Loewen, E. Ruth, Edward Bauer, Mary E. Thompson, Nadia Martin, Anne C.K. Quah, and Geoffrey T. Fong. 2022. "Timing Estimates for Complex Programmed Surveys." *Survey Practice* 15 (1). https://doi.org/10.29115/SP-2022-0011.

ARTICLES

Timing estimates for complex programmed surveys

E. Ruth Loewen[1], Edward Bauer[1], Mary E. Thompson[2][a], Nadia Martin[1], Anne C.K. Quah[1], Geoffrey T. Fong[3]

[1] Department of Psychology, University of Waterloo, [2] Department of Statistics and Actuarial Science, University of Waterloo, [3] Department of Psychology, University of Waterloo; School of Public Health Sciences, University of Waterloo; Ontario Institute for Cancer Research

Keywords: ITC Project, complex programmed surveys, branching, completion time, user groups, length estimation methods, survey length

https://doi.org/10.29115/SP-2022-0011

Survey Practice

Vol. 15, Issue 1, 2022

Developing a method for estimating average survey completion time—easily, accurately, and repeatedly—was a many-years-long aspiration for the International Tobacco Control Policy Evaluation (ITC) Project. This article describes the three survey length estimation methods used for the ITC Project's complex programmed surveys, with their respective merits and disadvantages. Method I, Read-Throughs, required staff to read through different versions of the survey. Automating this approach in Method II, Word Counts, replaced staff reading with question word counts and a postulated reading speed, but still required creation of multiple versions that were timed separately. Method III, Weighted Questions, also automated, was streamlined by weighting individual questions, rather than survey versions, by the proportion of respondents likely to answer them. The two automated methods greatly reduced staff time cost, provided surprisingly accurate time estimates, and made it possible to quickly re-estimate the length after each survey revision. Other survey researchers using surveys with complex branching may benefit from our experimentation and results.

Introduction

A chronic challenge for survey researchers is the need to estimate completion time for programmed surveys in telephone (computer-assisted telephone interview [CATI]), online, or in-person (computer-assisted personal interview [CAPI]) modes. Investigators' natural desire to include as many questions as possible is often countered by budget constraints, since survey firms charge more for longer surveys. In addition, longer surveys are associated with increases in respondent breakoffs and decreased respondent attentiveness, which lead to reduced data quality (Galesic and Bosnjak 2009; Hoerger 2010; Loosveldt and Beullens 2013; Qualtrics Support, n.d.). It is therefore essential to obtain reasonably accurate estimates of survey completion time.

There is surprisingly little literature, even less of it academic, on how to go about such estimation. Several survey research firms offer methods of varying complexity and specificity. Versta Research (2011) describes a method that assigns points to questions based on their length, format, complexity, and the nature of the cognitive task. Puleston (2012), in his blog, Question Science, suggests multiple formulas, varying from the simplest—considering question count only—to the most complex—taking into consideration word count,

[a] Corresponding author:
Mary Thompson; University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1; (519) 888-4567, ext 45543; methompson@uwaterloo.ca

reading speed, and question format. Qualtrics Research offers an online length estimation of customers' surveys as part of their Expert Review (n.d.), which, according to their website, includes reading speed, decision time, response entry time, as well as time to transition between questions.

Although no data are provided, the above approaches are intuitively logical and likely would work for simple surveys with little or no branching—i.e., surveys in which all respondents answer all questions. However, none of them is adequate for the more complex surveys conducted these days; in fact, Puleston (2012) states that most errors in estimating survey length centre on the issue of branching. Carter, Bennett, and Sims (2021) systematically tested six such formulas, including Versta's and all of Puleston's, comparing estimated durations to the actual completion times of a Health Risk Assessment Survey. This survey was relatively short (133 questions) but almost one third of the questions were asked of only some respondents, so branching would have undermined the accuracy of all estimates. The formula-derived estimates ranged from 7.6 to 39.6 minutes, while the observed duration was 14.0 minutes, and all formula-derived estimates were statistically significantly different from the observed time. The closest estimate, 16.0 minutes, was computed by Puleston's most complex formula, which included number of words, number of questions, and question format; this 2-minute difference (14% overestimate) is reasonable for such a short survey, but a 14% overestimate for longer surveys would represent considerable inaccuracy, and the discrepancy would be even greater for surveys with more complex skip patterns.

The surveys of the International Tobacco Control Policy Evaluation (ITC) Project provide excellent real-life examples of such branching complexity, and a context in which duration estimates are vitally important. The ITC Project (www.itcproject.com) was created in 2002 to evaluate the impact of governmental tobacco control policies such as large graphic warning labels, smoke-free laws, higher taxes, advertising and promotion bans, and support for cessation (Fong et al. 2006). It has conducted longitudinal surveys across all six World Health Organization (WHO) regions, in 31 countries, which include over half the world's population and over two-thirds of the world's tobacco users. To allow for cross-country comparison, questions across all ITC countries are identical or functionally similar; they deal with tobacco use behaviors as well as measures assessing the impact of the tobacco demand-reduction of the WHO Framework Convention on Tobacco Control. They also assess factors related to tobacco and nicotine use, such as cessation methods, social norms, health conditions and personality traits. Therefore, these surveys are long (between 25 and 50 minutes), though not as long as some other tobacco and health surveys (e.g., the Population Assessment of Tobacco and Health [PATH] Study; Hyland et al. 2016). This does result in some respondent burden, but ITC compensates respondents accordingly, the

compensation amount depending on the country, the specific survey firm's policy, and the number of questions asked (e.g., tobacco users are asked more questions and are more generously compensated).

Because of ITC's international scope, multiple tobacco and nicotine products are typically included in each survey, depending on the country. These products include bidis and areca nut in South Asia (India and Bangladesh), shisha and midwakh/dokha in the Middle East (Abu Dhabi), and e-cigarettes and newer nicotine delivery products in an increasing number of countries. The ITC surveys ask respondents, for each product used, a long list of parallel questions—for example, past and current use, last purchase of the product (quantity, source and price), quit attempts (methods/aids used and outcomes). This results in an enormous number of points where branching, and branches within branches occur. Indeed, the research firms that ITC employs frequently comment that they have never encountered such a complicated survey.

Developing a method for estimating survey completion time—accurately, relatively easily, and repeatedly—was a many-years-long aspiration for the International Tobacco Control Policy Evaluation (ITC) Project. The objective of this article is to describe the evolution of our methods. We describe the three survey length estimation methods used by ITC, with their respective merits and disadvantages. Our original method, Read-Throughs, required staff to read through different versions of the survey, either aloud or silently, followed by computation of a weighted mean time in minutes. Automating this approach in our second method, Word Counts, replaced staff reading with question word counts and a postulated reading speed, but still required creation of multiple versions that were timed separately and weighted. Our third method, Weighted Questions, was also automated, but the procedure was streamlined by weighting individual questions, rather than survey versions, by the proportion of respondents likely to answer them. We hope that other survey researchers and practitioners may benefit from our experimentation and results.

Method I: Read-Throughs

Our first method for timing ITC surveys was to time read-throughs of the surveys, simulating a survey interview. Staff read a draft of the survey, question by question, with a stopwatch, allowing time for imagined responses. For interviewer surveys (telephone/CATI and in-person/CAPI), the reading is aloud; for online surveys, the read-throughs are silent. This method is reasonably successful in helping to determine how much of a survey must be omitted to stay within time limits. However, there are difficulties of reliability, time cost, and accuracy.

First, there is substantial variability in reading speed among readers, so a slower/faster-than-average reader could seriously skew the estimate. This problem can be reduced by using multiple readers, but that strategy increases the demand on staff time.

Second, because of the extensive branching, there is no single pathway through the survey that can be read and timed. Nearly all questions are conditional on previous answers, in large part because the surveys ask about multiple products. To accommodate this challenge, the read-through method uses 3 to 5 different survey versions, each representing the pathway of one of the main user groups, in which product-relevant questions are included and others are skipped. For example, where the products are cigarettes and e-cigarettes, the user groups would be exclusive smokers, exclusive vapers, and dual users (those who smoke and vape), and each user group has a different pathway through the survey.

Even within a user group pathway, however, the questions answered by a respondent vary enormously. Many questions are conditional on something other than product usage. These other gateway questions, if answered in a certain way, will cause the respondent to be asked other downstream questions. For example, answering "yes" to the gateway question "Have you seen a doctor in the last 12 months?" will lead to downstream questions about the doctor providing quitting advice. Responding "married" to the marital status question will result in downstream questions about partner smoking, partner support for quitting, etc. This source of variability has at least as much influence on a respondent's completion time as product usage and is far less predictable. Therefore, two survey versions are created for each user group pathway, a Max and a Min. The Max version assumes that gateway questions are answered in a way that maximizes the number of downstream questions seen, while the Min minimizes the number of downstream questions seen. The mean of the Min and Max read-through times is taken as a time estimate for each user group, and each user group's estimate is weighted by the proportion of respondents expected to belong to that group—proportions usually provided by the previous wave's data. For instance, in our Spain project's second wave, the largest user group was smoker/heard of e-cigarettes, with a proportion of 0.77. The proportions for smoker/not heard of e-cigarettes, quitter/heard of e-cigarettes, and quitter/not heard of e-cigarettes were 0.12, 0.08 and 0.03, respectively. These proportions were used to weight the user group timing estimates for the following wave.

The creation of these many versions requires considerable time, easily 15–30 hours in total, from an ITC staff member skilled at interpreting complex branching. Then each of the versions is read through and timed by 3–4 people, typically involving a total of 20–40 hours. Total staff time cost is therefore in the range of 35–70 hours.

But even this careful estimate of completion time omits all user types other than the few groups that are explicitly timed. And, more critically, taking the mean of the Min and the Max is equivalent to assigning a weight of 0.5 to all downstream questions whose inclusion is not determined by user group,

effectively assuming that they are answered by half the respondents. In many cases, this results in an overestimated average completion time because the downstream questions are answered by far fewer than half the respondents.

These factors unavoidably reduce the accuracy of the resulting average times. Nonetheless, this method produces useful estimates and has been invaluable for many projects. Its greatest disadvantage has been time cost rather than inaccuracy. We must allow a week or two for the total timing process, which provides an estimate of how much needs to be cut from the survey; once the survey has been shortened, it would be hugely beneficial to re-estimate the time, but this is almost never done because of the time required.

Method II: Word Counts

To reduce time costs and to make repeated estimates feasible, we decided to attempt to automate the Read-Throughs approach. The automation not only replaced read-throughs with word counts and a likely reading speed, but it also generated the user group pathways and the Min/Max versions.

The Word Counts method comprised three stages. The first stage was creating the multiple survey versions programmatically rather than manually, by employing information that existed in the ITC Survey Information System (SIS), a relational database that contains the exact wording and much other information for every question asked in every ITC survey; to date it includes more than 175 survey waves, 11,000 distinct questions and 26,000 survey-question combinations.

SIS, the ITC survey database, documents the filter for each question—i.e., the prior question(s) and response(s) that result in that question being asked. However, the questions named in a filter are usually themselves filtered on other earlier responses and those on still others. For example, a question about the respondent's favorite e-cigarette flavor would be filtered on having used more than one flavor, which would be filtered on having used an e-cigarette recently, which would be filtered on having ever used an e-cigarette. The programming challenge was to integrate all of these nested filters and to obtain for each question an exact universe—a single statement, expressed as variable names and responses, of the prior answers that would result in that question being asked. Then user groups had to also be defined in terms of variable names and responses, so that the program could determine which questions would be answered by each user group. In this way, each question in the survey either was, or was not, assigned to the various user groups. This allowed the automatic creation of the survey pathways that had been previously drafted by hand.

The second stage was to automate the Min and Max versions. The program could compute exactly which responses to gateway questions would result in the most downstream questions, the Max version, and which responses

resulted in the fewest downstream questions, the Min version. Thus, the absolute longest and shortest survey versions for each user group could be determined far more accurately than was ever possible by hand.

The third stage was to automate the completion time estimates, using word counts and reading speed to replace staff read-throughs. Since the SIS database contains the exact wording of every question, it was simple to obtain word counts for questions.

However, these counts needed to be adjusted for different types of questions. Table 1 illustrates how word counts and timing estimates were obtained for different types of questions from the Japan Wave 3 survey.

Standalone questions were assumed to be read completely, including preamble, core question and response options, so their word counts are simple totals of all words. Questions in series were handled differently. All the questions in a series are shown on-screen at once, whether the series is a grid (i.e., a response is required for each item; also known as forced choice) or a checklist (i.e., only true items are checked off). The preamble and the response options appear once for the entire list, so their words were counted only once. For grid series, the core question—the wording that is unique to each item in the series—contributes all its words to the count. However, because checklists involve a simpler task and, according to Smyth et al. (2006) and Callegaro et al. (2015), are about twice as fast to complete, their word count was halved to compensate.

We decided not to include word counts for "respondent notes"—i.e., extra information on-screen to help the respondent understand how to answer—because analysis showed that any notes in a given survey were usually used in multiple related questions (e.g., how to enter price data or a list of e-cigarette brands) and would certainly not be read more than once, and we believed that many respondents would not read a note even the first time it appeared. Such notes are uncommon, as well as generally very short, and would have essentially no effect on the overall word count even if included. We also omitted entirely from word counts any questions that required a verbal answer—i.e., open-ended questions such as "specify other brand" after a list of brand names—because ITC experience is that few respondents answer such questions.

The word count for each survey version was then divided by a probable reading speed (in words per minute [wpm]) to compute a completion time estimate. Reading speeds, both silent and aloud, are available online (Brysbaert 2019; Nation 2009), and our first attempts at this method used such estimates. The computation of an overall survey completion time estimate proceeded exactly the same as for the Read-Throughs method: each user group's Min and Max

Table 1. Sample questions with word counts and estimated completion time for experienced survey panelists.

Question Text	Word Count	Reading Time (sec)	Notes on Word Count Method
Speed: 234 wpm			
Select all that apply. What effect has the coronavirus outbreak had on your smoking? Because of it, I quit smoking.	16		**Checklist series.** • Respondents check off as many items in the series as apply. There are no responses to contribute to the word count. • Preamble is counted in full. • Items in the series after the first have their word count halved.
Because of it, I'm smoking less.	6/2=3		
Because of it, I'm smoking more.	6/2=3		
It has had no effect at all on my smoking.	10/2=5		
Total for series question	**27**	**7.0**	
In the last 6 months, have you noticed advertising or information that talks about the dangers of smoking cigarettes, or encourages quitting, in any of the following places: On television? 1 Yes 2 No 8 Prefer not to answer 9 Don't know	34		**Yes/no grid series.** • Full word count used for all items, not just the first. • Yes & no counted only for the first item, since they are unlikely to be read more than once. • Non-responses – prefer not to answer, don't know – are never counted.
On radio?	2		
In newspapers or magazines?	4		
On posters or billboards?	4		
Total for series question	**44**	**11.2**	
Please tell us to what extent you agree or disagree with each of the following statements. Smoking cigarettes helps you control your weight. 1 Strongly agree 2 Agree 3 Neither agree nor disagree 4 Disagree 5 Strongly disagree 8 Prefer not to answer 9 Don't know	38		**Rating grid series.** • Full word count used for all items. • Responses counted only for the first item.
Smoking cigarettes is an important part of your life.	9		
Cigarette smoke is dangerous to non-smokers.	6		
Total for series question	**53**	**13.5**	
Does the e-liquid that you currently use most contain nicotine? 1 Yes 2 No 8 Prefer not to answer 9 Don't know *The e-liquid could come in disposable e-cigarettes, cartridges, pods or bottles.*	**14**	**3.6**	**Standalone question.** • All words in question and responses are included in full. • Respondent notes are not counted (here 11 words, in italics).
Now some questions about both heated tobacco products and ordinary cigarettes. Compared to smoking ordinary cigarettes, how harmful do you think it is to use a heated tobacco product? 1 Much less harmful than smoking ordinary cigarettes 2 Somewhat less harmful than smoking ordinary cigarettes 3 Equally harmful to smoking ordinary cigarettes 4 Somewhat more harmful than smoking ordinary cigarettes 5 Much more harmful than smoking ordinary cigarettes 8 Prefer not to answer 9 Don't know	**68**	**17.4**	**Standalone question.** • All words in question and responses are included in full. • These responses are much longer and contribute substantially to the word count and to the completion time.
Total for example questions	**206**	**52.7 sec, < 1 min**	

versions were timed separately, then averaged to provide a mean for that group. Then each user group mean was weighted by the proportion of respondents who had belonged to that group in the previous wave.

However, we quickly realized that online reading speeds were providing timing overestimates. In our experience, survey respondents who are members of survey panels, and therefore practiced at completing surveys, read faster than some reading speed estimates (e.g., for our Japan Wave 3 Survey, we obtained a reading speed of 219 words per minute for an inexperienced group of research assistants vs. 235 wpm for actual panelists). Therefore, we reverse engineered a more accurate reading speed from the previous wave's data, using essentially the same computation that provides the time estimate for the current survey but calculating speed from known time, rather than time from known speed. Specifically, we generated the user group pathways and Min/Max versions for the previous wave and obtained the mean word counts for each group. We knew the actual completion time for each group from the previous wave's data set and therefore could calculate the speed for each as Total Word Count / Completion Time = Words per Minute. These speeds were weighted by each group's proportion in the previous wave's sample.

Method III: Question Weights

We had long realized that the contribution of a question to overall survey completion time is primarily determined by how many people answer it. The longest question will have little effect on completion time if almost no one answers it. This is, of course, why branching is the most common problem in estimating survey completion time.

In Methods I and II, we accounted for branching to a degree when a question's inclusion depended on user group, by weighting of user group times, but that was a fairly rough estimate. When the inclusion was instead the downstream result of a gateway-question response not determined by user group, it was—via the Min/Max versions—effectively assigned a weight of 0.5, equivalent to assuming that it would be answered on average by half the respondents. For many such questions the result was an overestimate, and for others, it might be an underestimate.

To avoid this, we worked out a simpler and more accurate approach. We realized that it is possible to weight each individual question by the proportion of respondents who are likely to answer it. Clearly the accuracy of the method depends on how accurate this weighting information is, but the advantages are great: it completely eliminates the need to create user group pathways and Min/Max versions and to weight their time estimates; it automatically includes all such weighting and improves on it.

For ITC, the best source of question weights is almost always the data from the previous wave's survey. Since surveys of successive waves are usually very similar—comparability over waves being essential to tracking change over time—most of the questions in the current survey will have appeared in the previous wave. Unless the waves are quite far apart in time, and product usage

has changed substantially, the proportion of respondents who answered a question in the last wave is generally a fairly accurate estimate of the proportion who will answer the same question in the current survey.

Even for questions that are entirely new in the current survey, it is often possible to determine the proportion of respondents who **would have** answered them if they had been present in the last survey. This applies when new questions are filtered on questions that existed in the last survey. If question B is new this wave, but it is filtered on question A being answered with yes, and question A appeared in the last wave's survey, we can obtain the proportion of respondents who answered yes to question A last wave and assume that a similar proportion will answer question B this wave.

Where questions are completely new and not directly filtered on previous-wave questions, previous data will often still provide a guide. For example, a new series of questions asked respondents to speculate on their behavioral response if menthol was banned from tobacco products. The weight for these questions was based on the proportion of respondents in the prior wave who said their usual brand was a menthol brand. Where there are no previous data for a given country, other countries' data can help. For instance, to estimate the proportion of respondents who would have quit smoking between Wave 1 and Wave 2 of the Netherlands survey and would therefore be answering the new questions for quitters, we looked at the proportion of respondents who had quit between Waves 1 and 2 in several other countries.

External sources, i.e., published data from other studies, can also be used. It is, to use tobacco control examples, often possible to obtain information about prevalence of product usage, health conditions, cessation methods, and so on, for a given population. Obviously, though, the more precise one's knowledge, the more accurate the final timing estimate.

As was done for Method II, Word Counts, we realized that we could obtain a customized completion speed from the previous wave, since completion time is included in the respondent data. Completion speed varies by country, language and survey firm (Loosveldt and Beullens 2013; Trauzettel-Klosinski and Dietz 2012), so using the speed from the previous wave should increase accuracy of the estimate. The weighted-question computation was still appropriate, but instead of using speed to compute completion time, we used completion time to compute speed. Specifically, word counts for the past-wave survey were weighted by the proportion of respondents who **did** answer each question, and the resulting total word count was divided by the known median completion time to provide a speed in words per minute. This speed could then be inserted into the formula for the current survey: Total current-survey weighted word count / estimated median speed from past wave = estimated median completion time in minutes.

Table 2. Validation data for two automated estimation methods.

	Method		
Survey	II. Word Counts (min)	III. Question Weights (min)	Actual Duration (min)
Japan Wave 3	36.3	35.3	35.8
Netherlands Wave 2	32.2	32.0	28.7

Assuming access to question proportions from the previous wave, this is a very fast method of computing an estimate of survey completion time. Once all the proportions have been entered (we import them directly from SAS [Statistical Analysis Software] results), the calculation completes in seconds. And since the word counts for each question come from the SIS database records, once a survey has been shortened through question deletions, the program automatically obtains new word counts for the new estimate, again in seconds. As expected, this increased efficiency has been greatly appreciated and much used by the survey development team; for instance, the Japan Wave 4 survey was shortened in stages, and the length in minutes re-estimated 3 times, before it was deemed short enough.

Validating the Automated Methods

The automated methods are very successful in saving staff time and allowing repeated estimates. But we wanted to know: were they accurate? We have begun to collect validation data. To date, we have estimated completion times for three in-progress surveys: Spain Wave 3, New Zealand Wave 4, and Japan Wave 4. None of these surveys has completed fieldwork, so we cannot use them to check accuracy. Therefore, for validation we turned to recent surveys that have completed fieldwork and for which we already have actual completion times. In order to use the previous wave's reading speed and its proportions as weights, we needed projects that had completed at least two waves. Japan Waves 2 and 3 (fielded in 2018–2019 and 2020) and the Netherlands Waves 1 and 2 (both fielded in 2020) met this condition.

We used Japan Wave 2 data to estimate the median completion time of Japan Wave 3. That is, from the Wave 2 data, we derived user group weights and reading speed for the Word Counts method, and question weights and reading speed for the Question Weights method, as described above in the relevant sections. With the Word Counts method, the estimate was 36.3 minutes, and with the Question Weights method, it was 35.3 minutes. The two estimates were both within 1 minute of the actual median time of 35.8 minutes.

The Netherlands Wave 1 data were similarly used to estimate the median completion time of the Netherlands Wave 2. The time estimates were 32.2 (Word Counts) and 32.0 min (Question Weights); actual median time was 28.7 minutes. The reason for the somewhat larger discrepancy between actual and estimated times for the Netherlands (about 3 minutes; relative to Japan's less than 1 minute) is unknown. We theorized that at Wave 1 the survey was new to

all, but at Wave 2, only 4 months later, it was familiar to the many recontacted respondents, resulting in greater speed. However, this was not supported by the data: the speed for recontact respondents was only minimally faster than that of replenishment respondents. In any case, an estimate within 3–4 minutes of actual length was considered satisfactory.

For both countries, the Question Weights and Word Counts methods were equally accurate. This does not support our belief that averaging the Min and Max versions in effect overestimates the number of respondents who answer most downstream questions and therefore overestimates completion time. However, this hypothesis is likely the case for some other surveys, even if not for the two we used for validation. It is also possible that in some cases the Min/Max mean underestimates the responses for downstream questions. The underlying issue here is still true—that giving Min and Max versions equal weight (in the Word Counts method) makes assumptions that can reduce accuracy, assumptions which the Question Weights method does not require.

Discussion

Despite the importance of estimating survey completion time, there is little guidance in the research literature or from online sources such as the websites of survey research firms. The few methods documented online may work well for simple surveys. But as Puleston (2012) acknowledges in presenting his method, assessing survey length is very difficult when branching is complex. The Qualtrics method accommodates branching to a very limited degree by accounting for both the shortest and longest paths. No other method offers any adjustment for branching. Therefore, the ITC Project, with extensive skip patterns in computerized surveys, had to invent its own method for length estimation.

All three of ITC's methods do allow for branching complexity. Our first two approaches, Read Throughs and Word Counts, determine the most common user group pathways and weight them by their respondent proportions, and Min and Max versions are computed separately for each user group. The third method, Question Weights, eliminates the need for user groups and pathways entirely by simply weighting each individual question by its expected respondent proportion.

The second and third methods, by automating what is otherwise a very labor-intensive process, also produce estimates quite quickly, which allows the swift re-estimation of length whenever changes to the survey have been made.

Our methods work well for repeated surveys, whether longitudinal or cross-sectional (e.g., surveillance surveys), because past surveys provide the necessary reading speed and respondent proportions. However, in the absence of previous survey data, other sources of both types of information are necessary.

Common reading speeds can be found online (Brysbaert 2019; Nation 2009) for a variety of languages. However, the published reading speeds for English vary a great deal (Brysbaert 2019) and reading speeds specifically for surveys do not seem to have been published to date. Carver (1992) suggested that reading speed depends on the task and that reading in order to answer multiple choice questions, which he called 'Gear 2', would be done at about 200 wpm.

Large and complex surveys are almost invariably conducted with survey firms' panels, and panelists often complete surveys for the compensation, so they are motivated to invest as little time as possible for their reward. We would expect speeds greater than Carver's 200 wpm, and indeed, our survey data consistently show this. As a check on the effect of survey completion experience, we compared the speed of inexperienced readers (five research assistants who work for ITC and are familiar with tobacco topics, but do not complete surveys as part of their work) on about one-fifth of the Japan 3 survey to the speed of the actual respondents in that survey; the inexperienced readers were a little slower, with a speed of 219 wpm vs. the 235 for panelists.

The proportion of respondents likely to answer a question is a more difficult value to obtain, since it is dependent on the specific question and the specific respondent group. However, it is still possible to come up with likely proportions by turning to external sources such as other country-specific data sets, either the researcher's own or free-access data sets. Naturally, the more recent and the more group-appropriate the data available, the more accurate the resulting time estimate will be.

Our use of reading speed as the sole speed determinant gives the somewhat misleading impression that we are not allowing for deliberation time and response time as well. The other methods cited, by Versta, Qualtrics, and Question Science, allow specifically for response time and transition between questions, which we do not. Their calculations also differ for questions with different levels of cognitive complexity. We originally believed it would be necessary to create a more complex algorithm, and we could have done so, since we have the necessary information in our SIS database, but the accuracy of our estimates has persuaded us that it is unnecessary to complicate our method.

On the other hand, one of the benefits of our two automated methods is that the algorithm can be adapted as necessary for different survey conditions if desired. We have already allowed for faster completion of checklist series, as mentioned above. We could similarly allow for increased cognitive demand—e.g., questions with longer words or longer sentences, or those that require mental deliberation—by assigning greater word-count weights. Currently, we do not include time for open-text responses, because such fields are seldom completed by respondents, but it would not be difficult to add these to the algorithm.

Table 3. Characteristics of three ITC methods for estimating survey completion time.

Method feature	I. Read-Throughs	II. Word Counts	III. Question Weights
User group pathways	• Created manually • Very time-consuming	• Automated • Somewhat time-consuming	• Unnecessary • Questions, not user groups, are weighted
Min/Max versions	• Partial • Created manually • Very time-consuming	• Exact • Automated • Fast	• Unnecessary • Questions, not user groups, are weighted
Accommodation of complex branching	• Partial • By creation of multiple versions	• Partial • By creation of multiple versions	• Comprehensive • Individual questions are weighted
Read-T hrough time	20–40 hours	N/A	N/A
Weighting	Approximate, by user group	Approximate, by user group	More accurate, by question
Knowledge required for weights	Estimated proportions for major user groups	Estimated proportions for major user groups	Estimated proportions for individual questions
Retiming after changes to survey	Not feasible due to time costs	Almost instantaneous	Almost instantaneous
Adaptability to future conditions	Requires complete redo	Requires complete redo	Reading speed, word counts, weights can be easily modified

We may find that certain countries, certain languages, or panelists from specific survey firms, require faster or slower reading speeds. As long as we derive our reading speed from the previous wave of the same project, the effects of country, survey language and panel are automatically included in the calculation. But if we wanted to add a new language to an existing project, as when we added Chinese and English to our Malaysia project, we could compensate for different reading speeds quite easily. Speed conversion factors are available for many languages (Brysbaert 2019; Loosveldt and Beullens 2013; Puleston 2012; Trauzettel-Klosinski and Dietz 2012).

Although we have only applied our automated methods, Word Counts and Question Weights, to online surveys, they would likely have been just as relevant and useful in ITC's early days, when most of our surveys were CATI (i.e., telephone) . The main difference would be speed, since the interviewer reads the question aloud—which is much slower than reading silently—and the respondent also answers aloud. There is also opportunity in telephone surveys for the respondent to ask the interviewer for clarification, and such exchanges could add substantially to the survey duration. None of this would preclude using our methods; the reading speed (which would include the interviewer reading and the respondent answering) would simply be considerably slower. Previous waves could still provide an estimated speed. Reading-aloud speeds are also available, for situations in which no previous waves or relevant data exist (Brysbaert 2019), although they vary a great deal; allowance might also need to be made for interviewer-respondent exchanges, depending on interviewer instructions.

Table 3 presents a summary of the advantages and disadvantages of the three ITC methods.

In summary, previously published methods for timing estimates are likely to work well for simple surveys; the necessary reading speeds can be obtained online and adjusted as necessary to allow for the respondents' degree of survey completion experience. However, complex programmed surveys need a technique that allows for the effects of branching, and in that case, the ITC methods are more appropriate. Question weighting makes for greater accuracy of timing, and if question weights are not available from a previous wave, then approximate weights can be obtained from other sources, which is still likely to be more accurate than any non-weighted method.

Conclusion

The length estimates obtained by these methods are just that: estimates. Simplifications inherent in the automated methods—e.g., not accounting explicitly for deliberation time and response time, assuming that the current wave will be similar to the previous wave in user group distribution—potentially decrease accuracy. But we never expected the results to be minute-perfect. They are, like the original Read-Throughs method, simply guidance for the investigators in selecting content for a survey.

The primary intent was to free up staff time, produce more consistent estimates, and, in particular, to make it possible to compute revised estimates as often as necessary. In those respects, our automation efforts, particularly the Question Weights method, have been enormously successful. That the estimates are also surprisingly accurate is a bonus.

With programmed surveys now employing more complex branching, there is a greater need for timing estimation methods that are applicable and accurate under such conditions. Survey researchers are well aware of the effect of overlong surveys on cost, respondent retention, and data quality. They need to be sure that their surveys are no more than the intended length. The three methods described in this paper are tools that may assist others in achieving that.

Funding

This study was supported by the Canadian Institutes of Health Research (FDN-148477). Additional support to Geoffrey T. Fong was provided by a Senior Investigator Grant from the Ontario Institute for Cancer Research (IA-004).

Ethics

The survey questionnaires of ITC Netherlands (REB#41704), ITC Japan (REB#22508/31428), ITC New Zealand (REB#21211/30726 and REB#42549), and ITC Spain (REB#41105) were cleared by the Research Ethics Board at the Office of Research Ethics, University of Waterloo, Canada. Ethics clearance from each country's internal review board were also obtained:

Japan National Cancer Center (IRB 2021-069) and Osaka International Cancer Institute (IRB 21054) for ITC Japan, University of Otago (IRB 15/126 and IRB 20/020) for ITC New Zealand, and the Spain Hospital Universitari de Bellvitge - Clinical Research Ethics Committee of Bellvitge (IRB PR100/16) for ITC Spain. Ethics clearance at Maastricht University in the Netherlands was waived due to minimal risk.

Declaration of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data statement

In each country participating in the International Tobacco Control Policy Evaluation (ITC) Project, the data are jointly owned by the lead researcher(s) in that country and the ITC Project at the University of Waterloo. Data from the ITC Project are available to approved researchers 2 years after the date of issuance of cleaned data sets by the ITC Data Management Centre. Researchers interested in using ITC data are required to apply for approval by submitting an International Tobacco Control Data Repository (ITCDR) request application and subsequently to sign an ITCDR Data Usage Agreement. The criteria for data usage approval and the contents of the Data Usage Agreement are described online (http://www.itcproject.org).

Submitted: February 09, 2022 EDT, Accepted: September 14, 2022 EDT

REFERENCES

Brysbaert, Marc. 2019. "How Many Words Do We Read per Minute? A Review and Meta-Analysis of Reading Rate." *Journal of Memory and Language* 109: 104047. https://doi.org/10.1016/j.jml.2019.104047.

Callegaro, Mario, Michael H. Murakami, Ziv Tepman, and Vani Henderson. 2015. "Yes–No Answers versus Check-All in Self-Administered Modes: A Systematic Review and Analyses." *International Journal of Market Research* 57 (2): 203–24. https://doi.org/10.2501/ijmr-2015-014a.

Carter, Brittany, James Bennett, and Elric Sims. 2021. "Evaluation of Estimated Survey Duration Equations Using a Health Risk Assessment." *Survey Research Methods* 15 (2): 187–94. https://doi.org/10.18148/SRM/2021.V15I2.7800.

Carver, R.P. 1992. "Reading Rate: Theory, Research, and Practical Implications." *Journal of Reading* 36 (2): 84–95. https://www.jstor.org/stable/40016440.

Fong, Geoffrey T., K. Michael Cummings, Ron Borland, Gerard Hastings, Andrew Hyland, Gary A. Giovino, David Hammond, and Mary E. Thompson. 2006. "The Conceptual Framework of the International Tobacco Control (ITC) Policy Evaluation Project." *Tobacco Control* 15 (Suppl III): iii3–11. https://doi.org/10.1136/tc.2005.015438.

Galesic, Mirta, and Michael Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Public Opinion Quarterly* 73 (2): 349–60. https://doi.org/10.1093/poq/nfp031.

Hoerger, Michael. 2010. "Participant Dropout as a Function of Survey Length in Internet-Mediated University Studies: Implications for Study Design and Voluntary Participation in Psychological Research." *Cyberpsychology, Behavior, and Social Networking* 13 (6): 697–700. https://doi.org/10.1089/cyber.2009.0445.

Hyland, Andrew, Bridget K. Ambrose, Kevin P. Conway, Nicolette Borek, Elizabeth Lambert, Charles Carusi, Kristie Taylor, et al. 2016. "Design and Methods of the Population Assessment of Tobacco and Health (PATH) Study." *Tobacco Control* 26 (4): 371–78. https://doi.org/10.1136/tobaccocontrol-2016-052934.

Loosveldt, Geert, and Koen Beullens. 2013. "'How Long Will It Take?' An Analysis of Interview Length in the Fifth Round of the European Social Survey." *Survey Research Methods* 7 (2): 69–78. https://doi.org/10.18148/SRM/2013.V7I2.5086.

Nation, P. 2009. "Reading Faster." *International Journal of English Studies* 9 (2). https://revistas.um.es/ijes/article/view/90791.

Puleston, Jon. 2012. "How to Calculate the Length of a Survey." In *Question Science*. http://question-science.blogspot.com/2012/07/how-to-calculate-length-of-survey.html.

Qualtrics Support. n.d. "ExpertReview Functionality." https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/quality-iq-functionality/.

———. n.d. "Predicted Duration." https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/survey-methodology-compliance-best-practices/#PredictedDuration.

Smyth, Jolene D., Don A. Dillman, Leah Melani Christian, and Michael J. Stern. 2006. "Comparing Check-All and Forced-Choice Question Formats in Web Surveys." *Public Opinion Quarterly* 70 (1): 66–77. https://doi.org/10.1093/poq/nfj007.

Trauzettel-Klosinski, Susanne, and Klaus Dietz. 2012. "Standardized Assessment of Reading Performance: The New International Reading Speed Texts IReST." *Investigative Opthalmology & Visual Science* 53 (9): 5452–61. https://doi.org/10.1167/iovs.11-8284.

Versta Research Blog. 2011. “How to Estimate the Length of a Survey.” https://verstaresearch.com/newsletters/how-to-estimate-the-length-of-a-survey/.