Improving the Question Appraisal System (QAS): Moving Further Away from Black Magic and Black Boxes

Ashley Schaad[3], Matt Jans[2], Melinda Scott[2]
[3]InMoment, 424 Fairfax Ave, Fayetteville, NC 28303
[2]ICF, 530 Gaither Road, Rockville, MD 20850

Abstract
Questionnaire design can be the least transparent survey lifecycle phase, often being conducted by one person (or a very small group), under tight timelines, and without any qualitative or quantitative pretesting. Some researchers still view it as "black magic" that can only be accomplished by highly trained and skilled survey methodologists. For many surveys, the best possible scenario is to have a trusted questionnaire designer review their questionnaire, potentially providing justification with their revisions. Regardless, the process often remains a "black box" that is difficult to audit or replicate. The Questionnaire Appraisal System (QAS-99) was developed to a) make this process replicable and transparent, and b) allow questionnaire revision by survey staff with less training and experience. The QAS consists of eight steps focusing on question characteristics that can lead to difficulty answering accurately, such as the question's readability, instruction presence and complexity, implicit assumptions in the question, and topic sensitivity. The QAS-04 improved the appraisal process by adding steps to assess question translatability, cross-cultural assumptions, and challenges beyond language, and a step to assess issues across questions within the instrument. This paper presents new developments in the QAS process that incorporate a) a questionnaire-level details and flow review to assess the entire instrument, and b) a step for each question reviewed that assesses whether the information requested would have ever been encoded by the respondent. The new review stage and new question-specific step, along with the original QAS-99 and QAS-04 steps, were incorporated into a single Excel file that reviewers used to complete the assessment. This innovation will be discussed in the context of time-sensitive questionnaire development, the overall survey development process, and survey transparency.

Key Words: question pretesting, questionnaire design, Question Appraisal System (QAS)

1. Problems with Questionnaire Design

In recent years, the survey research field, and AAPOR specifically, have taken major steps to promote and normalize transparency of the methods used to collect, analyze, and report data.[1] However, questionnaire design remains one of the least transparent aspects of a survey's lifecycle, and one of the most difficult to make more transparent. This is partly because it is often conducted by one person (or a very small group of people), under tight timelines, and without any qualitative or quantitative pretesting, leaving little

[1] https://www.aapor.org/Transparency_Initiative.htm

in terms of processes or artifacts to make transparent. In some contexts, questionnaire development is still viewed as "black magic" that can only be accomplished by highly trained and skilled survey methodologists. Not only does this "black box" approach lead to revisions and decisions that are not replicable, it can lead to errors of omission when a questionnaire reviewer fails to catch issues that would have been captured through a more systematic and auditable approach.

There are many ways to evaluate and improve questionnaire quality, including both quantitative and qualitative methods (Czaja, R., 1998; Beatty, Collins, Kay, Padilla, et al, 2019). Quantitative pretesting methods often include field pilot tests, behavior coding, and embedded question wording experiments. Qualitative pretesting methods include expert review, cognitive interviews, and interview observation. The quality and transparency of these methods depends largely on the tools and documentation used during their implementation.

1.1 Question Appraisal System (QAS) History and Goals

A major development in pretesting methods that addresses some of the transparency and quality challenges is the Question Appraisal System (QAS), which is a set of steps designed to identify and isolate various cognitive and implementation issues that are commonly seen in draft questions. It provides a framework that helps to ensure important review criteria (e.g., timeframe references and question sensitivity) are not missed, and creates documentation of the identified problems for each question. This structure naturally lends transparency and replicability to the review process.

1.1.1 QAS Origins (QAS-99 and Precursors)

For the purposes of this project we began with the QAS-99 (Willis & Lessler, 1999) because it was the most detailed publicly available version of the QAS that we could find. QAS-99 was refined from earlier question appraisal systems (Forsyth & Hubbard, 1992; Lessler and Forsyth, 1996) by the Research Triangle Institute (RTI) as a tool to simplify and error-proof the review of new questions proposed for the Behavioral Risk Factor Surveillance System (BRFSS) surveys. The QAS-99 was designed to be a systematic draft question evaluation tool that is easy to follow and understand, with a user manual spelling out specific review criteria for users to identify common sources of question problems that can result in response error. Mapped to stages of the question-answer response process, QAS-99 divides the question review into 8 steps assessing Reading, Instructions, Clarity, Assumptions, Knowledge/Memory, Sensitivity/Bias, Response Categories, and Other, to document any issues that do not fall into the previous categories. Within each step are subcodes to break down the components of that review category. For example, the Reading step includes "what to read," "missing information," and "how to read." The interested reader can review Lessler and Willis (1999) for detailed descriptions of QAS-99 steps and subcodes. Earlier QAS models additionally included codes that focused on the nature of the response task (i.e. mnemonic or judgement processes).

Rather than a system for reliably coding specific types of error accurately, the authors describe QAS-99 as "a series of fishing nets—if one net misses, another one may 'make the catch'"(Lessler and Willis, 1999, p 3-2). Suggested uses include not only identifying potential sources of response error and improving questions, but flagging questions for further testing, and using the tool for collaborative review. This approach ensures that major question issues are identified, and that QAS results will be useful in practice, but leaves in question the reliability of the coding system.

1.1.2 QAS Grows to Include Cross-Cultural Considerations (QAS-04)
In response to the increasing need for multilingual surveys, QAS-04 (Dean, Caspar, McAvinchey, Reed, & Quiroz, 2005) is ideal for surveys that incorporate parallel question development early in the survey lifecycle, alerting survey designers to potential issues that arise in cross-cultural/translated survey administration. In addition to the eight original QAS-99 review steps, QAS-04 includes Cross-Cultural Considerations (to assess questions for inappropriate or ineffective cross-cultural references), Potential Translation Problems (to identify potential problems translating questions into the languages required for the survey), and Cross-Question issues (which helps identify potential conflicts or confusion due to differences in questions across the entire questionnaire, such as question placement, data collection mode, inconsistency with other questions, content of previous questions affecting the intended meaning, skip pattern problems, and formatting).

Table 1 includes QAS-99 and QAS-04 review steps and their overlap. For a full understanding of QAS history, readers should also review Lessler and Forsyth (1996) for the original cognitive issues included in QAS coding schemes and the evolution of the system.

Table 1: QAS-99 Steps and QAS-04 Steps

		Steps included in…	
QAS Step	*Step Definition*	*QAS-99*	*QAS-04*
Reading	Determine if it is *difficult for the interviewers to read the question* uniformly to all respondents	X	X
Instructions	Look for problems with any *introductions, instructions, or explanations* from the *respondent's* point of view	X	X
Clarity	Identify problems related to communicating the *intent* or *meaning* of the question to the respondent	X	X
Assumptions	Determine if there are problems with *assumptions* made or the *underlying logic*	X	X
Knowledge/Memory	Check whether respondents are likely to *not know* or have *trouble remembering information*	X	X
Sensitivity/Bias	Assess questions *for sensitive nature or wording*, and for bias	X	X
Response Categories	Assess the adequacy of *the range of responses to be recorded*	X	X
Other	Look for problems not identified in Steps 1 - 7	X	X
Cross-Cultural Considerations	Assess questions for inappropriate or ineffective cross-cultural references		X
Potential Translation Problems	Identify problematic question characteristics		X
Cross-Question	Look for cross-question problems in the entire questionnaire		X

1.2 Research and Implementation Questions

Based on our review of the QAS to date, we focused this project on the five research and implementation questions below.

1) Can the QAS be easily adopted by staff with questionnaire design experience, but no QAS experience per se, and used with junior staff with little to no questionnaire design experience? In other words, how "plug and play" is the QAS?
2) Are there any question-specific QAS rating steps that we would want to add or revise?
3) Would a questionnaire-wide overview and flow review checklist help with the overall review?
4) Would putting the QAS into an Excel spreadsheet make it more useful?
5) How reliable are the question-specific QAS coding steps?

2. Putting QAS Expansions into Practice

2.1 Assessing the QAS for Revision Potential and Making Modifications

While QAS-99 and QAS-04 (and their precursors) took the first big steps in systematizing the question review and revision processes by developing a method that could be used a) by staff with little training and experience in questionnaire design, and b) in a fast-paced environment, we noticed a few elements of both systems that could be improved upon to make this excellent tool even better.

We decided that the QAS needed to be split into two "assessments," the first focused on the questionnaire as a whole, and the second focused on individual questions similar to the original QAS. Essentially, this separated the Cross-Question review step into its own assessment process (referred to as "Assessment 1" below) to address overall questionnaire flow and other considerations that apply to all questions in the questionnaire.

Assessment 1 inspects the questionnaire in its entirety much like the Cross-Question step developed by QAS-04, however, we expanded the evaluation. We call this assessment the Questionnaire Flow Review. First, we added steps to review screener scripts and instructions, interviewer scripts and help screens, informed consent language/instructions, and programming logic. Second, question order issues were expanded to include asking about topic relevance, topic interest, timeframes, assimilation, contrast, and priming effects. Third, skip pattern problems were expanded to include how to deal with refusal and "don't know" responses, and mandatory questions. Fourth, we expanded on formatting issues by asking about piped and filled values. Last, the subcode to review question wording and response category inconsistencies was moved to the second assessment to be combined with other question-level appraisal steps and is called Cross-Question, as the step in QAS-04.

Assessment 2 is the individual question review including the original QAS-99 and QAS-04 review steps with modifications, or Question-Specific Review Steps and Cognitive-Response Process Mapping. We considered whether the review steps needed any further clarification or revision. First, we created Encoding as an explicit separate step from the Knowledge/Memory step. The Encoding step expands on the type of memory or recall failure by assessing whether the information requested is something a respondent would likely have stored in memory, which is critical in question answering. QAS-99 has a step to address Knowledge/Memory but does not overtly mention encoding. However, encoding has been included in several cognitive models of question answering (Cannel, Marquis, and Laurent, 1977; Strube, 1987; Groves, 1989; Eisenhower, Mathiowetz, and Morganstein, 1991; Lee, Brittingham, Tourangeau, et al. 1999; Tourangeau, 2018;

Tourangeau, Rips, and Rasinski, 2000; Callegaro, 2005). Second, we simplified the Cross-Cultural Considerations and Potential Translation Problems to ease use for non-experts. Third, while our review step definitions are generally quite similar to the original, instead of subcodes to break down the review category, we collapsed the components into short cohesive examples of potential error risks.

With these two assessments, we created an Excel-based tool (rather than paper form) that facilitates using the QAS as a comprehensive end-to-end questionnaire content and flow review tool.

Table 2: Assessment 1: Questionnaire Flow Review Steps and Definitions

Questionnaire Flow Issues	*Error or Inefficiency Risk*
Are screener scripts and instructions clear?	If screening and eligibility scripts are not clear, interviewers may interview ineligible households or fail to interview eligible households. This scenario produces coverage errors.
Are interviewer help screens and tailored rebuttals provided and easily accessible?	Cooperation is more likely when interviewers can tailor responses to questions asked of them and the tone of the potential respondent. The CATI questionnaire should support such tailoring.
Is informed consent language provided? Is it clear and easy to read? Is it clear to the interviewer that this must be read verbatim or that certain sections must be read verbatim?	Informed consent language that is long and not written in plain language[2] will be difficult for interviewers to read and risks hang-ups from potential respondents who would otherwise be willing to participate (Dillman, Smyth, and Christian, 2014).
Is the first question in the questionnaire topic-relevant? If so, are respondents told why less interesting and relevant questions are being asked first?	If the first few questions are difficult, or not topic-relevant, respondents may not continue the interview.
Are questions with similar timeframes and topics logically grouped together?	Timeframe grouping allows respondents to think about events over a specific timeframe, reducing one source of measurement error. Topic or experience grouping allows respondents to think about all similar topics or experiences together. If questions within a topic reference varying timeframes, those should be ordered or introduced to reduce burden and confusion that can lead to measurement error.
How are "don't know" and refusal options dealt with? Are they ever read aloud? Are any questions mandatory?	In general, respondents should be allowed to skip any questions they want to. Core questions, such as those required for within household selection are mandatory, and the interview is discontinued if the respondent will not answer. Rules need to be set for each question.
Are there any risks of assimilation, contrast, or priming effects due to question order?	When general questions come first, answers to some survey questions can be influenced by questions asked before them. This risk can be particularly high in multilingual surveys (Lee and Grant, 2009).

[2] Information about the U.S. government's plain language guidelines can be found here: https://www.plainlanguage.gov/

Table 2: Assessment 1: Questionnaire Flow Review Steps and Definitions

Questionnaire Flow Issues	*Error or Inefficiency Risk*
Which specific sections or topics are likely to be sensitive? Describe sensitive responses or groups with whom the question will be sensitive.	Traditionally sensitive topics include sexuality, drug use, alcohol use, and income and finances. Some questions are sensitive simply because of their content (e.g., income), while others may only be sensitive for respondents who provide certain answers. Confidentiality reminders can help avoid item nonresponse on these questions.
Is programming logic clearly described?	Lack of clarity in programming logic can lead to inaccurate programming and processing error or item nonresponse error when items are not asked.
Are respondents only asked questions relevant to them and is skip logic clearly specified?	Limiting questions to only those relevant to respondents by using filter questions makes the entire interview process less burdensome reducing measurement error and hang-ups due to fatigue.
Are fills or piped values used optimally?	Filling relevant text into questions (e.g., "You told me earlier that…") can reduce measurement error by making questions more specific and using information already provided during the interview.

Table 3: Assessment 2: Individual Question Review Steps and Definitions

QAS Review Step	*Potential Error Risks*
Encoding: Does the question ask about something that the respondent has noticed and committed to memory?	Respondents must have encoded information previously in order to accurately answer a survey question about it. For example, asking respondents "What was the first thing your doctor said to you at your last visit?" will be difficult to answer accurately. Asking about information that has not been encoded leads to measurement error due to guessing, item nonresponse, and potentially unit nonresponse.
Reading: Determine if it is difficult for the interviewers to read the question uniformly to all respondents.	If interviewers cannot read questions as worded, they will change wording and meaning, introducing measurement error.
Instructions: Look for problems with any introductions, instructions, or explanations from the respondent's point of view.	Comprehension problems can include: • Definitions that are read after the question instead of at the beginning • Definitions that are read to some respondents and not others, at interviewer discretion
Clarity: Identify problems related to communicating the intent or meaning of the question to the respondent.	Comprehension problems can include: • Vagueness in meaning that leaves respondents to intuit the meaning of questions in a way that is different from the intended meaning
Assumptions: Determine if there are problems with assumptions made or the underlying logic.	Comprehension problems can include: • Respondent has not had the specific experience required to answer

Table 3: Assessment 2: Individual Question Review Steps and Definitions

QAS Review Step	*Potential Error Risks*
Knowledge/Memory: Check whether respondents are likely to not know or have trouble remembering information.	Memory/retrieval problems can include: • Recalling rare events occurring in the past • Recalling specific instances of very frequent events
Sensitivity/Bias: Assess questions for sensitive nature or wording, and for bias.	Judgement problems can include: • Wording that encourages a socially desirable response • Question wording that routes respondents with rare experiences into "other" • Response options that are not sensitive to respondents' situations and disclosure risks
Response Categories: Assess the adequacy of the range of responses to be recorded.	Response selection problems can include: • Categories that are not mutually exclusive or exhaustive • Lack of ordering in naturally ordered categories • Category ranges that do not reflect actual behavior or map to reporting needs • Too many response options • Overreliance on "other, specify" options
Cross-Cultural Considerations: Assess questions for inappropriate or ineffective cross-cultural references.	Comprehension problems can include: • The use of "family" to refer only to "nuclear family" • Gender stereotyped or gender-specific behaviors and experiences
Potential Translation Problems: Identify problematic question characteristics.	Comprehension problems, can include: • Idioms and turns of phrase that do not translate from English
Cross-Question: Look for cross-question problems in the entire questionnaire.	Comprehension problems, can include: • Making sure terminology in the question is consistent with terminology used in similar questions in the questionnaire, or differences are explained clearly

2.2 Using the Excel-based QAS Tool

Our revised QAS was programmed into Excel so that question text could be referenced across worksheets easily, and rater feedback could be processed and summarized in various ways within the workbook.

2.2.1 Assessment 1: Questionnaire Flow Review

The Questionnaire Flow Review walks the reviewer through the entire questionnaire in a natural progression, starting with the survey instructions, informed consent, interviewer script prompts, and first question of the survey. The assessment addresses question grouping based on timeframe and topic, refusal responses, and question order effects. In

the last steps, topic sensitivity, programming logic, skip logic, and piped values are addressed. Assessment 1 includes a guide to facilitate use describing each category and commonly associated pitfalls.

Figure 1 displays Assessment 1 in our QAS Excel worksheet. The far right column (Column C) shows results from our sample assessment from an opioid use survey.

	A	B	C
1	**Questionnaire Flow Issues**	**Error or Inefficiency Risk**	**Assessment**
2	Are screener scripts and instructions clear?	If screening and eligibility scripts are not clear, interviewers may interview ineligible households or fail to interview eligible households. This scenario produces coverage errors.	no issues
3	Are interviewer help screens and tailored rebuttals provided and easily accessible?	Cooperation is more likely when interviewers can tailor responses to questions asked of them and the tone of the potential respondent. The CATI questionnaire should support such tailoring.	N/A - web only
4	Is informed consent language provided? Is it clear and easy to read? Is it clear to the interviewer that this must be read verbatim or that certain sections must be read verbatim?	Informed consent language that is long and not written in plain language[1] will be difficult for interviewers to read and risks hang-ups from potential respondents who would otherwise be willing to participate.	Plain language is used, but order could be altered. **Suggest dividing last bullet into two, beginning at "There is no direct benefit..." and moving the statement "If you don't want to answer a particular question, you can skip it." to either the beginning or end of bullet #3.**
5	Is the first question in the questionnaire topic-relevant? If so, are respondents told why less interesting and relevant questions are being asked first?	If the first few questions are difficult, or not topic-relevant, respondents may not continue the interview.	First questions are not topic-relevant but they are easy to answer and there is language used to introduce the section that identifies it as being non-topic relevant. Intro does not explicitly state why the first questions are not topic-relevant however.
6	Are questions with similar timeframes and topics logically grouped together?	Timeframe grouping allows respondents to think about events over a specific timeframe, reducing one source of measurement error. Topic or experience grouping allows respondents to think about all similar topics or experiences together. If questions within a topic reference varying timeframes, those should be ordered or introduced to reduce burden and confusion that can lead to measurement error.	Questions are grouped by topic. Cigarettes section is ordered by timeframe. **Opioid use section should be re-ordered by timeframe (i.e., have 10340 come after 10350).** Campaign recall sections are ordered by timeframe.
7	How are "don't know" and refusal options dealt with? Are they ever read aloud? Are any questions mandatory?	In general, respondents should be allowed to skip any questions they want to. Core questions, such as those required for within household selection are mandatory, and the interview is discontinued if the respondent will not answer. Rules need to be set for each question.	**nearly all questions are mandatory**

Figure 1: Assessment 1: Questionnaire Flow Review in Excel – 1st tab

2.2.2 Assessment 2: Question-Specific Review Steps and Cognitive-Response Process Mapping

Assessment 2 is comprised of QAS-99 and QAS-04 steps for individual question review, including Reading, Instructions, Clarity, Assumptions, Knowledge/Memory, Sensitivity/Bias, Response Categories, Cross-Cultural Considerations, Potential Translation Issues, and Cross-Question problems. In addition to these steps, we added Encoding as a new step and placed it at the beginning of the review steps because it is the first cognitive pre-requisite for any fact or behavior questions. Respondents cannot be expected to answer questions about information they never encoded.

After reading instructions for Assessment 2 in the second tab, the reviewer completed the individual question review in the third tab, documenting any issues found. This structure makes it very easy to share comments with your team.

Figure 2a displays Assessment 2 in our QAS Excel worksheet. Column A contains descriptions of each review step while column B lists the associated potential error risks.

	A	B
1	**QAS Review Step**	**Potential Error Risks**
2	*Encoding* : Does the question ask about something that the respondent has noticed and committed to memory?	Respondents must have encoded information previously in order to accurately answer a survey question about it. For example, asking respondents "What was the first thing your doctor said to you at your last visit?" will be difficult to answer accurately. Asking about information that has not been encoded leads to measurement error due to guessing, item nonresponse, and potentially unit
3	*Reading:* Determine if it is difficult for the interviewers to read the question uniformly to all respondents.	If interviewers cannot read questions as worded, they will change wording and meaning, introducing measurement error.
4	*Instructions:* Look for problems with any introductions, instructions, or explanations from the respondent's point of view.	Comprehension problems can include: • Definitions that are read after the question instead of at the beginning • Definitions that are read to some respondents and not others, at interviewer discretion
5	*Clarity:* Identify problems related to communicating the intent or meaning of the question to the respondent.	Comprehension problems can include: • Vagueness in meaning that leaves respondents to intuit the meaning of questions in a way that is different from the intended meaning
6	*Assumptions:* Determine if there are problems with assumptions made or the underlying logic.	Comprehension problems can include: • Respondent has not had the specific experience required to answer
7	*Knowledge/Memory:* Check whether respondents are likely to not know or have trouble remembering information.	Memory/retrieval problems can include: • Recalling rare events occurring in the past • Recalling specific instances of very frequent events
8	*Sensitivity/Bias:* Assess questions for sensitive nature or wording, and for bias.	Judgement problems can include: • Wording that encourages a socially desirable response • Question wording that routes respondents with rare experiences into "other" • Response options that are not sensitive to respondents' situations and disclosure risks

Figure 2a: Assessment 2: Individual question review in Excel including instructions worksheet (a) and review worksheet (b)

Figure 2b displays the worksheet with review steps in each column and question text/response options in each row.

	A	B	C	D	E	F
1		*Encoding* : Does the question ask about something that the respondent has noticed and committed to memory?	*Reading:* Determine if it is difficult for the interviewers to read the question uniformly to all respondents.	*Instructions:* Look for problems with any introductions, instructions, or explanations from the respondent's point of view.	*Clarity:* Identify problems related to communicating the intent or meaning of the question to the respondent.	*Assumptions:* Determine if there are problems with assumptions made or the underlying logic.
2	*Question 1* [text] [response options]					
3	*Question 2* [text] [response options]					
4	*Question 3* [text] [response options]					
5	*Question 4* [text] [response options]					
6	*Question 5* [text] [response options]					

Figure 2b: Assessment 2: Individual question review worksheet in Excel

2.3 Revised QAS Implementation

Our modified QAS was implemented on a web survey of opioid awareness and attitudes that we conducted for a health nonprofit organization. All questions in the questionnaire received expert review, and 10 were selected for review in the QAS (see Table 4 for the original and revised questions).

The questionnaire team consisted of a junior and senior survey researcher led by a senior methodologist. The junior researcher had about 2 years of survey experience and this was her first questionnaire development experience. The senior researcher had about 7 years

of survey experience and had worked on questionnaire development for several surveys, including question writing and cognitive testing. The junior and senior researchers served as QAS raters in this pilot. The senior survey methodologist had about 20 years of experience, largely in questionnaire design and testing. Training on the revised QAS was minimal and no test ratings were conducted to assess accuracy of codes. Thus, this research should be interpreted as exploratory, but also reflects how the tool can be used in a fast-paced survey development environment.

First, the senior methodologist and senior researcher reviewed the entire questionnaire and picked 10 questions for review in the QAS. After reviewing the questionnaire, they jointly completed the Assessment 1: Questionnaire Flow Review sheet. Second, the junior and senior researchers rated each question using the Assessment 2: Question-Specific Review Steps and Cognitive-Response Process Mapping worksheet. Upon completion, the senior methodologist reviewed the raters' results of Assessment 2, and revisions were made to the original questions.

Table 4: Questions (n = 10) and Response Options Selected from Draft Questionnaire for Review and Revised after QAS Review

Question and Response Options	*Revised Question and Revised Options*
Student (Q10145) Are you…? 1) Currently a college student that lives on or near a college campus during the school year 2) Currently a college student who does not live on or near a college campus 3) Not currently a college student	Even if you also work, are you currently a college student or in any kind of technical training program after high school? 1) Yes 2) No
Opioids: Self use (Q10350) On how many occasions (if any) in your lifetime, have you taken prescription opioids without a doctor specifically prescribing them to you? 1) 0 occasions 2) 1 occasion 3) 2 occasions 4) 3-5 occasions 5) More than 5 occasions	Have you ever taken prescription opioids without a medical professional prescribing them to you? Remember that your answers will be kept private. By prescription opioids, we mean any opioid/narcotic drug that may be prescribed by a medical professional to relieve pain. Some examples: Oxycodone, Hydrocodone, Acetaminophen/aspirin with codeine, Morphine, Vicodin, Oxycontin, Percocet, Fentanyl, Hydromorphone, Methadone, Buprenorphine, Oxymorphone. These are sometimes also called lean, percs, oxy and other nicknames. 1) Yes 2) No

Table 4: Questions (n = 10) and Response Options Selected from Draft Questionnaire for Review and Revised after QAS Review

Question and Response Options	*Revised Question and Revised Options*
Opioids: Other use (Q10360) In the past 6 months, have any of the following people close to you used prescription opioids without a doctor specifically prescribing them? Select all that apply. 1) A family member 2) A friend/ peer/ acquaintance 3) A significant other 4) None of these	To the best of your knowledge, in the past 6 months, has a family member who lives with you used prescription opioids without a medical professional prescribing them? (split into multiple question for relationship) 1) Yes 2) No
Attitude: Risk Harm (Q10400) How much do you think people risk harming themselves (physically or in other ways) if they try prescription opioids once or twice without a doctor telling them to? 1) No risk 2) Slight risk 3) Moderate risk 4) Great risk	The next few questions ask about how harmful or unharmful you think it is to take opioids. For this survey there are no right or wrong answers. We are interested in what you know without looking anything up online or asking anyone about it. When thinking about "harm", please think about physical, social, or emotional ways that people can harm themselves by taking opioids. For example, physical harm could include addiction, side effects, or death. Social harm could include losing friends or family members. Emotional harm could include depression or other psychological problems. How much do you think people risk harming themselves if they try prescription opioids once or twice without a medical professional telling them to or in a different way from what the medical professional prescribed? 1) No risk 2) Low risk 3) Moderate risk 4) High risk 5) Very high risk
Attitude: Take a Stand (Q10500, 17) How much do you agree with the following? Taking a stand against prescription opioids is important to me. 1) Do not agree 2) Somewhat agree 3) Strongly agree 4) Very strongly agree	Taking a stand against misuse of prescription opioids is important to me. 1) Strongly Disagree 2) Disagree 3) Neither Agree/Disagree 4) Agree 5) Strongly Agree

Table 4: Questions (n = 10) and Response Options Selected from Draft Questionnaire for Review and Revised after QAS Review

Question and Response Options	*Revised Question and Revised Options*
Media Exposure (Q10750) In a typical week, what percentage of the time do you watch TV or video using video services that…? Must sum to 100% [SHOW TOTAL MUST SUM TO 100%] 1) Have advertising (such as Cable TV, Fiber Optic TV [FIOS], Hulu, Satellite TV, YouTube) 2) Do not have advertising (such as Amazon Instant Video, Hulu Plus, Netflix, HBO GO)	In the past 30 days where have you seen, heard, or read any messages about the opioid epidemic? a) Television you watch through cable or antenna 1) Yes 2) No b) Video streaming services such as Hulu, Youtube and others 1) Yes 2) No c) Websites 1) Yes 2) No
Adult Mentor (Q12330) If you had a serious problem, is there an adult family member you could talk to? 1) Yes 2) No	If you had a serious problem, is there an adult family member or friend, such as a mentor, coach, or teacher, that you could talk to? 1) Yes 2) No
Negative Well-being (Q12370) For each item, please indicate how often the following are true for you during the past 6 months. Please answer all items as well as you can even if some do not seem to apply to you. a) You lie or cheat. b) You have trouble sleeping. c) You are unhappy, sad, or depressed. d) You have trouble concentrating or paying attention. e) You don't get along with other people your age. 1) Never True 2) Rarely True 3) Sometimes True 4) Often True	How often have the following been true for you during the past 6 months? Please answer all the questions even if you think some don't apply to you. a) I have trouble sleeping. b) I am unhappy, sad, or depressed. c) I have trouble concentrating or paying attention. 1) Never True 2) Rarely True 3) Sometimes True 4) Often True 5) Always True

Table 4: Questions (n = 10) and Response Options Selected from Draft Questionnaire for Review and Revised after QAS Review

Question and Response Options	*Revised Question and Revised Options*
Home Location (Q13200) Now we have a few more general questions. Do you live…? 1) In an urban or city area 2) In a suburban area next to a city 3) In a small town or rural area 4) Not sure	Do you live…? 1) In an urban or city area 2) In a suburban area next to a city 3) In a small town or rural area
Income (Q13230) Considering your own income and the income from any other people who help you, how would you describe [your family's/your] overall financial situation? Would you say you…? 1) Don't meet basic expenses 2) Just meet basic expenses with nothing left over 3) Meet needs with a little left over 4) Live comfortably	How would you describe your household's overall financial situation? Would you say you…? 1) Can't pay for basic expenses like rent, water, electricity, clothes, and food 2) Can pay for basic expenses with nothing left over 3) Can pay for basic expenses with a little left over to save or spend on yourself 4) Can pay for basic expenses with more than a little left over

3. Implementation Results and Interrater Agreement

3.1 Qualitative Assessment of Implementation and Questionnaire Flow Review Checklist

The questionnaire flow checklist that we incorporated as the first part (Assessment 1) of our QAS proved to be a concise way to organize review and feedback on overarching questionnaire issues. Rather than relying on the "black box" expertise of the senior survey methodologist, that expertise was communicated through the checklist in such a way that the senior rater could fill it out, the senior methodologist could review and modify it, and the team could recommend changes to the client. More specifically, the checklist helped us recommend things such as revised bullets in the consent script to clarify informed consent points, recommending new opening questions because the first few questions in the draft questionnaire were not topic-relevant, to reorder the opioid section by timeframe to match ordering in the tobacco use section, and to consider whether and which questions should be mandatory versus optional. In addition to identifying changes, the checklist helped us confirm that key components of the questionnaire were already present (e.g., that screening criteria were included, that there were no obvious assimilation, contrast, or priming effect risks, that the tone was good for a survey with many sensitive questions, and that clear and appropriate skip logic was provided). Finally, we were able to identify that at least in the initial draft, there were no planned or hidden fills or piping, which can become challenging if discovered later in the specification and programming process.

3.2 Agreement Between Raters with More and Less Experience

We calculated interrater reliability as simple agreement between the junior and senior raters' assessments of each question reviewed. We used the following assessments of interrater reliability.

1) Overall agreement about whether there was an issue present across all 10 questions rated
2) Agreement on at least one specific problem type on each question
3) Correlation between each rater on the number of issues identified on each question
4) Agreement rate between each rater on each QAS step across all questions

Table 5 is a crosstabulation of whether the junior and senior raters agreed on whether there was an issue present on each of the 10 questions. There was high agreement that at least one issue was present (0.9 agreement rate), meaning that they agreed that there was at least one problem present on nine of the 10 questions. On the one question where there was disagreement, the senior rater found potential problems with Instructions, Knowledge/Memory, and Cross-Cultural Considerations that the junior rater did not. The problems identified were helpful for revising the question. This question asked about the percentage of time that the respondent watched TV, specifically streaming video that contains ads, and similar services that do not contain ads (see Table 5 below). First, the frame of reference for the percentage was unclear (i.e., the question asked "percentage of time," but the response options were supposed to add up to 100%, so a clearer wording would have asked "percentage of the time you watch TV or videos." Second, this seems like a difficult quantity for a respondent to estimate and report, assuming they understood the percentage correctly and encoded the information in the first place. The definitions were located in the response options, meaning the respondent may come to an answer without actually reviewing the response options. Further, the response options were not completely accurate (e.g., YouTube, like Hulu, has levels of service contain ads and those that do not). In line with Willis and Lessler's "wide net" goal for the QAS, we think it is better that the senior rater found these problems, than that neither rater found them. In other words, were the QAS was not reliable, a more experienced reviewer would likely find problems anyway.

Table 5: Cross-Tabulation Between Senior and Junior Raters' Assessment of a Problem Present on Each of 10 Questions

		Agreement	
		Junior Rater	
		Issue not present	*Issue present*
Senior Rater	*Issue not present*	0	0
	Issue present	1	9

We also assessed whether the two raters identified the same types of problems on each question (no table shown). This measured whether the raters agreed on at least one specific cognitive issue. For example, if both reviewers identified Clarity and Knowledge/Memory problems, regardless of other problems that either reviewer identified, they were considered to be "in agreement" for the purposes of this calculation. For the college student status question, the senior rater identified Clarity and Assumptions, while the junior rater identified Response Categories, Cross-Cultural

Considerations, and Cross-Question. For that question, the two raters were not in agreement because neither reviewer identified the same issue as the other. Comparatively, on the question asking about friends and family who have used unprescribed prescription opioids, the senior rater identified Encoding, Instructions, ***Clarity, Knowledge/Memory, Response Categories,*** and Cross-Question, while the junior rater identified ***Clarity, Knowledge/Memory, Response Categories***, and Cross-Cultural Considerations (overlapping problems in ***bold italics***). Because one or more cognitive problems overlapped between the two raters, we called this agreement. We observed a 0.5 agreement rate in at least one specific issue across questions (i.e., raters were in agreement on at least one specific cognitive issue on five of the 10 questions, basically chance.) While the junior and senior raters did not agree on any specific problems, the fact that they both identified problems aided in question revisions.

With categories that overlap, inter-rater reliability may not be very high, but this tool nonetheless meets the goal of catching issues and organizing rater comments, facilitating team collaboration and documentation for the client.

Table 6: Junior and Senior Raters' Assessments of Student Status

Senior Rater		*Junior Rater*		
Clarity	*Assumptions*	*Response Categories*	*Cross-Cultural Considerations*	*Cross-Question*
Unclear if the goal is to determine student status in general or living arrangement for full-time student or perhaps to differentiate online versus on-campus students. Need to define "college student," either through written definition or intro to explain the purpose of the question	See column E response (Clarity)	Should an answer for technical/trade school/training be included? May or may not consider this "college" (4 yr vs.2 yr) depending on the certification or degree, and relevant to lower SES. Many certifications and vocational training also take place in "colleges." What is the goal here?	See comment in response categories, definition of "college" per SES	Depending on definition of college intended, completing high school may not be necessary but is required in base for this question. Definitions should be consistent with Q140

Next, we looked at whether the number of problems identified on each question was correlated between raters. Figure 3 shows the number of problems identified by the junior and senior rater by question. Each data point represents one of the 10 questions rated. The y-axis represents the junior rater's assessment of the number of problems on that question, and the x-axis represents the senior rater's assessment of the number of problems on that question. For example, the junior rater identified four problems on the "Adult mentor" question, while the senior rater only identified one. The scatterplot shows no clear correlation between the two raters. They identified the same, or nearly the same number of problems on some questions (e.g., "Income [3,3], "Student" [2,3]). Interestingly "Opioids: Other use" and "Opioids: Self use" both received [6,4], and three questions ("Attitudes: Take a stand", "Home location", and "Negative well-being"

received a [2,1]. The only clear pattern is that the senior rater tended to find more problems than the junior rater, with the exceptions being "Adult mentor" [4,1] and "Student" [3,2]).

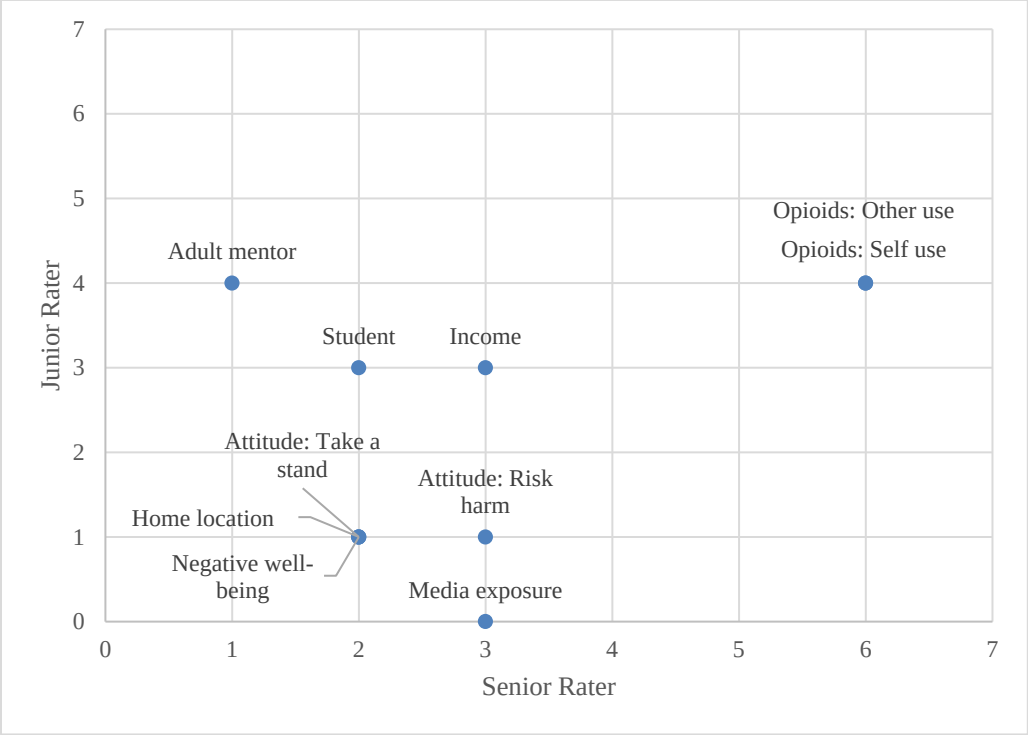

Figure 3: Number of question problems identified by a junior and senior rater, by question

Finally, we assessed how reliable each of the QAS rating steps was by reviewing the agreement rate on each step across questions (see Table 7). Reading and Translation issues had perfect agreement (i.e., whenever one rater noticed one of those issues, the other rater did too). We observed very high agreement (i.e., 0.7 to 0.9) on Assumptions, Encoding, Knowledge/Memory, Sensitivity/Bias, and Clarity. Agreement was less than chance on Response Categories, Cross-Question, Instructions, and Cross-Cultural Considerations. Less agreement suggests the raters were using different understandings of the review steps or seeing different issues in the questions.

Table 7: Agreement Between Junior and Senior Rater by QAS Step Across 10 Questions

QAS Step	*Agreement*
Reading	1.0
Potential Translation Problems	1.0
Assumptions	0.9
Encoding	0.8
Knowledge/Memory	0.8
Sensitivity/Bias	0.8
Clarity	0.7
Response Categories	0.5
Cross-Question	0.5

Table 7: Agreement Between Junior and Senior Rater by QAS Step Across 10 Questions

QAS Step	*Agreement*
Instructions	0.4
Cross-Cultural Considerations	0.4

4. Results Summary, Recommendations, and Future Directions

4.1 Results Summary

First, our modified QAS can be used by staff with little training or experience in questionnaire design, and in a fast-paced questionnaire design environment. This is not surprising given the QAS's origins, but it is reassuring. Certainly, the QAS takes more time than a quick expert review, but not so much time that it should be overlooked on that criteria alone. The QAS can help shift expensive expert reviewer time to less expensive junior or mid-level staff with expert oversight. Further, the QAS makes a good training tool for junior questionnaire design staff. Readers should keep in mind that we did not review all questions in the questionnaire with the QAS steps. Rather, we identified 10 questions that we thought were the most problematic and needed closer review. We also only used the QAS review steps, but not the detailed subcodes. For example, raters only needed to identify whether there were problems with the instructions (QAS Step 2) and not whether those problems were "conflicting or inaccurate instructions" (2a) versus "complicated instructions" (2b). Both of these decisions meant that raters could work very quickly, providing their reviews in less than one day.

Second, adding a separate Encoding step helps isolate this type of problem in a clearer way. While this type of question problem might otherwise be captured under Step 5 Knowledge/Memory in past QAS protocols, neither QAS-99 nor QAS-04 actually instruct the user to look for encoding problems specifically. Given the frequency at which we see this type of problem in practice, and the potential for the QAS to be used as a question review training tool, we felt it was important to include encoding as its own review step. Two of the 10 questions reviewed had encoding problems, and this step exhibited relatively high agreement between raters. Although, it is worth noting that raters agreed that eight of 10 questions did not have an encoding problem, but did not agree on either of the two questions that the senior rater assessed as having an encoding problem. Encoding problems tend to require a unique approach to revisions, usually involving discussions with lead researchers or clients to isolate what similar experience or behavior that respondents *do have* encoded would meet their measurement goals. Thus, we think this will continue to be a very useful new QAS step in future implementations.

Third, the questionnaire flow checklist (Assessment 1) proves to be a useful tool to provide a quick overview of potential problems within the questionnaire and missing components. We found it helpful to us, operationally speaking, to have this checklist outside of the question-specific review steps (instead of within the Cross-Question original QAS-04 step), in terms of reviewer completion, for providing feedback to the client, and for developing specifications for our programmers. We found it helpful to keep the Cross-Question review separate as well, as the two elements address different things, and our Assessment 1 addressed issues that were not captured under the Cross-Question step.

The Excel-based tool proved to be essential for quick and organized data entry and analysis. While Lessler and Forsyth (1996) mention electronic versions of the QAS, the QAS-99 manual (i.e., Willis and Lessler, 1999; the most detailed manual we could find online) appears to be developed to be used as a paper-and-pencil form, or in a Word document (i.e., it looks like a paper form, and asks the user to create one sheet per question). By implementing the QAS in Excel, we could create a database that lets us view all the reviewed questions at one time, and that could be more easily incorporated into other questionnaire databases, inventories, review systems, programing specifications, and code books. We have seen similar Excel-based questionnaire development tools like this in other contexts and would be surprised if we were the first to put the QAS into Excel but have not seen an example of such an implementation.

Finally, our reliability assessment, which is the first of its kind to our knowledge, showed that there was high reliability about the presence of problems overall, but not necessarily on the type or number of problems. While a reliability study was not the original or primary goal of this project, it is a useful side product, and provides some insight into the codes. Our senior rater tended to find more problems than our junior rater, which is to be expected because she had several more years of experience in writing and testing questionnaires. However, we were encouraged that the two raters were in agreement on the presence of any problem and found the same or nearly the same number of problems on many questions. Yet, agreement on specific problem categories across items was low (0.5), and there was a wide range in agreement across review categories. Interestingly some rating categories that could be thought of as highly subjective (e.g., Sensitivity/Bias) had high agreement, while other categories that seem like they would be easy to identify (e.g., the appropriateness of Response Categories or helpfulness and clarity of Instructions) had low agreement. Thus, despite not being a primary goal of this original study, the analysis provides some insight into QAS steps that may be easier and harder to assess without extensive training. Further research should be conducted on QAS reliability, including the possibility of establishing reliability standards as is commonly done for other coding schemes. No effort was taken in this study to make sure that raters were reliable on QAS codes before conducting their reviews.

4.2 Reflections and Recommendations for Future Use and Evaluation of the QAS

Willis and Lessler (1999) describe the QAS as a system of "overlapping fishing nets" (p. 3-2) to catch errors. High reliability is a nice feature to have, but not an essential feature. The goal of the QAS is to find errors (and potential fixes), regardless of what category they are perceived to fall into. Even with low reliability, this approach is more replicable and transparent, and less error prone than expert review alone. No doubt, expert input will still be an essential part of any QAS, particularly if inter-rater reliability is low. Even if higher levels of reliability can be established, it is probably a good idea to involve questionnaire design experts in any QAS implementation. While the QAS is good at identifying problems, it does not necessarily tell the user how to fix any specific problem on a specific question.

Our QAS innovation is just the beginning of our improvements to this tool. First, we envision developing systematic rules for a process for selecting which questions to evaluate. The QAS was developed to assess individual questions. However, using the QAS in the context of full questionnaire reviews provides a different challenge, specifically, how many and which questions should be reviewed. Our question triage process was relatively ad hoc, but the QAS could conceivably include a "quick" review of all questions (i.e., are there any problems of any type, or a gut feeling of "something

wrong"), followed by a "detailed" review that uses the problem-specific coding steps. Such a revision would potentially make the QAS easier to use in a fast-paced questionnaire development environment in which an entire questionnaire needs to be reviewed.

Second, following QAS-99 origins, we plan to apply the question-specific QAS review step to our Behavioral Risk Factor Surveillance System (BRFSS) surveys, and other surveys that do not have time for cognitive testing or more in-depth question-specific testing before a field pilot. We plan to use the questionnaire flow checklist as standard operating procedure on all new surveys.

Third, we think the QAS would benefit from dedicated, rigorous reliability studies, similar to those conducted for other rating systems. To our knowledge, these have never been conducted with the QAS prior to the humble attempt presented here. Such studies could a) help refine code definitions and user instructions, and b) provide reliability training standards for raters learning to use the system.

Fourth, and admittedly the most complex ambition of these future directions involves integrating the QAS into other questionnaire development and documentation tools, such as question inventories, programming specs, codebooks, or metadata systems that capture question and response option wording. Every questionnaire designer wishes they had such an integrated and comprehensive questionnaire system or database, but examples are few and far between. Efficiencies can be gained in the overall questionnaire development and documentation process by connecting even two of these things, and we encourage readers to make attempts to do so.

In summary, we found that even a quick and abbreviated implementation of the QAS question-specific review steps made question review easier to delegate across the questionnaire design team, and made findings easier to communicate to the client due to the structured format. Despite the breadth of earlier QAS versions, implementing only the major tasks may be sufficient in some contexts, and combined with an overarching full questionnaire review checklist, the QAS is a very useful and efficient tool that we encourage others to try.

Acknowledgements

We sincerely thank Gordon Willis and Liz Dean for answering a few key QAS history questions for us when developing our 2020 AAPOR presentation. Any errors or misrepresentations are ours.

References

Beatty, P. C., Collins, D., Kaye, L., Padilla, J.-L., Willis, G. B., & Wilmot, A. (Eds.). (2019). *Advances in Questionnaire Design, Development, Evaluation and Testing* (1st Edition). Wiley.

Callegaro, M. (2005, 18-22 July). Origins and developments of the cognitive models of answering questions in survey research. Paper presented at the First annual meeting of the European Association for Survey Research (EASR), Barcelona.

Cannel, C. F., Marquis, K. H., & Laurent, A. (1977). A summary of studies of interviewing methodology. Vital and Health Statistics, Series 2(69), i-68.

Czaja, R. (1998). Questionnaire Pretesting Comes of Age. Marketing Bulletin, 9, 52–66.

Dean, E., Caspar, R., McAvinchey, G., Reed, L., & Quiroz, R. (2005). Developing a low-cost technique for parallel cross-cultural instrument development: the Question Appraisal System (QAS-04). In J. H. P. Hoffmeyer-Zlotnik, & J. Harkness (Eds.), Methodological aspects in cross-national research (pp. 31-46). Mannheim: GESIS-ZUMA.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, phone, mail, and mixed-mode surveys: The Tailored Design Method (4th ed.). Hoboken, NJ: Wiley.

Eisenhower, D., Mathiowetz, N., & Morganstein, D. (1991). Recall error: Sources and bias reduction techniques. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. Mathiowetz & S. Sudman (Eds.), Measurement errors in surveys (pp. 127-144). New York: Wiley.

Forsyth, B. H., & Hubbard, M. (1992). A method for identifying cognitive properties of survey items. In American Statistical Association (Ed.), Proceedings of the section on Survey Research Methods (pp. 470-475). Washington DC: American Statistical Association.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: John Wiley and Sons.

Lee, L., Brittingham, A., Tourangeau, R., Willis, G., Ching, P., Jobe, J., & Black, S. (1999). Are reporting errors due to encoding limitations or retrieval failure? Surveys of child vaccination as a case study. Applied Cognitive Psychology, 13(1), 43–63. https://doi.org/10.1002/(SICI)1099-0720(199902)13:1<43::AID-ACP543>3.0.CO;2-A

Lee, S., & Grant, D. (2009). The effect of question order on self-rated general health status in a multilingual survey context. American Journal of Epidemiology, 169(12), 1525–1530.

Lessler, J. T., & Forsyth, B. H. (1996). A Coding System for Appraising Questionnaires. In N. Schwarz & S. Sudman (Eds.), Answering questions: Methodology for determining cognitive and communicative processes in survey research (1st ed, pp. 259–291). Jossey-Bass Publishers.

Strube, G. (1987). Answering survey questions: The role of memory. In H.-J. Hippler, N. Schwarz & S. Sudman (Eds.), Social information processing and survey methodology (pp. 86-101). New York: Springer-Verlag.

Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. Quality Assurance in Education, 26(2), 169–181. https://doi.org/10.1108/QAE-06-2017-0034

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The Psychology of Survey Response (1st edition). Cambridge University Press.

Willis, Gordon & Lessler, Judith. (1999). Question Appraisal System QAS-99.