questions about oneself had the shortest response times whereas factual questions about others had a longer response time; both types of factual items had shorter response times than questions requiring the expression of attitudes (Bassili, 1996; Bassili & Fletcher, 1991). Both our findings and Bassili's seem to support the conventional wisdom that attitudinal questions are harder to answer and need longer time than the other two types of questions (even after controlling for other question properties), perhaps because answers to simple factual questions are more likely to be based on pre-stored answers or other readily retrieved information whereas answers to attitude questions require more difficult retrieval and integration (see Sudman, Bradburn, & Schwarz, 1996; Tourangeau et al., 2000). Furthermore, the contrast of demographic questions against other two types of survey questions shows significant variation across survey respondents (with a variance of 0.020). Our speculation is that, in our study, respondents' experience with web surveys improved respondents' speed with the demographic and factual questions more than with attitudinal questions. The major component of the processing time with these demographic and factual items may be identifying the correct answer on screen and making the appropriate response (e.g. clicking on the radio button) rather than with formulating the answer itself; as a result, experienced web respondents show greater gains with these items than novices do.

We thought it likely that questions with a larger number of answer categories might impose greater burden on respondents' working memories. The results show a significant effect of number of answer categories on the response time—the more options, the slower the answer, holding everything else constant. This effect could be a result of increased burden of processing or it may simply reflect added reading time. If the burden of processing the options is the issue, the impact of the number of options might depend on whether the options formed an ordered scale; if reading time is the issue it may matter whether each scale point was labelled or just the end points. Our analysis shows that

DOI: 10.1002/acp

whether response options formed a response scale or not has no significant impact on response times. Neither does it make a difference in response times whether the response scale is a frequency scale or a rating scale and whether every point of the frequency scale is labelled or just the end points. However, there is a marginally significant effect of a fully labelled rating scale on response times. On top of the nonsignificant fixed effects, the random effects conditional model revealed that respondents vary significantly in the difference in response times between frequency scales and rating scales and in the impact of a fully labelled frequency (or rating) scale. We interpret these findings to mean that the impact of greater numbers of response options on overall response times mostly reflects the increased time to read the options and that the increased reading time varies across respondents.

The final question characteristic that affected response times was the position of the question within the questionnaire. We classified each item according to whether it came in the first fourth of the questionnaire, the second fourth and so on. Our models showed that, conditioning on other variables in the model, respondents tended to answer more quickly as they got closer to the end of the questionnaire—the data show a significant negative impact of location on response times. However, readers are cautioned that we did not attempt to vary the order of the survey questions. Future studies experimenting with the order of the survey items are needed to fully address the location effect we observed.

## DISCUSSION

Although response times have long been studied by psychologists, their use in survey settings is still relatively rare. This paper attempts to take a systematic look at the variables that affect response times in web surveys. We examine a number of potential predictors of response times and construct a preliminary cross-classified model that demonstrates the effects of both question and respondent characteristics on response times. Of course, there are probably many additional item- and respondent- level variables that affect response times that we were unable to incorporate in our models.

Our models were based on some rather simple assumptions. First, we thought that the items systematically varied in how difficult they were to read and understand, and we included a number of variables (e.g. number of clauses, number of words per clause, number of response categories) that would reflect these item-level differences. Second, we thought that some questions would be likely to elicit ready-made answers and others would require new judgments or estimates. Our models included several type-of-question variables (e.g. demographic questions vs. attitudinal questions) as indirect indicators of the type of processing needed. Finally, we thought that response speed would depend on the cognitive capacity of the respondents and their expertise in completing survey questions, and we included a number of variables related to that hypothesis (such as respondent age, education and prior experience with web questionnaires). For the most part, we found differences in line with our initial assumptions.

A couple of points warrant additional discussion. First of all, our findings demonstrate the importance of using the multilevel modelling approach. Most of the previous works on response times have ignored the cross-classified structure in response time datasets. Typically, analyses are carried out either at the item level or the respondent level. However, as our results demonstrate, both item and respondent characteristics affect response times. In addition, the effects of item characteristics can vary across respondents and the effects of

DOI: 10.1002/acp

respondent characteristics can vary across items. The significant random effects suggest that ignoring the nested nature of the data is inappropriate and can lead to serious underestimation of the error terms. Our work is the first we are aware of to incorporate both levels in examining response time data.

Second, our results regarding item characteristics also parallel some of Saris's (2005) findings from his meta-analysis of question wording and format experiments and the diagnoses generated by QUAID program developed by Graesser, Bommareddy, Swamer, and Golding (1996). Saris fit structural equation models to datasets that collected multiple indicators of several traits via two or more methods (e.g. using two or more response formats). His meta-analysis of these multi-trait, multi-method experiments shows that the number of answer categories and the use of fully labelled scales, both of which tended to increase response times in our datasets, contributed positively to item reliability. This finding suggests that increases in response times do not necessarily signal increased difficulty or decrements in data quality. We also compared our simple quantitative measure of question complexity with the measure available from the QUAID program. QUAID was designed to analyse potential problems with draft survey questions (Graesser et al., 1996). We ran 55 survey questions from the first two web surveys through QUAID and correlated our quantitative measures with QUAID's diagnoses. We found that the 'complex syntax' indicator generated by QUAID correlated highly with the number of clauses ($r = 0.48$) and the average clause length (0.41). If a scale is a frequency scale it has a correlation of 0.43 with the QUAID diagnosis 'vague or imprecise relative term'. This is presumably because frequency scales tend to include such terms as 'often', or 'most of the time', which are considered problematic by QUAID.

Third, our work differs from some of the previous studies of response times in surveys (Bassili, 1993), which have tended to employ active timers. Our response times use a latent timer on the server where surveys resided; they capture the entire time from the moment the server computer displayed the question until the moment it received an answer. Thus, our results reflect reading times and other comprehension processes. Still, many of our other results are quite compatible with findings from studies employing active timers. Bassili and Fletcher (1991) showed an ordered list of response times corresponding to various types of questions:

> On average, simple questions about salient facts seem to take less than 1 second to answer. Questions about facts that are less salient or that require a simple frequency estimate take between 1 and 1.4 seconds. Simple attitude questions take between 1.4 and 2 seconds, whereas more complex attitude questions take between 2 and 2.6 seconds (1991: 339).

We ran the same cross-classified random effects conditional model on the same trimmed dataset in its raw scale (without log transformation). The estimated fixed effect of the contrast of attitudinal versus factual questions was $-0.24$; that is, conditioning on other variables in the model, factual questions are about 0.24 seconds faster than the attitudinal questions. This is comparable to the difference in response times between questions 'about facts that are less salient or that require a simple frequency estimate' and 'simple attitude questions' reported by Bassili and Fletcher (1991). The estimated fixed effect of the contrast of demographic questions versus the other two types of questions is $-1.02$, which suggests that demographic questions are about 1 second faster than the other types of questions holding everything else constant. Again, this estimate of difference is similar to the one between ‘simple questions about salient facts’ and ‘simple attitude questions’ reported in the passage quoted above. Active timers have the drawback that they start timing after processing has begun. Our results indicate that latent timers give sensible results that converge with those from studies employing active timers (for similar conclusions, see Grant et al., 2000; Mulligan et al., 2003). In addition, our approach allows us to examine factors affecting comprehension.

Fourth, we acknowledge that server-side response times used in our analysis included additional ‘nuisance’ components (such as download times) besides the cognitive processes that were of primary interest. In the last web study, we also recorded client-side response times for several items; these excluded the download times. We correlated our server-side response time measures with the client-side response times on 16 questions. On average, the client-side response times were about 3 to 4 seconds shorter than the server-side times. Still, the correlations between the two response time measures range from 0.91 to 0.99 across the items with the average correlation of 0.96. Thus, we did not attempt to control for differences across respondents in connection or processing speed (the only variable we controlled in our model was whether the respondent used a telephone modem to connect to the Internet or not). We suspect that these nuisance components are only likely to affect our estimates of the grand mean ($\theta_0$), leading to an overestimate of $\theta_0$. The high correlations between the server-side and the client-side response times suggest that the noise in server-side response times did not necessarily affect various regression coefficients.

A key limitation of this study is that the findings are correlational. We did not attempt to experimentally manipulate the length or position of the items or the number of response options they offered. Future studies that systematically vary such question characteristics would help to advance the literature on response times in surveys.

## ACKNOWLEDGEMENTS

The work reported here was supported by grants from the National Science Foundation (SES-0106222) and NIH (R01 HD041386-01A1) to Roger Tourangeau, Mick Couper, Fred Conrad and Reg Baker. All opinions, findings and conclusions expressed here are those of the authors and do not reflect the views of the National Science Foundation or the National Institutes of Health. We would like to extend our personal thanks to Mick Couper, Fred Conrad, Frauke Kreuter and Reg Baker for their comments, advice, and support.

## REFERENCES

American Association for Public Opinion Research, (AAPOR). (2000). *Standard definitions: Final disposition of case codes and outcome rates for surveys*. Lenexa, KS: AAPOR.

Bassili, J. N. (1993). Response latency versus certainty as indexes of the strength of voting intentions in a CATI survey. *Public Opinion Quarterly*, *57*, 54–61.

Bassili, J. N. (1996). The how and the why of response latency measurement in telephone surveys. In N. Schwarz, & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 319–346). San Francisco: Jossey-Bass.

Bassili, J. N. (2000). Editorial’s introduction: On response latency measurement in telephone surveys. *Political Psychology*, *21*, 1–6.

DOI: 10.1002/acp

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.

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly*, *60*, 390–399.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, NJ: Sage Publications.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359.

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

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*, 464–494.

Donders, F. C. (1868). Over de snelheid van psychische processen. *Onderzoekingen gedaan in het Physiologisch Laboratorium der Utrechtsche Hoogeschool, 1868–1869, Tweede reeks, II*, 92–120.

Dovidio, J., & Fazio, R. H. (1992). New technologies for the direct and indirect assessment of attitudes. In J. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys* (pp. 204–237). New York: Russell Sage Foundation.

Draisma, S., & Dijkstra, W. (2004). Response latency and (para) linguistic expressions as indicators of response error. In S. Presser, J. Rogthgeb, M. Couper, J. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 131–147). Hoboken, NJ: John Wiley & Sons.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior. *Advances in Experimental Social Psychology*, *23*, 75–109.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*, 229–238.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, *69*, 370–392.

Graesser, A. C., Bommareddy, S., Swamer, S., & Golding, J. (1996). Integrating questionnaire design with a cognitive computational model of human question answering. In N. Schwarz, & S. Sudman (Eds.), *Answering questions* (pp. 143–174). San Francisco: Jossey-Bass.

Grant, J. T., Mulligan, K., Mockabee, S. T., & Monson, Q. (2000). *Response time methodology for survey research*. Paper Presented at the Annual Meeting of the Midwest Political Science Association, April, Chicago.

Heerwegh, D. (2002.). *Describing response behavior in web surveys using client-side paradata*. Paper presented at the International Workshop on Web Surveys, Mannheim, Germany.

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

Just, M. A., Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122–149.

Mulligan, K., Grant, J. T., Mockabee, S. T., & Monson, J. Q. (2003). Response latency methodology for survey research: Measurement and modeling strategies. *Political Analysis*, *11*, 289–301.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Salthouse, T. A. (1991). *Theoretical perspectives on cognitive aging*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Saris, W. (2005). The structural equation modeling approach: The effects of survey characteristics on random and systematic errors in surveys. Talk given to the Total Survey Error workshop, Washington, DC.

Schwarz, N., Park, D., Knäuper, B., & Sudman, S. (1999). *Cognition, aging, and self-reports*. Washington, DC: Psychology Press.

Strack, F., & Martin, L. L. (1987). Thinking, judging, and communicating: A process account of context effects in attitude surveys. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 123–148). New York: Springer-Verlag.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

DOI: 10.1002/acp

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, position, and order: Interpretative heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*, 368–393.

Tourangeau, R., Rasinski, K., & D'Andrade, R. (1991). Attitude structure and belief accessibility. *Journal of Experimental Social Psychology*, *27*, 48–75.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.

van Zandt, T. (2002). Analysis of response time distributions. In H. Pashler, & J. Wixted (Eds.), *Stevens' handbook of experimental psychology* (3rd ed., Vol. 4). *Methodology in experimental psychology*. New York: John Wiley & Sons.

DOI: 10.1002/acp