methods, data, analyses | Vol. 19(1), 2025, pp. 137-165 DOI: 10.12758/mda.2025.09

Improving Understanding of Survey Questions with Multimodal Clarification

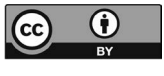
Maura Spiegelman[1] & Frederick Conrad[2]

[1] Independent Researcher
[2] University of Michigan

Abstract

If survey respondents do not interpret a question as it was intended, they may, in effect, answer the wrong question, increasing the chances of inaccurate data. Researchers can bring respondents' interpretations into alignment with what is intended by defining the terms that respondents are at risk of misunderstanding. This article explores strategies to increase alignment between researchers' intentions and respondents' answers by taking advantage of the unique affordances of online surveys compared to paper or other analog formats. Web surveys are often text-based, but allow for the seamless integration of embedded audio material so that users may read, listen to, or both read and listen to survey instructions. Unimodal definitions are either spoken or textual, while multimodal definitions are both spoken and textual. Further, definitions can be designed to take advantage of the affordances of each mode. While *mode-invariant* definitions contain the same words irrespective of whether they are textual or spoken, *mode-optimized* definitions are designed to take advantage of the affordances of written and spoken communication. For example, definitions optimized for textual presentation use fewer words than corresponding mode-invariant definitions and are designed so the key information is visually salient, while definitions optimized for spoken presentation are shorter and more colloquial than corresponding mode-invariant definitions. In this study, both mode-optimized and mode-invariant formats improved alignment. Multimodal, mode-optimized definitions produced improved alignment over both types of unimodal definitions. This study suggests that multimodal definitions, when thoughtfully designed, can improve data quality in online surveys without negatively impacting respondents.

Keywords: question definitions, questionnaire instructions, audio input, multimodal input, data quality, web survey



Ensuring that survey respondents interpret survey questions as their authors intended is a prerequisite for producing high quality data. Otherwise, respondents may, in effect, answer a different question than the one the researchers believed they were asking, potentially resulting in inaccurate answers. One way to align respondents' and researchers' interpretations is to clarify terms that may not map cleanly to respondents' circumstances. For example, if a respondent is unsure whether to include TV programming streamed to their laptop computer when answering a question about their recent TV watching, defining exactly what is meant by TV watching should resolve the respondent's uncertainty about how to answer. Explicitly clarifying terms can help assure that respondents understand survey questions—whether asked by interviewers or self-administered—as intended and in a way that fits their situation. In everyday conversation, participants *ground* what has been said (Clark, 1996) by discussing the speaker's intentions until both parties agree they understand each other well enough to accomplish the goals of the conversation. The benefits of grounding meaning have been explored in survey interviews, self-administered online questionnaires, virtual interviews, and speech dialog systems (see Conrad & Schober (2021) for a summary and review). This prior research concerns the delivery of unimodal, that is, solely spoken or solely textual definitions. However, there may be value in exploring multimodal delivery of definitions. In educational psychology, researchers have found that multimodal communication can improve comprehension and information retention compared to unimodal communication in some, though not all, circumstances (Moreno & Mayer, 2002; Mousavi et al., 1995). This paper builds upon the research in both conversational grounding and multimodal communication to explore whether multimodal definitions, that is, definitions that are both spoken and textual, can improve the quality of survey responses relative to unimodal definitions (either spoken or textual) or no clarification. This study also tests the conditions under which multimodal definitions might be most effective, that is, multimodal definitions that are identically worded across the two modes or that exploit the affordances of each mode in which they are implemented.

Notes

This work was conducted at the Joint Program in Survey Methodology, University of Maryland, USA and was supported a Rensis Likert Dissertation Research Award, University of Michigan. The authors declare there is no conflict of interest.

Direct correspondence to

Maura Spiegelman
E-mail: mspiegelman@gmail.com

Survey Definitions

Surveys ask respondents about conditions or situations of varying complexity, clarity, and familiarity. When respondents' understanding of ideas or terms is different from what researchers intend, data quality is likely to suffer unless understanding can be aligned (Conrad & Schober, 2000; Schober & Conrad, 1997; Schober et al., 2018). For example, the concept of how many people live in a household is straightforward for respondents in most living situations. However, for respondents with a child living away at college, it is not clear whether they should include their child in their response, potentially introducing bias if such misalignments occur in one direction. For the portion of respondents who have children living at college, a definition of who should be counted as living in the household can correct respondents' misconceptions, leading to an accurate answer.

Although it can improve comprehension of questions and response accuracy, providing definitions generally increases the amount of time needed to answer survey items (Conrad & Schober, 2020, 2001; Conrad et al., 2007; Schober et al., 2004; Schober & Conrad, 1997), although West et al. (2018) found no effect on response times. Respondents need to listen to or read the definitions and incorporate them into their interpretation of the question—which could potentially reduce their satisfaction with the interview experience, potentially reducing completion rates, and likely inflating sampling variance.

While definitions can certainly help align respondents' and researchers' understanding of survey questions, providing definitions will only provide these benefits if respondents use them. One reason respondents might not use a definition is if it is hard to understand. This might be the case if, for example, the definition is presented in textual form and the respondent is not a strong reader, or because a spoken definition is so complicated and long that a momentary lapse in the respondent's attention might result in their not understanding the definition. The first of these might be addressed by presenting the definition in both textual and spoken forms, a multimodal definition. Not only might this increase the chances that the content of the definition is interpretable by most respondents, but it emphasizes the importance of the definition by conveying it in two ways. The second issue, that the definition is long and complicated, might be addressed through improved design, such as simplifying the content and presenting the definition in a way that is most appropriate to its mode.

Multimodal Communication

Multimodal communication typically involves the simultaneous presentation of information in two or more channels of communication. In the case of survey questionnaires this can involve the way content such as questions and definitions is presented, how respondents report their answers, or both (Johnston,

2008). For example, online data collection can be designed so that respondents can both read a textually presented question and hear the corresponding spoken question; or to enable respondents to answer by either typing/clicking or speaking (e.g., "Type 1 or say 'Yes'"). This study explores the former: multimodal presentation of information, in particular, definitions of survey concepts.

Research on multimodal communication has combined spoken information with a variety of visual presentations and has found that, under the right circumstances, combining audio and visual material can be more effective than using only one mode. For example, in educational psychology, researchers observed that presenting students with both audio and visual material is more effective pedagogically than using only spoken communication for certain types of information and presentations. For example, Moreno and Mayer (2002) noted that participants showed higher levels of retention and were more effective at applying information in a new context (rather than simply recalling it) when taught in a multimodal, rather than unimodal spoken format. Mousavi, Low, and Sweller (1995) found that students needed less time to accurately solve geometry problems using combined diagrams—which are visual—with orally presented verbal information rather than textually presented verbal material that competes with diagram processing for limited visual attention. By comparing sequential and simultaneous presentation, they attributed these results to the relatively low cognitive load of using multiple communication channels, due to partial independence of visual and verbal processing (Mayer, 2014; Sweller et al., 2019).

Extending these findings to processing survey questions, multimodal presentation could help respondents understand and apply survey definitions, presumably improving the quality of their answers. By dividing content between the textual and spoken material, the amount of content presented in either mode is reduced relative to unimodal presentation. This division by mode is particularly helpful for processing spoken information which is ephemeral and (unless it is audio-recorded and can be replayed) will be gone after it is presented. In contrast, textual information is persistent, (i.e., remains visible over time) and can be read while the spoken information is presented, after it is presented, or both. Thus, textual information does not need to be stored in working memory initially, the way spoken information does, because it is preserved externally (i.e., on the screen). Moreover, working memory is hypothesized to consist of separate storage mechanisms for textual ("visuo-spatial") and spoken ("phonological") information, coordinated by a central executive system (Baddeley, 1992; also see Dumas et al., 2009), suggesting that it is well suited to multimodal presentation of information.

However, in the psychological literature, the benefits of multimodal communication depend on the extent to which the information in the different modes is redundant and conveys the same information. When information is simultaneously conveyed both orally and textually, redundancy can potentially reduce

comprehension. For example, when an animated technical explanation was combined with either spoken narration only or identical, simultaneous spoken and textual narration, the latter treatment resulted in poor comprehension, evident in reduced retention and transfer of information (Mayer et al., 2001). That is, participants who were exposed to redundant spoken and written words and a complementary animation had lower comprehension than participants who were exposed to only spoken words and a complementary animation. Note that many of these studies involve visual stimuli that created substantial cognitive demands, such as combinations of written instructional text, numerical tables, and graphs or diagrams (e.g., Kalyuga et al., 2004). Since survey researchers generally attempt to convey less complicated material to respondents, rarely requiring animated instruction, these findings are unlikely to limit the effectiveness of multimodal material in surveys, but they do point out that redundant content across modes can degrade respondents' ability to process additional information.

When spoken and textual content does not consist of exactly the same words but instead conveys the same underlying message, this kind of semantic redundancy does not seem to harm comprehension the way literal redundancy does (Kalyuga et al., 2004). Mild levels of redundancy, for example key words or phrases, have been shown to increase retention (Mayer & Johnson, 2008). This suggests that multimodal definitions of survey concepts can yield higher rates of comprehension when the text emphasizes somewhat different ideas than the spoken content, rather than simply duplicating the information. More specifically, identical spoken and textual definitions may reduce comprehension, while complementary definitions seem likely to improve comprehension.

This study tests whether multimodal definitions for key concepts in survey questions can improve the quality of responses (i.e., their alignment with definitions) compared to unimodal (either spoken or textual) definitions. Two types of multimodal definitions were tested: mode-invariant definitions, with fully redundant spoken and textual information (i.e., the same *words* presented visually and via speech) and mode-optimized definitions, designed specifically for each mode with partially redundant content, i.e., the same *concepts* conveyed textually and orally using complementary wording and exploiting the affordances of each mode. If both types of multimodal definitions outperform unimodal definitions, this would likely be due to the partial independence of communication channels. If only mode-optimized multimodal definitions outperform unimodal definitions, this would likely be due to the literal redundancy of mode-invariant definitions, which provide no additional information or perspective on the underlying concept from their multimodality, even potentially interfering with comprehension.

Methods

Experimental Design

Respondents completed a web survey in one of seven experimental conditions distinguished by the type of definitions made available: (1) none (i.e., the control condition), (2) textual, mode-invariant, (3) textual, mode-optimized, (4) spoken, mode-invariant, (5) spoken, mode-optimized, (6) multimodal (i.e., both textual and spoken), mode-invariant), or (7) multimodal (i.e., both textual and spoken), mode-optimized. Irrespective of the mode(s) and optimization of definitions, all survey questions were presented textually. These seven conditions comprise a fraction of all possible combinations of mode, multimodality, and format, but allow for the most important comparisons: whether multimodal definitions were more effective, i.e., promoted greater alignment of respondents' understanding of each question with the question's intended interpretation, than unimodal definitions and whether mode-optimized, multimodal definitions—in which the information in each mode was complementary and relatively non-redundant—were more effective than mode-invariant, multimodal definitions.

Respondents were asked to provide numeric responses to 15 survey questions, each accompanied by a definition for the key concept (except in the control condition). Definitions were either "inclusive" (five questions) or "exclusive" (seven questions). Inclusive definitions were designed to expand the scope of what behaviors could be counted as examples of the concept in question (for example, *including* commuting when reporting the amount of work for which the respondent was paid), and exclusive definitions were designed to reduce the scope (for example, *excluding* streamed or recorded content when reporting on the amount of television watched; Schober & Conrad, 2000). To promote the questionnaire's coherence, questions on similar topic areas were grouped together, for example, hours spent watching television and listening to the radio. Thus, all respondents viewed questions in the same order. Finally, all respondents were asked a series of debriefing questions about their demographics and experience during the study. All surveys were identical except for the type of definition made available.

Mode-invariant definitions were designed to emulate the format of data collection instruments from many government statistical agencies. The definitions in these types of surveys contain detailed information, and when presented in textual format, often appear as a dense paragraph; they are not designed for respondents to identify the subcomponents most relevant to their situations. When these same definitions are read aloud, they do not flow like a conversation or other spoken communication. Instead, the experience is reminiscent of questionnaires designed to be self-administered (either on paper or online) but which are administered by an interviewer over the phone to some respondents. For multimodal mode-invariant definitions, identical wording was used for both the spoken and textual components leading to fully redundant information. For

multimodal mode-optimized definitions, spoken optimized and textual optimized components were presented together.

Mode-optimized definitions were designed to be easier for respondents to either read or listen to and to help respondents identify relevant information by following best practices of written and spoken communication. For textual mode-optimized definitions, factors known to facilitate text comprehension (White, 2012) were used: bolded text to draw attention to key words and phrases, bullets and other organizational devices to divide text into logical groupings. For each question, mode-optimized textual definitions had lower Flesch-Kincaid grade level reading scores (shorter sentence length and fewer syllables per word) than their mode-invariant counterparts (Flesch, 1948).

For spoken mode-optimized definitions, the scripts were designed to follow best practices for spoken communication. For example, in order to facilitate comprehension in spoken mode-optimized definitions, extraneous information that was included in their mode-invariant counterparts was removed (Sweller et al., 1990). Shorter spoken definitions are also less taxing on respondents' working memory, and require relatively little effort to comprehend compared to longer, mode-invariant definitions (Leahy & Sweller, 2011). For each question, mode-optimized spoken definitions were shorter in duration than their mode-invariant counterparts (an average of 11.4 seconds compared to 23.1 seconds). Spoken mode-optimized definitions were read aloud, audio-recorded, and played back by the researchers to judge their flow and ease of comprehension, then adjusted iteratively, if needed in the researchers' judgment. The displayed text and scripts for all mode-invariant and mode-optimized definitions are shown in the Appendix and screenshots of each condition are shown in Table 1.

Data Collection

We implemented the experimental conditions—the seven different web-based questionnaires—in Qualtrics, using TurkPrime (now CloudResearch) to recruit and manage participants from Amazon Mechanical Turk (MTurk). Each condition was posted as a separate "task" within MTurk, with identical descriptions, and participants were only eligible to complete one of these tasks, essentially randomizing respondents across treatment groups. A $1 incentive was provided to respondents upon completion of the survey. The median completion time for the surveys, including debriefing and other non-experimental questions, was just under 10 minutes, resulting in a median hourly rate of $6.17. The University of Maryland Institutional Review Board approved this study, and we collected data in the summer of 2018.

In total, 1,014 respondents completed the study. For the 12 experimental survey questions, 11,988 total observations were retained for analysis after removing impossible and implausible values (for example, reports of participating in

any given activity for close to 168 hours per week since there are, in total, only 168 hours per week). The distribution of respondents and observations by experimental condition is shown in Table 2.

Table 1 Selected screenshots by experimental condition

Experimental condition	Screenshot
Control	In the past 7 days, how many hours of television did you watch?
Spoken mode-invariant	In the past 7 days, how many hours of television did you watch? Play for more information: 
Spoken mode-optimized	In the past 7 days, how many hours of television did you watch? Play for more information: 
Textual mode-invariant	In the past 7 days, how many hours of television did you watch? Watching television includes any programs other than films. This may include sitcoms, dramas, news, sports, and reality shows. Television is watched on a television set at the time it is broadcast and does not include programming recorded with a DVR, viewed on-demand, or streamed. Include content viewed on a television set only and exclude any content viewed on a computer or mobile device.
Textual mode-optimized	In the past 7 days, how many hours of television did you watch? • Content is broadcast. Exclude DVRed, on-demand, and streamed shows. • TV set. Exclude shows watched on a computer or mobile device. • TV shows. Exclude films, even if watched while they air.

Table 1 (continued)

Multimodal mode-invariant	In the past 7 days, how many hours of television did you watch? Play for more information:  Watching television includes any programs other than films. This may include sitcoms, dramas, news, sports, and reality shows. Television is watched on a television set at the time it is broadcast and does not include programming recorded with a DVR, viewed on-demand, or streamed. Include content viewed on a television set only and exclude any content viewed on a computer or mobile device. ☐
Multimodal mode-optimized	In the past 7 days, how many hours of television did you watch? Play for more information:  • Content is broadcast. Exclude DVRed, on-demand, and streamed shows. • TV set. Exclude shows watched on a computer or mobile device. • TV shows. Exclude films, even if watched while they air. ☐

Table 2 Sample sizes and number of observations by experimental condition

	# Respondents	# Observations
Control	104	1,239
Spoken mode-invariant	200	2,356
Spoken mode-optimized	162	1,920
Textual mode-invariant	101	1,196
Textual mode-optimized	80	952
Multimodal mode-invariant	160	1,890
Multimodal mode-optimized	207	2,435
Total	1,014	11,988

The number of respondents and observations varies by condition for two primary reasons. First, a subset of participants were recruited for a pilot test with control, spoken mode-invariant, textual mode-invariant, and multimodal

mode-optimized definitions. When it was evident that the procedure worked as expected, these cases were pooled with newer cases. In addition, more respondents were recruited into treatment groups with spoken components (both unimodal and multimodal) under the assumption that not all respondents would comply with instructions and play the spoken definitions.

Analytic Strategy

Alignment of Question Interpretation and Intended Meaning

Because different questions asked about different reference periods and different types of activities and measured the target behavior on different scales, responses could not be averaged in raw form. Moreover, for some questions—namely those for which definitions were inclusive—higher numeric responses indicated consistency with definitions, while for other questions—namely those for which definitions were exclusive—lower numeric responses indicated consistency with definitions. So that we could compare across questions and conditions, we converted responses to each question to a z-score, trimmed to +4 and -4, and then multiplied these z-scores by -1 for questions with exclusive definitions. Because of this trimming, the mean z-score per question deviates slightly from 0. This conversion allows responses to be pooled across questions, using a standard scale, and for results from each condition to be pooled, namely higher values indicate greater alignment with definitions (more standard deviations from the mean response) while lower values indicate that responses are less aligned with definitions, irrespective of whether a definition was inclusive or exclusive.

We used a general linear mixed model in SAS 9.4 to compare the effects of different definition treatments by modeling z-scores while accounting for clustering of observations within respondents. Questions (denoted with subscript q) are nested within respondents (denoted with subscript i). Both questions and respondents were given random intercepts, allowing for baseline differences in question difficulty and respondents' behavior, though respondents are treated as random effects and questions as fixed effects.

$$Y_{iq} = \gamma_{00} + \gamma_{10}\,(D_i) + U_{i0} + e_{iq}$$

We conducted F-tests at the observation level, rather than the respondent level, when analyzing alignment of responses with definitions (D_i) using type 3 F-tests of fixed effects unless otherwise stated. For pairwise comparisons of point estimates, we use Student's t-tests unless otherwise stated.

Respondent Use of Definitions

Some respondents did not fully comply with the instructions to attend to definitions, raising the possibility that the effect of definition type on alignment might be stronger for those who comply. To address this, observations can be compared both overall and by examining only observations based on (inferred) compliance with the experimental treatment. For spoken definitions, we captured whether audio clips were fully played, and for textual definitions, we estimated the time that would be required to read a particular question and its associated definition and compared this to the actual time each respondent spent on the page. Note that compliance is relevant for the control group even though control respondents were not provided with any definitions; these respondents were expected to spend sufficient time on each page to read survey questions.

To measure respondents' exposure to spoken definitions, the online survey captured how many times a spoken definition was fully played by using embedded JavaScript code. It is important to note that this measure could not record whether a respondent's audio was muted, nor whether they truly attended to the spoken information, but instead serves as a proxy for respondent compliance in a self-interview setting that involved auditory information. Respondents could play a definition by clicking the "play" icon (right arrow) in a standard media bar. For spoken definitions, whether as part of a unimodal or multimodal format, a given response was considered "compliant" if the audio file was fully played.

We inferred whether respondents read the textual information they were presented by determining if the time spent on any given question was at least as long as the estimated reading time for that text, in which case they were determined to have complied with instructions. We calculated the reading time threshold for each question and each textual definition (in the relevant conditions) by counting the words and multiplied the counts by 200 msec. This is the average reading speed, according to Carver (1992), for adult Americans when reading to retain content for relatively short intervals, as is needed when answering survey questions[1]. Thus, the word count for the control group and groups with unimodal spoken definitions were identical (question word count only), and the word count for the textual only and multimodal definition conditions were identical for mode-invariant versions (question and definition word count) as was also the case for the mode-optimized versions (question and definition word count). However, like the proxy for spoken definition compliance, this criterion does

[1] Conrad et al. (2017) and Zhang and Conrad (2014) used 300 msec/word for similar purposes. However, their thresholds were intended to account for reading plus thinking time, so faster responses could be considered speeding. Because the tasks in our study were less cognitively burdensome, we selected a more conservative threshold in order to avoid inflating our estimates of the impact of compliance on alignment of question interpretation and intended meaning.

not guarantee that respondents truly attended to and absorbed the textual information presented to them. Eye-tracking could help determine whether respondents viewed the textual information, for example, whether they fixated on textual information in left-to-right, top-to-bottom order or whether they skipped or sped through information. However, even knowing what they looked at would not capture whether they deeply comprehended and internalized the information or merely scanned the text. In a self-interview setting, time per page is the best available measure of reading time and thus proxy for respondent compliance. Note that for both spoken and textual definitions, compliance was treated as a binary metric for which a given observation either met compliance criteria or did not.

Results

Alignment of Question Interpretation and Intended Meaning

The average *z*-scores by mode and optimization of definitions are shown in Table 3. *Z*-scores indicate the number of standard deviations by which observations for a given definition type varied from the average response across all questions and definition modes. Higher values indicate more alignment with definitions, while lower values indicate less alignment with definitions. For example, the average *z*-score for responses to questions in the control group was about -0.13, indicating that those responses were less aligned with definitions than the average response by 0.13 standard deviations.

Table 3 Mean *z*-score by definition mode and optimization

Definition mode	Optimization	Mean z-score
Control (no definition)	n/a	-0.126
Spoken	All	-0.009
	Mode-invariant	-0.025
	Mode-optimized	0.012
Textual	All	-0.014
	Mode-invariant	-0.032
	Mode-optimized	0.008
Multimodal	All	0.041
	Mode-invariant	0.013
	Mode-optimized	0.063

Responses to questions with multimodal definitions were more aligned with definitions than the average response by about 0.041 standard deviations, and they were significantly more aligned than responses to questions with only unimodal definitions. That is, average *z*-scores were higher for the multimodal group than unimodal textual definitions ($t(996) = 2.33$, $p = .020$), and unimodal spoken definitions ($t(1002) = 2.56$, $p = .011$). Overall, definition mode was a significant predictor of the degree to which responses were aligned with definitions ($F(3,994) = 11.37$, $p < .001$). As expected, responses were least aligned with definitions for the control group, under which no definitions were available.

The effectiveness of multimodal definitions appears to be driven by their optimization. That is, average *z*-scores were higher for the multimodal mode-optimized definitions than for unimodal mode-invariant definitions, both spoken ($t(1003) = 3.39$, $p < .001$) and textual $t(997) = 2.98$, $p = .003$). *Z*-scores for multimodal mode-invariant definitions were higher, though not significantly so, than *z*-scores for unimodal mode-optimized definitions both spoken ($t(997) = 1.87$, $p = .062$) and textual ($t(990) = 1.59$, $p = .111$). They were marginally higher than for multimodal, mode-invariant definitions ($t(1000) = 1.80$, $p = .072$) making it somewhat ambiguous to what extent the presence alone of complimentary, rather than redundant, multimodal information can improve data quality.

Respondents' Use of Definitions

Compliance Rates

For the four definition types with a spoken component (spoken mode-invariant, spoken mode-optimized, multimodal mode-invariant, multimodal mode-optimized), compliance with spoken definitions (that is, fully playing a definition's audio file) ranged from the relatively low rate of 29% for multimodal mode-invariant definitions to 47% for spoken mode-optimized definitions (Table 4).

For all four treatment groups in which spoken definitions were available, compliance was highest with the first definition presented in the survey, ranging from 61% for multimodal mode-invariant to 84% for multimodal mode-optimized. Across all spoken mode conditions, the compliance rate for the first survey question was significantly higher than the compliance rate for every other question at the $p < .05$ level. A steady decline in compliance might reflect respondent fatigue; the reason for this abrupt drop is unclear but could have occurred if respondents noted there were no direct repercussions of answering without playing the entire spoken definition. This drop-off in compliance occurred both overall and within each of the 4 treatment groups with spoken definitions. For both mode-invariant and mode-optimized definitions, compliance with instructions to play the audio was higher for respondents who only received spoken definitions, rather than multimodal respondents who were encouraged to both read and listen to definitions. Compliance was higher for the spoken mode-optimized

group than for either of the multimodal conditions, and higher for the spoken mode-invariant than multimodal mode-invariant group. However, it should be noted that these overall compliance rates were less than 50% for each condition.

Differences in compliance between unimodal and multimodal groups may be driven by the presence of an alternative way of acquiring multimodal definition content. For respondents in unimodal spoken groups who were inclined to use definitions in their responses, their only choice was to listen to spoken definitions. Respondents in multimodal groups could have given responses consistent with definitions by reading textual definitions, even if they did not fully play an audio clip.

Table 4 Compliance with spoken and textual portions of definitions by question number and definition type

Question		Textual mode-invariant	Textual mode-optimized	Spoken mode-invariant	Spoken mode-optimized	Multimodal mode-invariant	Multimodal mode-optimized
1	Listened	n/a	n/a	73%	78%	61%	84%
	Read	55%	88%	n/a	n/a	97%	99%
2	Listened	n/a	n/a	59%	54%	40%	42%
	Read	21%	40%	n/a	n/a	58%	58%
3	Listened	n/a	n/a	41%	50%	32%	32%
	Read	23%	76%	n/a	n/a	53%	84%
4	Listened	n/a	n/a	44%	52%	30%	38%
	Read	42%	61%	n/a	n/a	54%	74%
5	Listened	n/a	n/a	39%	46%	27%	38%
	Read	31%	66%	n/a	n/a	50%	82%
6	Listened	n/a	n/a	36%	45%	26%	26%
	Read	36%	70%	n/a	n/a	62%	82%
7	Listened	n/a	n/a	26%	33%	24%	26%
	Read	21%	64%	n/a	n/a	44%	84%
8	Listened	n/a	n/a	34%	40%	22%	29%
	Read	29%	64%	n/a	n/a	49%	74%
9	Listened	n/a	n/a	37%	38%	26%	25%
	Read	33%	74%	n/a	n/a	50%	74%
10	Listened	n/a	n/a	40%	46%	20%	35%
	Read	29%	59%	n/a	n/a	42%	67%
11	Listened	n/a	n/a	25%	39%	21%	26%
	Read	18%	48%	n/a	n/a	41%	67%
12	Listened	n/a	n/a	31%	40%	21%	25%
	Read	17%	70%	n/a	n/a	43%	78%
Overall	Listened	n/a	n/a	39%	47%	29%	35%
	Read	29%	65%	n/a	n/a	53%	78%

Comparing overall compliance for mode-optimized and mode-invariant definitions, the rate was higher for respondents in the former than latter (47% and 39%, respectively, for unimodal; 35% and 29%, respectively, for multimodal), though this difference was only significant when comparing the two unimodal conditions.

Compliance with textual definitions followed a similar pattern. Again, this type of compliance was operationalized as at least as much time spent on a given question as the estimated reading time for the question and definition text. Compliance was significantly higher for the first question than every other subsequent question ($p < .001$) for each definition type, similar to the pattern shown for compliance with spoken definitions (see Table 4). In addition, compliance rates differed by condition for each pairwise comparison between the 4 groups with textual definitions. In particular, the 78% compliance rate for multimodal mode-optimized definitions was significantly higher than the 65% compliance rate for textual mode-optimized definitions ($t(544) = 3.54$, $p < .001$), which was significantly higher in turn than the 53% compliance rate for multimodal mode-invariant definitions ($t(544) = 2.74$, $p = .006$), which was significantly higher than the 29% compliance rate for textual mode-invariant definitions ($t(544) = 6.54$, $p < .001$). So, compliance was highest for mode-optimized definitions. For both mode-invariant and mode-optimized definitions, compliance was higher for multimodal than unimodal definitions. However, as with spoken definitions, all mode-optimized textual definitions had fewer words than all mode-invariant textual definitions, and presumably as a result, shorter estimated reading times. As a result, length and optimization are confounded and prevent us from distinguishing the effects of optimization per se from reduced text on compliance.

Compliance with multimodal definitions depends on whether respondents only read, only listened to, or both read and listened to definitions. However, in this study it is important to note that the duration of each spoken definition was at least as long as the estimated reading time for the corresponding textual definition, so all respondents who fully listened to a multimodal definition's spoken component were coded as being in full compliance.

Alignment of Question Interpretation and Intended Meaning When Respondents Access Definitions

We have suggested that definitions—in either mode—can help align respondents' understanding of questions with the questions' intended meaning, However, increased alignment could be due to the mere presence of the definitions rather than the content of the definitions. For example, multimodal definitions may signify to respondents that the information is important, or because the content of the definitions is better understood by respondents. These possible explanations cannot be disentangled without examining responses while considering whether individuals accessed the definitions that were available to them.

We can treat noncompliant responses in two different ways. They may be dropped from analysis, for example, a response to a textual definition that did not meet the criteria for reading can be omitted entirely. Alternatively, that response effectively had the same de facto treatment as a control group response and could be analyzed with the others from that group, though, respondents may have read part of a definition or read all text more quickly than the estimated reading speed threshold, so their experiences may not be identical to those of actual control group participants. Any observations for which a respondent in a multimodal condition did not both fully listen to and read the definition can be analyzed with the control, unimodal textual or unimodal spoken groups (although the latter is theoretical given that audio clips were longer than estimated reading duration so playing an audio file will lead to a compliant reading classification even if the respondent did not read the definition). As with spoken definitions, observations were categorized based on whether they fully met compliance criteria, so observations for which definitions may have been partially played or read were considered noncompliant and analyzed accordingly. Observations that did not meet criteria for the control group, that is, the amount of time spent on the page was less than the compliance cutoff for fully reading the question text, were excluded since they could not be treated as compliant with any treatment group. Table 5 shows the average *z*-score for both methods of categorizing compliant and noncompliant responses.

Table 5 Mean *z*-score by definition mode and optimization for compliant responses and de facto treatment

Definition mode	Optimization	Mean z-score (compliant responses only)	Mean z-score (by de facto treatment)
Control (no definition)	n/a	-0.129	-0.071
Spoken	All	0.076	0.075
	Mode-invariant	0.073	0.071
	Mode-optimized	0.079	0.079
Textual	All	0.059	0.076
	Mode-invariant	0.018	0.041
	Mode-optimized	0.082	0.093
Multimodal	All	0.163	0.163
	Mode-invariant	0.079	0.079
	Mode-optimized	0.217	0.217

When limiting analysis to only observations that met compliance criteria, responses to questions with multimodal definitions were more aligned with definitions than the average response by about 0.16 standard deviations (Table 5).

That is, average z-scores were higher for definitions that were multimodal than unimodal textual ($t(691) = 2.86$, $p = .004$) and unimodal spoken ($t(619) = 2.78$, $p = .006$) definitions.

As with the analysis of all observations (irrespective of compliance), this difference is driven by multimodal mode-optimized definitions. Responses to these questions were more aligned with the underlying concepts than the average response by about 0.22 standard deviations. That is, average z-scores were higher for the mode-optimized multimodal group than for all other conditions: spoken mode-invariant ($t(649) = 3.41$, $p < .001$), spoken mode-optimized ($t(601) = 3.26$, $p = .001$), textual mode-invariant ($t(959) = 3.46$, $p < .001$), textual mode-optimized ($t(614) = 2.97$, $p < .001$), and multimodal mode-invariant groups ($t(643) = 2.78$, $p = .006$) groups. The increased data quality with multimodal definitions is primarily attributed to presenting complementary, rather than redundant, information.

Looking at de facto treatment, that is, categorizing responses based on the treatment they effectively received rather than the group to which they were originally assigned, we observed a similar pattern. Answers reported when respondents were compliant with multimodal definitions were significantly more aligned with definitions than each of the other de facto definition types. That is, average z-scores were higher for the multimodal group than when the effective treatment was textual definitions ($t(2567) = 2.89$, $p = .004$), spoken definitions ($t(1517) = 2.88$, $p = .004$), and the control treatment with no definitions ($t(1713) = 8.99$, $p < .001$). Overall, de facto definition mode significantly predicted the degree to which responses were aligned with definitions ($F(3,2057) = 37.57$, $p < .001$).

Observations produced when respondents complied with multimodal mode-optimized definitions were significantly more aligned with definitions than each of the other types of definition. That is, with de facto categorization, average z-scores were higher for the multimodal mode-optimized group than when the treatment received was mode-invariant multimodal ($t(1566) = 2.87$, $p = .004$), spoken mode-invariant ($t(1585) = 3.55$, $p < .001$), spoken mode-optimized ($t(1467) = 3.34$, $p < .001$), textual mode-invariant ($t(2517) = 3.80$, $p < .001$), or textual mode-optimized ($t(2471) = 3.29$, $p = .001$) definitions, as well as the control treatment with no definitions ($t(2471) = 8.99$, $p < .001$). Once again, the effectiveness of multimodal definitions is due to the use of complementary, rather than mode-invariant, instructions.

Respondent Burden

Survey respondents' acceptance of multimodal clarification, particularly compared to unimodal formats, is crucial if multimodal definitions are realistically to be deployed in production research. If respondents react negatively to

multimodal communication, potentially abandoning the survey, these perceptions must be weighed against the increase in data quality brought about by this approach to clarification in online surveys, at least as demonstrated here.

To explore this, we asked respondents to rate their satisfaction with the survey and how burdensome they found the process; we also measured the amount of time respondents spent on each page of the web survey. The number of seconds respondents spent on the 12 survey items with definitions is shown in Table 6. Comparing mean response times with a Tukey adjustment, respondents with spoken mode-invariant definitions spent significantly more time completing the questionnaire than spoken mode-optimized or textual respondents. Respondents with multimodal definitions spent significantly more time than textual mode-invariant respondents, but not significantly longer than other types of definitions.

Table 6 Time spent on 12 definition questions by definition mode and optimization (in seconds)

Definition mode	25th percentile	Median	75th percentile	Mean	SD
Control (no definition)	72	93	136	107	51
Spoken mode-invariant	142	288	410	299	205
Spoken mode-optimized	133	198	279	237	175
Textual mode-invariant	90	130	209	169	120
Textual mode-optimized	105	160	195	192	219
Multimodal mode-invariant	140	222	363	264	171
Multimodal mode-optimized	142	192	281	249	235

However, a longer survey duration does not necessarily indicate that respondents *feel* more burdened. Respondents who were presented with unimodal spoken and multimodal definitions were asked to describe how burdensome they found the process of accessing spoken definitions (Not at all burdensome, slightly burdensome, moderately burdensome, very burdensome, extremely burdensome). This question was designed to measure the effort required to play spoken definitions, and so is not applicable to respondents in the control group, who saw no definitions, or respondents who were assigned to view unimodal textual definitions, since textual definitions appeared by default with no additional action needed from respondents. Overall, respondents did not indicate that playing definitions was notably burdensome. Most reported that accessing definitions was not at all burdensome (61%) or slightly burdensome (21%), while few found the process to be very (4%) or extremely (3%) burdensome. Multimodal respondents had the option of reading definitions without deliberately playing spoken definitions, so it is notable that the perceived level of burden did not vary between these four

types of definitions ($\chi^2(4) = 1.33$, $p = .856$). That is, respondents found the process of playing definitions to impose little burden regardless of whether they had another option for obtaining that information.

We also asked respondents to rate their overall satisfaction with the survey (Overall, how satisfied were you with your experience when responding to this survey? Very dissatisfied, somewhat dissatisfied, neither dissatisfied nor satisfied, somewhat satisfied, very satisfied). Respondents provided positive feedback about their survey experience. Almost half (48%) were very satisfied, and one-third (33%) were somewhat satisfied. The remainder were neither dissatisfied nor satisfied (14%), somewhat dissatisfied (4%), or very dissatisfied (1%). This distribution differed by definition mode ($\chi^2(12) = 32.28$, $p < .001$), with relatively higher proportions of respondents who were exposed to unimodal spoken and multimodal definitions reporting they were very satisfied when compared to unimodal textual respondents and those who were not shown definitions (53%, 52%, 39%, and 31%, respectively). Together, these suggest that multimodal definitions can be implemented in online surveys without overburdening respondents or otherwise causing a negative survey experience.

Discussion

Why were multimodal definitions—especially when optimized—more effective than unimodal definitions? On the one hand, if speech and written text are processed at least somewhat independently, then any multimodal communication (fully redundant or complementary) would improve comprehension when compared to unimodal communication, since more information would be available to a respondent, potentially compensating for an attentional lapse and providing more opportunity to internalize the content. Alternatively, if redundant definitions are less effective than complementary (i.e., mode-optimized) multimodal definitions at conveying the intended meaning of the question, then the latter should improve response quality more than unimodal definitions. Responses based on multimodal definitions were more aligned with survey concepts than responses based on unimodal definitions, and this was driven by mode-optimized definitions. This suggests that it is primarily complementary, rather than redundant multimodal content, that is effective (and supporting the idea that conveying identical information through multiple channels can reduce—or at least not facilitate—comprehension). The increased alignment with multimodal, and particularly mode-optimized multimodal definitions, appeared when comparing all observations. While the presence of multimodal definitions (regardless of whether they were used) increased data quality, these cues alone did not prompt respondents to attend to definitions; instead, the effect of multimodal definitions was sharpened when analyses were restricted to all compliant obser-

vations, as compliance with instructions about how to use definitions provided a purer measure of their impact on comprehension.

Overall, compliance was higher for mode-optimized than for mode-invariant definitions. Because the features of optimization (e.g., concision, increased salience of key material) were presented as a package and not experimentally varied, we cannot determine which of these features may have been most responsible for its benefits in multimodal definitions. In fact, for all definition types, compliance was highest for the first survey item than for subsequent questions, but respondents were willing to play spoken definitions in a survey mode that typically includes only text. While compliance could perhaps increase with shorter or more visually appealing definitions (two features that differentiated mode-invariant and mode-optimized presentations), these findings are promising for the efficacy of multimodal definitions, particularly given the strict compliance criteria for spoken definitions (i.e., respondents were required to fully play an audio clip). If respondents only minimally complied with multimodal definitions, or if they provided negative feedback about their experiences, those drawbacks would have to be carefully weighed against the increased alignment with definitions for responses to multimodal instructions. Instead, these results suggest that respondents do not find multimodal definitions to be burdensome, are willing to comply with instructions to both read and listen to them, and will apply these definitions to their formulation of survey responses. In an online survey, multimodal definitions can improve data quality without negatively impacting respondents. It is reassuring that the presence of spoken information did not decrease respondent satisfaction, and in fact, respondents who were presented with spoken definitions either alone or as part of multimodal definitions reported the highest levels of satisfaction.

Future Research

The sample for this study was drawn from Amazon Mechanical Turk. This study provides a proof-of-concept that multimodal definitions can improve data quality, but more research is needed to determine the degree to which these findings can be replicated in samples from other sources and whether unpaid participants are as amenable to play integrated audio clips in an online survey. We were unable to capture the type of device on which surveys were completed, for example, a laptop computer or smartphone, and these findings may vary further by device type.

Compliance was inferred without truly knowing whether respondents attended to definitions. For spoken definitions, compliance may have been underestimated for respondents who partially listened to spoken definitions. For textual definitions, compliance may have been over- or under-estimated if

respondent reading speed was miscalculated by our use of response latency as a measure, or if they simply did not attend to their screen. For spoken definitions, a more robust tracking mechanism could assess how much of spoken definitions were played. For textual definitions, a lab study that tracks respondents' eye movements could more accurately measure whether on-screen text was read. All of these limitations can be addressed in straightforward ways in follow-up studies.

This study focuses on a fundamentally visual type of survey: a textual web survey, in which spoken definitions were embedded in some experimental conditions. While text is persistent, spoken communication is ephemeral, so improvements in data quality due to adding text to a communication format that is typically spoken (such as telephone surveys) is likely to be greater than the improvements due to adding spoken information to a communication format that is typically textual (such as web surveys). While some spoken surveys do have an added textual component (e.g., show cards), that text has typically been used to present response options, rather than questions and definitions. Telephone surveys rarely include a textual component, and this gap is particularly ripe for exploration. Respondents completing a telephone survey are often using an internet-enabled device. A respondent could receive text instructions from an interviewer, e.g., via a text message, particularly for survey items for which the underlying constructs are nuanced or potentially counterintuitive. While the effectiveness of multimodal communication may differ across these scenarios, particularly given differences in communication norms and respondent expectations, these uses warrant further exploration of multimodal definitions given its richness, the likelihood it will become more practical with technological advances, and the possibility that respondents will be more satisfied with their experience knowing they understand what they are asked.

References

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading, 36*(2), 84–95. http://www.jstor.org/stable/40016440

Clark, H. H. (1996). Community, commonalities, and communication. In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 324–355). Cambridge University Press.

Conrad, F. G., Couper, M. P., Tourangeau, R., & Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods, 11*(1), 45–61. https://doi.org/10.18148/srm/2017.v11i1.6304

Conrad, F. G., & Schober, M. F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly, 64*(1), 1–28. https://doi.org/10.1086/316757

Conrad, F. G., & Schober, M. F. (2021). Clarifying question meaning in standardized interviews can improve data quality even though wording may change: A review of the evidence. *International Journal of Social Research Methodology, 24*(2), 203–226. https://doi.org/10.1080/13645579.2020.1824627

Conrad, F. G., Schober, M. F., & Coiner, T. (2007). Bringing features of human dialogue to web surveys. *Applied Cognitive Psychology, 21*(2), 165–187. https://doi.org/10.1002/acp.1335

Dumas, B., Lalanne, D., & Oviatt, S. (2009). Multimodal interfaces: A survey of principles, models and frameworks. In D. Lalanne & J. Kohlas (Eds.), *Human machine interaction* (pp. 3–26). Springer. https://doi.org/10.1007/978-3-642-00437-7_1

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233. https://doi.org/10.1037/h0057532.

Johnston, M. (2008). Automating the survey interview with dynamic multimodal interfaces. In F. G. Conrad & M. F. Schober (Eds.), *Envisioning the survey interview of the future*. Wiley. https://doi.org/10.1002/9780470183373

Kalyuga, S., Chandler, P., & Sweller, J. (2004). When redundant on-screen text in multimedia technical instruction can interfere with learning. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*(3), 567–581. https://doi.org/10.1518/hfes.46.3.567.50405

Leahy, W., & Sweller, J. (2011). Cognitive load theory, modality of presentation and the transient information effect. *Applied Cognitive Psychology, 25*(6), 943–951. https://doi.org/10.1002/acp.1787

Mayer, R. E. (Ed.). (2014). Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning* (2nd ed., pp. 72–103). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93*(1), 187–198. https://doi.org/10.1037/0022-0663.93.1.187

Mayer, R. E., & Johnson, C. I. (2008). Revising the redundancy principle in multimedia learning. *Journal of Educational Psychology, 100*(2), 380–386. https://doi.org/10.1037/0022-0663.100.2.380

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology, 94*(1), 156–163. https://doi.org/10.1037/0022-0663.94.1.156

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology, 87*(2), 319–334. https://doi.org/10.1037/0022-0663.87.2.319

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *The Public Opinion Quarterly, 61*(4), 576–602. https://doi.org/10.1086/297818

Schober, M. F., Conrad, F. G., & Fricker, S. S. (2004). Misunderstanding standardized language in research interviews. *Applied Cognitive Psychology, 18*(2), 169–188. https://doi.org/10.1002/acp.955

Schober, M. F., Suessbrick, A. L., & Conrad, F. G. (2018). When do misunderstandings matter? Evidence from survey interviews about smoking. *Topics in Cognitive Science, 10*(2), 452–484. https://doi.org/10.1111/tops.12330

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. *Journal of Experimental Psychology: General, 119*(2), 176–192. https://doi.org/10.1037/0096-3445.119.2.176.

Sweller, J., Van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*, 261–292. https://doi.org/10.1007/s10648-019-09465-5

West, B. T., Conrad, F. G., Kreuter, F., & Mittereder, F. (2018). Can conversational interviewing improve survey response quality without increasing interviewer effects? *Journal of the Royal Statistical Society Series A: Statistics in Society, 181*(1), 181–203. https://doi.org/10.1111/rssa.12255

White, S. (2012). Mining the text: 34 text features that can ease or obstruct text comprehension and use. *Literacy Research and Instruction, 51*, 143–164. https://doi.org/10.1080/19388071.2011.553023

Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods, 8*, 127–135. https://doi.org/10.18148/srm/2014.v8i2.5453

Appendix

Definitions by Survey Question and Treatment Group

Question	In the past 7 days, how many hours of television did you watch?
Mode-invariant definition	Watching television includes any programs other than films. This may include sitcoms, dramas, news, sports, and reality shows. Television is watched on a television set at the time it is broadcast and does not include programming recorded with a DVR, viewed on-demand, or streamed. Include content viewed on a television set only and exclude any content viewed on a computer or mobile device.
Textual mode-optimized definition	• Content is broadcast. Exclude DVRed, on-demand, and streamed shows. • TV set. Exclude shows watched on a computer or mobile device. • TV shows. Exclude films, even if watched while they air.
Spoken mode-optimized definition	By television, we mean content watched on a TV set at the time it is broadcast. Exclude streamed, on –demand, and DVRed shows and anything watched on a computer or mobile device. Exclude films.
Inclusive/exclusive	Exclusive definition

Question	In the past 7 days, for how many hours did you listen to the radio?
Mode-invariant definition	Listening to the radio includes listening to programming transmitted and received through an antenna. Available stations and reception are restricted by signal strength and listener location. Programs are listened to live, that is, as they air, rather than played later by the listener, such as with podcasts and other downloadable content. Programming can include talk-based content, such as news or sports, but does not include music even if accessed by antenna.
Textual mode-optimized definition	• Antenna. Only count local stations through over-the-air access, not satellite or internet. • Live Content. Exclude podcasts or other content played on-demand. • Talk. Programming includes news, sports, and talk shows. Exclude music.
Spoken mode-optimized definition	By radio, we mean local programming listened to live, over-the-air, but not podcasts, satellite radio, or internet radio. Include news, sports, and talk shows, not music.
Inclusive/exclusive	Exclusive definition

Question	In the past 7 days, for how many hours did you use e-mail?
Mode-invariant definition	E-mail use includes composing, sending, and reading messages, as well as managing an inbox. Count time spent using an online mailbox, desktop mailbox, or mobile application, and do not count time spent reading attachments or linked content in a browser. Only count e-mail use when connected to the internet through a wired or wireless (Wi-Fi) connection. Exclude email use involving a cellular connection such as 3G or 4G. Exclude offline use.
Textual mode-optimized definition	Exclude • E-mail using a cellular network such as 3G or 4G. • Reading attachments or linked content. Include • Composing, sending, reading, and sorting messages. • Use of a Wi-fi or wired connection.
Spoken mode-optimized definition	By e-mail use, we mean writing, reading, sending and sorting messages. Only count time using an application, not time spent reading attachments or linked content. Only count access through a wired or Wi-Fi connection, so exclude cellular networks such as 3G and 4G.
Inclusive/exclusive	Exclusive definition

Question	Excluding e-mail use, in the past 7 days, for how many hours did you use the internet?
Mode-invariant definition	People may use the Internet to carry out personal or professional tasks and activities. Exclude internet use involving a cellular connection such as 3G or 4G. Include active tasks such as reading news articles, posting in online forums, and playing online games. Exclude passive tasks that do not involve direct attention or engagement such as streaming videos or music.
Textual mode-optimized definition	• Connection. Count Wi-fi and wired connections only. Exclude cellular networks such as 3G and 4G. • Active use. Count tasks such reading articles, posting in forums, and playing online games. Do not count passive activities such as streaming videos or music.
Spoken mode-optimized definition	By Internet, we mean access through a wired or Wi-Fi connection, so exclude cellular networks such as 3G and 4G. Only count time on tasks such as reading or posting content or playing games, and do not count passive activities such as streaming videos or music.
Inclusive/exclusive	Exclusive definition

Question	In the past 7 days, how many hours did you work in total?
Mode-invariant definition	Work is paid employment performed for an employer or, if self-employed, for oneself. Count paid internships or apprenticeships. Count time directly spent on work activities, such as time at an office or work site, as well as commuting to and from an office.
Textual mode-optimized definition	Include • Paid work or self-employment. • Work as an employee or paid intern. • Time at work and commuting to and from work.
Spoken mode-optimized definition	By work, we mean a paid job or internship, or self-employment. In addition to time at a job site, work includes commuting time.
Inclusive/exclusive	Inclusive definition

Question	In the past 7 days, how many miles did you travel by vehicle?
Mode-invariant definition	Vehicles have two or more wheels, are used for ground transportation and can include cars, trucks, taxis, buses, trains, subways, trams, motorcycles, and bicycles. All miles spent in a vehicle, regardless of seat location, should be considered. Miles as both a driver and passenger should be included.
Textual mode-optimized definition	• Vehicle. Count any ground travel by vehicle, including cars, trucks, taxis, buses, motorcycles, trains, subways, and bicycles. • Role. Count miles as both driver and passenger.
Spoken mode-optimized definition	By travel, we mean miles as a driver or passenger in a vehicle such as a car, truck, taxi, bus, train, subway, tram, motorcycle, or bicycle.
Inclusive/exclusive	Inclusive definition

Question	In the past year, how many plane trips did you take?
Mode-invariant definition	A plane trip begins at liftoff and ends at touchdown. If multiple legs (liftoffs and touchdowns) are involved, such as with non-direct or multi-city flights, each is counted separately. Similarly, for roundtrip flights, outbound and return flights are each counted separately, and all legs are counted separately.
Textual mode-optimized definition	• Count each leg of a trip separately. • Count roundtrip flights separately.
Spoken mode-optimized definition	Count each component of a trip separately. For example, layovers and roundtrip flights should be counted as multiple plane trips.
Inclusive/exclusive	Inclusive definition

Question	In the past 30 days, how many times have you had food or drinks at a restaurant?
Mode-invariant definition	Restaurants are dining establishments at which food and/or beverages are served. Include sit-down establishments, restaurants with and without table service, fast food restaurants, coffee shops and cafes, bars and pubs, food trucks, and street vendors. Food may be eaten at the restaurant or elsewhere, if ordered for take-out, to-go, or delivery.
Textual mode-optimized definition	• Type. Count sit-down restaurants, fast food, coffee shops, bars, food trucks and street vendors. • Location. Count dine-in, take-out, to-go orders, and delivery.
Spoken mode-optimized definition	We mean sit-down restaurants, fast food, coffee shops, bars, food trucks and street vendors. We mean dine-in, take-out, to-go orders, and delivery.
Inclusive/exclusive	Inclusive definition

Question	How many pairs of shoes do you own?
Mode-invariant definition	Shoes are footwear worn primarily outdoors and secured to a foot with some type of fastener, such as laces, zipper, Velcro, clasps, or buckles. For this question, footwear designed primarily for indoor use such as slippers does not qualify. For this question, non-fastening shoes such as flip flops, slides, clogs, pumps, and other unsecured footwear do not qualify.
Textual mode-optimized definition	Exclude shoes • Worn indoors, including slippers. • Unsecured, such as flip flops, slides, clogs, pumps, etc. Include shoes • Worn outside • Secured with laces, zippers, Velcro, clasps, buckles, etc.
Spoken mode-optimized definition	By shoes, we mean footwear worn primarily outside that can be secured with fasteners such as laces, zippers, Velcro, clasps, or buckles. Do not count unsecured footwear such as flip flops, slides, clogs, pumps, and other unsecured footwear.
Inclusive/exclusive	Exclusive definition

Question	How many hours of rest do you get on a typical weekday?
Mode-invariant definition	Include time spent in a state of sleep or time that has the potential to become sleep. This includes overnight sleep and daytime naps, as well as time when sleep is not necessarily intended, such as during class or a meeting, while reading a book, or while watching television.
Textual mode-optimized definition	• Time of day. Count evening and daytime rest. • Sleep state. Count time spent asleep or when sleep is possible, such as sitting while reading a book or watching television.
Spoken mode-optimized definition	By rest, we mean time when you are asleep or could fall asleep, such as sitting while reading a book or watching TV.
Inclusive/exclusive	Inclusive definition

Question	In the past 7 days, how many hours did you exercise?
Mode-invariant definition	Exercise is physical activity that results in an elevated heart rate. This can include vigorous activities such as running or biking and less vigorous activities such as walking, climbing up or down stairs, and yoga. Exercise can be performed alone, such as swimming or biking, or with a group or team, such as basketball or tennis. Include all physical activities, regardless of how long they lasted.
Textual mode-optimized definition	• Activities. Count all activities that result in an elevated heart rate. • Duration. Count all physical activities, regardless of how long they lasted.
Spoken mode-optimized definition	By exercise, we mean activities that result in an elevated heart rate, regardless of the duration of each activity.
Inclusive/exclusive	Inclusive definition

Question	In the past 7 days, how many caffeinated drinks did you have?
Mode-invariant definition	Caffeine is a stimulant often found in cacao plants and a variety of beverages. Common caffeinated beverages include coffee, tea, and sodas. While caffeinated beverages may be consumed in any amount or container size, for this question, 8 fluid ounces of a caffeinated beverage is one caffeinated drink.
Textual mode-optimized definition	• Count every 8 ounces as one drink. • Count coffee, tea, soda, and other caffeinated beverages.
Spoken mode-optimized definition	By caffeinated drinks, we mean 8 ounces of caffeinated beverages such as coffee, tea, and soda.
Inclusive/exclusive	Inclusive definition

Question	In the past 7 days, how many hours did you exercise?
Mode-invariant definition	Exercise is physical activity that results in an elevated heart rate. This can include vigorous activities such as running or biking and less vigorous activities such as walking, climbing up or down stairs, and yoga. Exercise can be performed alone, such as swimming or biking, or with a group or team, such as basketball or tennis. Include all physical activities, regardless of how long they lasted.
Textual mode-optimized definition	• Activities. Count all activities that result in an elevated heart rate. • Duration. Count all physical activities, regardless of how long they lasted.
Spoken mode-optimized definition	By exercise, we mean activities that result in an elevated heart rate, regardless of the duration of each activity.
Inclusive/exclusive	Inclusive definition

Question	In the past 7 days, how many caffeinated drinks did you have?
Mode-invariant definition	Caffeine is a stimulant often found in cacao plants and a variety of beverages. Common caffeinated beverages include coffee, tea, and sodas. While caffeinated beverages may be consumed in any amount or container size, for this question, 8 fluid ounces of a caffeinated beverage is one caffeinated drink.
Textual mode-optimized definition	• Count every 8 ounces as one drink. • Count coffee, tea, soda, and other caffeinated beverages.
Spoken mode-optimized definition	By caffeinated drinks, we mean 8 ounces of caffeinated beverages such as coffee, tea, and soda.
Inclusive/exclusive	Inclusive definition