Research in Social and Administrative Pharmacy xxx (xxxx) xxx

Contents lists available at ScienceDirect

Research in Social and Administrative Pharmacy

journal homepage: www.elsevier.com/locate/rsap

Towards a reconsideration of the use of agree-disagree questions in measuring subjective evaluations

Jennifer Dykema [a,b,*], Nora Cate Schaeffer [a,b], Dana Garbarski [c], Nadia Assad [a], Steven Blixt [d]

[a] University of Wisconsin Survey Center, University of Wisconsin, Madison, USA
[b] Department of Sociology, University of Wisconsin, Madison, USA
[c] Department of Sociology, Loyola University, Chicago, USA
[d] Bank of America, USA

A B S T R A C T

Agree-disagree (AD) or Likert questions (e.g., "I am extremely satisfied: strongly agree … strongly disagree") are among the most frequently used response formats to measure attitudes and opinions in the social and medical sciences. This review and research synthesis focuses on the measurement properties and potential limitations of AD questions. The research leads us to advocate for an alternative questioning strategy in which items are written to directly ask about their underlying response dimensions using response categories tailored to match the response dimension, which we refer to as item-specific (IS) (e.g., "How satisfied are you: not at all … extremely"). In this review we: 1) synthesize past research comparing data quality for AD and IS questions; 2) present conceptual models of and review research supporting respondents' cognitive processing of AD and IS questions; and 3) provide an overview of question characteristics that frequently differ between AD and IS questions and may affect respondents' cognitive processing and data quality. Although experimental studies directly comparing AD and IS questions yield some mixed results, more studies find IS questions are associated with desirable data quality outcomes (e.g., validity and reliability) and AD questions are associated with undesirable outcomes (e.g., acquiescence, response effects, etc.). Based on available research, models of cognitive processing, and a review of question characteristics, we recommended IS questions over AD questions for most purposes. For researchers considering the use of previously administered AD questions and instruments, issues surrounding the challenges of translating questions from AD to IS response formats are discussed.

Introduction

Credited to Rensis Likert in his seminal research on attitude measurement, agree-disagree (AD) or Likert questions are among the most frequently used response formats to assess attitudes and opinions, appearing in numerous studies and many national and federal surveys.[1–3] As illustrated by the following question, AD questions present respondents with statements and ask them to rate their level of agreement: *Medical researchers work extremely hard to make sure they keep information from participants private and secure. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?*[4]

While researchers have written about the positive psychometric properties of AD questions,[5] the ubiquity of these items is also likely due to their ease of use. Scales comprised of AD questions are practically appealing because the same response categories can be used for each statement regardless of the content or complexity of the statements, and for self-administered questionnaires, researchers can format multiple AD questions economically in a grid.[6,7]

These positive features, however, may be offset by increased burden for respondents, which may reduce data quality, and has led questionnaire designers to advocate for item-specific (IS) questions.[6–9] IS questions are written to directly ask about a question's underlying response dimension with response categories tailored to match the response dimension.[6,7,9] For example, an IS version of the example question would be written to measure the underlying response dimension of how hard medical researchers work using response categories that assess the intensity of working hard: *How hard do medical researchers work to make sure they keep information from participants private and secure: not at all hard, a little hard, somewhat hard, very hard, or extremely hard?*

In the following sections we: 1) review experimental studies comparing data quality for AD and IS questions; 2) present conceptual models of and review research concerning respondents' cognitive processing of AD and IS questions; 3) provide an overview of question characteristics that frequently differ between AD and IS questions and may affect respondents' cognitive processing and data quality; and 4) offer concluding comments and recommendations regarding the use and study of AD and IS questions.

* Corresponding author. University of Wisconsin Survey Center (UWSC), 4308 Sterling Hall 475 N. Charter St., Madison, WI, 53706-1582, USA.
E-mail address: dykema@ssc.wisc.edu (J. Dykema).

https://doi.org/10.1016/j.sapharm.2021.06.014
Received 21 November 2020; Received in revised form 6 May 2021; Accepted 20 June 2021
Available online 24 June 2021

Effects of AD versus IS questions on data quality

We identified 20 experimental studies that directly compare AD and IS questions and evaluate differences based on data quality or cognitive processing outcomes. Several studies examine the desirable data quality indicators of validity and reliability. Overall, a larger number of studies find IS questions are associated with higher validity and reliability. For example, while six studies reported no consistent difference between AD and IS questions,[3,4,10–13] three studies demonstrated validity was higher for IS questions,[8,14,15] and no studies reported higher validity for AD questions. For reliability, five studies demonstrated higher reliability for IS questions,[8,11,12,15,16] two for AD questions,[4,13] and two studies reported no difference.[3,17]

Studies have also examined undesirable data quality indicators including acquiescence (tendency to agree with a question regardless of its content),[18] response effects due to primacy (systematic selection of the first category), recency (systematic selection of the last category), and extreme responding (systematic selection of the first and last categories), straightlining (tendency to give similar answers to items in a battery of questions),[19] item nonresponse, and speeding and break-offs in online surveys. In general, more studies find AD questions are associated with these negative outcomes, but a number of studies find no differences, and a few studies find higher levels of undesirable outcomes for IS questions. For example, while four studies reported no or inconsistent differences between AD and IS questions for acquiescence,[13,16,20,80] four studies reported AD questions were more susceptible to acquiescence.[10,11,14,17] Findings for other response effects and straightlining are more mixed. Three studies uncovered primacy,[21] extreme responding,[22] and scale direction[23] effects for AD questions; one study reported recency effects[4] for IS questions; and a final study reported extreme responding was present for both AD and IS formats.[2] For straightlining, two studies reported more straightlining in AD scales,[10,12] one in IS scales,[22] and two studies reported no differences.[21,23] While three studies reported no consistent pattern in item-missing responses for AD and IS questions,[16,21,22] one study reported higher levels for IS questions.[4] Finally, while three studies reported higher levels of speeding among questions with AD formats,[21–23] neither an AD or IS format was more likely to affect the likelihood of break-offs in online surveys.[22,23]

Cognitive processing of AD and IS questions

Questionnaire designers argue that AD questions are more likely to lower data quality because they are more cognitively burdensome than IS questions.[6–8,24] A characteristic that contributes to the complexity of AD questions is that they often present respondents with a mismatch between the question's "offered" and "underlying" response dimensions. A response dimension is the continuum a question asks the respondent to consider when constructing their answer.[6,9,25] For questions about evaluations and judgments using rating scales, response dimensions can establish *valence* (whether the evaluation of a target object is positive or negative; e.g., "agree or disagree"), *intensity* (degree to which the evaluation is held; e.g., "not at all … extremely"), *quantity* (amount of the evaluation held; e.g., "none … a great deal"), or *relative frequency* of the target object (e.g., "never … always"). Consider the AD question in Table 1.[4] The offered response dimension presented by the response categories is the intensity of agreement. This conflicts with the underlying response dimension of the intensity of working hard presented in the statement. These mismatches force respondents to undertake complicated cognitive processing steps in order to "map" their naturally occurring responses to the statement onto the AD response categories.

Tourangeau et al.[26] describe four stages through which respondents construct answers to survey questions: comprehension, retrieval of relevant information from memory, use of retrieved information to make judgments, and selection and reporting of an answer. Others have expanded on this model, adding cognitive steps involved in responding to AD questions specifically,[6,8,23,27,28] and in Table 1, we present conceptual models of the cognitive processing steps undertaken to answer AD and IS questions.

Conceptual model of cognitive processing steps for AD questions

The first step is Comprehension in which the respondent must comprehend the literal meaning of the statement (e.g., "Medical researchers work extremely hard to make sure they keep information from participants private and secure") as well as its component parts (e.g., "medical researchers," "work [extremely] hard," etc.). Next, during Identification, the respondent identifies the question's underlying response dimension, which is accomplished by understanding the meaning of the statement as well as attending to threshold words, if included. Threshold words are intensifiers (e.g., "very"), quantifiers (e.g., "most"), or frequency markers (e.g., "rarely") often included in AD statements that establish a threshold on the underlying response dimension without presenting the full range of scale options. For example, the AD question includes the threshold word "extremely," which, by modifying "work hard," serves to reinforce the intensity of working hard as the underlying response dimension. After identifying the underlying response dimension, the respondent must generate their own internal value (response) on the dimension (Generation). For the current question, the respondent generates an internal value of "pretty hard." Ensuing steps encompass a set of complicated cognitive processes in which the respondent evaluates the distance between their internal value of "pretty hard" and the threshold value of "extremely hard" (Threshold evaluation), and then determines whether the distance between their internal value and the threshold value indicates "agreement," "disagreement," or "neutrality" (Polarity evaluation). Finally, guided by their evaluation of polarity, the respondent must map their internal value onto the offered response dimension using one of the offered categories (Mapping). For example, the respondent might select "agree" because their internal value "pretty hard" is close to the threshold value "extremely hard," or the respondent could select "disagree" because "pretty hard" is less intense than "extremely hard."

Conceptual model of cognitive processing steps for IS questions

The cognitive processing steps undertaken to answer a comparable IS question are simplified and predicted to be less burdensome. First, the respondent must comprehend the literal meaning of the question and its component parts (Comprehension). During Identification, the respondent determines the underlying response dimension, which is reinforced by the manner of questioning and the labeling and ordering of the response categories (e.g., "not at all hard," "a little hard," etc.). Next, the respondent generates an internal value of "pretty hard" (Generation), but placement of this value is done directly by mapping it to one of the offered categories (Mapping), thereby circumventing Threshold and Polarity evaluation. For the current question, the respondent could select "somewhat hard" or "very hard" because "pretty" lies between "somewhat" and "very" based on studies that scale adverbial phrases and intensifiers.[29,30]

Respondents' cognitive effort when processing AD and IS questions alone and in batteries

Studies examining respondents' cognitive effort processing AD and IS questions indicate two question characteristics moderate effort: whether questions appear alone or as part of a battery; and the extent to which IS response categories vary across questions.[23,28,31] While the model in Table 1 anticipates that a single AD question presented in isolation will require a higher level of cognitive processing, most AD questions appear in batteries in which the statements vary but the response categories remain constant. This presentation allows respondents to memorize the pattern of questioning and categories and

Table 1
Conceptual model of the cognitive processing steps undertaken by respondents to answer an AD versus IS question.

Question Characteristics					Cognitive Processing Steps					
Response format	Question wording	Response dimension		Threshold word	Comprehension	Identification	Generation	Threshold evaluation	Polarity evaluation	Mapping
		Offered	Underlying							
AD	Medical researchers work extremely hard to make sure they keep information from participants private and secure. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	intensity of agreement	intensity of working hard (i.e., "how hard medical researchers work")	extremely (hard)	comprehend literal meaning of statement and pragmatic goal of the task	indirectly determine underlying response dimension, which is reinforced by the threshold value ("extremely hard")	generate internal value ("pretty hard")	evaluate distance between internal value ("pretty hard") and threshold value ("extremely hard")	evaluate whether distance between internal value and threshold value indicates "agreement" "disagreement," or "neutrality"	map internal value onto the offered response dimension using one of the discrete categories offered in the "agreement" or "disagreement" range or select midpoint; e.g., select "agree" because "pretty hard" is close in value to "extremely hard" or "disagree" because "pretty hard" is less than "extremely hard"
IS	How hard do medical researchers work to make sure they keep information from participants private and secure: not at all hard, a little hard, somewhat hard, very hard, or extremely hard?	intensity of working hard (i.e., "how hard medical researchers work")	same as offered response dimension	(not applicable)	comprehend literal meaning of question and pragmatic goal of the task	directly determine underlying response dimension based on manner of questioning and response categories	generate internal value ("pretty hard")	(not applicable)	(not applicable)	map internal value onto one of the discrete categories; e.g., select "somewhat hard" or "very hard" because "pretty" lies between "somewhat" and "very"

Note: The "threshold evaluation" step is conditional upon the AD statement containing a threshold word.

may encourage a less thoughtful process of answering.[32] By contrast, when multiple IS questions are grouped together, they (often, but not always) use different response dimensions and response categories, requiring respondents exert more effort to process the variable response categories.[23]

Research examining variation in the time respondents spend processing and answering questions largely support these propositions. Response latencies (RLs) measure time spanning the end of an interviewer's question reading to the respondent's answer.[33] Researchers timed RLs for questions about trust[4] and political efficacy[11] in which categories varied across IS items, but were invariant across AD items. In both studies, RLs for the *first* question in the battery were significantly (or marginally so) longer for the AD item, providing some evidence that AD response formats imposed a more cognitively burdensome response task. Evaluated as a group, RLs were longer for the IS questions about trust, but not political efficacy.

Researchers have also examined response times (RTs; total time spent reading and answering) for questions presented as stand-alone items in which response categories were the same for AD items but varied for IS items.[21–23] Findings indicated RTs were longer for IS questions, regardless of the number or ordering of categories or whether the questions were answered on PCs or smartphones. By contrast, there were no differences in RTs for AD and IS questions presented in grids in which the response categories were held constant for both the AD and IS questions.[23] Taken together, studies of RTs indicate the changing nature of IS categories may increase the amount of cognitive effort respondents expend.

Other methodologies also provide evidence that the varying response categories of grouped IS questions require more cognitive effort while the repeated questioning pattern of grouped AD questions encourage more superficial processing. In an interviewer-administered study, researchers[4] reported that IS questions were associated with higher levels of behavioral indicators of response difficulty (e.g., higher levels of uncodable answers and answers with qualifications) because the IS response categories were harder for participants to remember, an issue exacerbated by the number of questions (11 questions were asked without show cards) and aural presentation of items. Using eye-tracking technology, researchers[28,34] examined cognitive effort by recording respondents' eye movements separately for question stems versus response categories, which were the same for the AD items, but varied for the IS questions. While findings indicated no differences in eye movements for the question stems, respondents processed IS response categories more intensively, viewing them more and for longer times.

Results from studies examining respondents' cognitive effort answering AD and IS questions suggest more research is needed to understand factors that lead to increased effort for IS questions and most importantly, whether that effort is associated with data quality. Response times alone can be difficult to interpret: "delays in responding could mean that a question is difficult to process (usually a bad sign) or that the question encourages thoughtful responding (typically a good sign) (p. 297)."[7] While longer times have been associated with less accurate answers,[35] an experimental study with a self-administered instrument suggested the relationship between time and accuracy may be curvilinear with longer and shorter times being less accurate.[36]

Overview of question characteristics that differ between AD and IS questions

In experiments, the AD-IS question pairs being evaluated often vary on multiple question characteristics that can affect cognitive processing and data quality. For example, both the offered response dimension (intensity for the AD question and frequency for the IS question) and direction of the response categories (high to low agreement versus low to high frequency) vary for following AD-IS pair: "Doctors rarely keep the whole truth from their patients: agree strongly … disagree strongly" and "Doctors keep the whole truth from their patients: never … always."[8] Some characteristics, such as response dimensions, often co-vary in studies comparing AD and IS questions in ways that are not controlled experimentally, making it impossible to isolate unique or moderating effects of the characteristics. Other characteristics, such as the number and use of verbal labels for response categories, are usually held constant within an AD-IS experiment; but these features vary across studies, complicating the task of generalizing findings. We compiled questions included in AD-IS experiments and systematically coded their features to identify key characteristics that differ between AD-IS question pairs (summarized in Table 2). We describe how these characteristics vary within and across AD-IS experiments, and for select characteristics, we briefly summarize findings regarding data quality.[9,25]

Manner of questioning

Questioning manner – whether the sentence with the content to be evaluated is structured as a statement or question – is fundamental to the nature of what distinguishes AD and IS items and always differs across AD-IS comparisons. Researchers cite the indirect question structure of AD items as a reason to avoid them,[6] and findings from experimental studies support these recommendations. While subjects in eye-tracking studies appeared to exert equivalent cognitive effort processing AD and IS question stems,[28,34] subjects in a laboratory setting processed the content of items less deeply when they were written as assertions versus interrogatives.[37]

Acquiescence

Research indicates the offered response dimension of agreement may cause AD questions to be more vulnerable to acquiescence, particularly among respondents with lower levels of education,[18,38,39] whereas IS response dimensions make this much less of a concern. Acquiescence for AD questions could arise because listeners have a pre-disposition to "agree" unless they have a reason to disagree, perhaps due to politeness, deference, or because of conversational practices.[18,40] Such tendencies might be exacerbated if AD statements are complex or part of a large group of items that are repetitious or not salient to the respondent. In addition, the "agree" or positive end of the response dimension is usually offered first,[18] and may receive more processing or be perceived more favorably, and thus more likely to be selected.[31]

Threshold words

Threshold words, the selection of which is typically arbitrary,[8] may complicate respondents' efforts to map internal values onto AD response categories, and ultimately lead to answers that violate the principle of monotonic equivalence.[7] An item possesses monotonic equivalence when increasing (or decreasing) values for the answers correlate with increasing (or decreasing) values on the underlying scale of the construct being measured. For example, consider the statement "non-adherence is mostly due to people being careless," designed to measure patients' reasons for medication non-adherence.[41] The underlying response dimension implied by the statement is *how much* non-adherence is due to carelessness. However, one respondent could answer "disagree" because they believe non-adherence is "not at all" due to carelessness while another could "disagree" because they feel non-adherence is due to carelessness "a great deal." While both respondents report a value of "disagree," the first respondent's internal value of "not at all" is clearly much lower on the underlying response dimension than "a great deal." An IS version of this item provides a direct method of asking this question and more readily ensures that respondents order themselves accurately on the response continuum: "How much is non-adherence due to people being careless: not at all, a little, somewhat, quite a bit, or a great deal?"

Because measurement requires monotonic equivalence, some argue

Table 2
Comparison of differences in key question characteristics for agree-disagree (AD) and item-specific (IS) questions.

Question Characteristics		Operationalization within and across studies	
Category	Description	AD	IS
Manner of questioning	Whether the sentence to be evaluated is structured as declarative (a statement) or interrogative (a question)	Indirect, structured as a statement	Direct, structured as a question
Response dimensions	Continuum that a question asks respondents to consider when constructing their answer	Offered response dimension (intensity of agreement) and underlying response dimension (intensity, quantity, or frequency) do not match	Offered and underlying response dimensions match and measure intensity, quantity, or frequency
Threshold words	Intensifiers (e.g., "very"), quantifiers (e.g., "most"), or frequency markers (e.g., "rarely") that establish a threshold for comparison	Often, but not always, included in the statement	Not applicable
Polarity	Whether response dimension is bipolar with both poles or ends of the response dimension presented or unipolar with only one pole presented	Usually bipolar ("strongly agree … strongly disagree")	Usually unipolar ("not at all satisfied … extremely satisfied"), but may be bipolar ("extremely satisfied … extremely dissatisfied")
Number of response categories	Number of categories or points offered on the response continuum	Category number in AD-IS experiments is almost always held constant between AD-IS comparisons within a study; across studies, category number varies from 4 to 11, with 5 categories being the most common implementation	
Labeling of response categories	Labeling of all or only some of the categories using various combinations of words and numbers	Category labeling in AD-IS experiments is almost always held constant between AD-IS comparisons within a study; across studies, labeling varies, with categories fully labeled with words and no numbers being the most common implementation	
Direction of response categories	Whether the categories increase in value (e.g., "strongly disagree … strongly agree," "not at all" … "extremely") or decrease in value (e.g., "strongly agree … strongly disagree," "extremely" … "not at all")	Varies, but categories often decrease in value (e.g., "strongly agree … strongly disagree")	Varies, but categories often increase in value particularly for unipolar quantity (e.g., "none … a great deal") and frequency (e.g., "never … always") dimensions
Middle category	For bipolar questions, whether the response categories include a conceptual middle where the dimension transitions from positive to negative; category may be neutral category or mixed	Commonly used bipolar questions often include a middle category (e.g., "neither agree nor disagree")	If unipolar, no conceptual middle category
Battery	Whether questions appear alone or as part of a battery of topically-related items with a common response format	Use same response categories for all items included in a battery	Response categories for items in the battery will likely vary depending on the underlying response dimension
Valence of the construct and target objects	Whether the construct and target objects asked about in the questions are inherently positive, negative, neutral, or ambiguous	Valence of the construct and objects in AD-IS experiments is almost always held constant between AD-IS comparisons within a study; across studies, valence varies	
Alignment with the construct	Whether the construct and the response categories are positively aligned (e.g., higher-valued response categories indicate higher levels of the construct) or negatively aligned (e.g., higher-valued response categories indicate lower levels of the construct)	Alignment of a construct for an AD-IS experimental pair sometimes varies; across studies alignment varies	

that responses to AD questions are only interpretable if they include threshold values at either end of the response continuum.[42] For some response dimensions, such as frequency, extreme values may be obvious (e.g., "never" or "always"). For other response dimensions, such as quantity using "how much," it is not absolutely clear what the extreme positive value should be. Is "a great deal" the highest positive value on a "how much" scale? Further, the literature is replete with examples of instruments using AD questions that fail to include a threshold value at all, allowing respondents to superimpose their own interpretations.

Polarity

AD items are almost always bipolar and present both poles or ends of the response dimension (e.g., "agree strongly … disagree strongly"). While IS items can be bipolar (e.g., "extremely dissatisfied … extremely satisfied"), they are usually unipolar, presenting only one possible pole (e.g., "not at all satisfied … extremely satisfied" or "not at all satisfied … extremely dissatisfied"). Whenever the underlying response dimension for an AD question is quantity or frequency, the corresponding IS question will always be unipolar because quantities do not contain values less than "none" or "not at all" and frequencies do not possess values lower than "never." Only intensity response dimensions can be bipolar and there are some dimensions (e.g., "important") where it is unclear whether the negative polar-value (e.g., "unimportant") is equivalent to the positive polar-value.

In an analysis of measurement error for items from the General Social Survey (GSS), which included a number of AD questions, results indicated unipolar questions were more reliable than bipolar questions.[43] Differences in polarity alone are also likely to generate differences in the marginal distributions,[44] which limit the maximum correlations among the items. IS items offer the possibility of using a variety of positive and negative response dimensions as recommended by some;[45,46] and the items may have lower correlated method variance than AD items.

Compared to bipolar AD items, unipolar IS items also offer more points of differentiation on a particular side of the response dimension and may increase variation for scale scores.[12]

Response categories

Response categories differ in terms of their number, labeling, and direction. While the number and labeling of categories within a study is almost always held constant between AD-IS pairs, these characteristics vary considerably across studies. By contrast, category direction – whether the categories increase or decrease in value – sometimes varies across AD-IS pairs in the same study. In AD-IS experiments, categories for AD questions more often decrease in value (e.g., "agree … disagree"), while categories for IS questions more often increase (e.g., "never … always"). Some research indicates data quality for both AD and IS items may be optimized using five categories, fully labeled with words, and presented in increasing order.[9,22,47] In other research, respondents had difficulty distinguishing between "strongly disagree" and "disagree."[17] "Strongly" may be problematic as a modifier because it potentially conflates the extremity of a respondent's evaluation with their certainty.[48]

Middle category

In contrast to unipolar IS items, AD questions often include a clear conceptual middle category (e.g., "neither agree nor disagree"). While experiments evaluating data quality for the inclusion of middle categories for bipolar questions have had mixed results,[7,49–51] studies indicate respondents use the middle category when answering AD questions in unwanted ways. For example, when probed, respondents overwhelmingly reported selecting the middle category because they did not have an opinion on the issue.[52,53] Research indicates respondents may use the AD's middle "neither agree nor disagree" category to indicate uncertainty or deal with a lack of knowledge and express ambivalence.[4,54,55] From a measurement perspective, respondents use of the "neither/nor" middle category is problematic: while respondents may reliably select this option, their response is not a valid measure of the construct being assessed. Researchers have noted problems with the interpretation of an AD middle category and often suggest analyzing responses using this category separately and not as a middle value.[5]

Battery

As described in the section on cognitive processing, when AD questions appear in batteries their presentation as variable statements with repeated response categories allows respondents to memorize the questioning pattern and response categories.[32] By contrast, when multiple IS questions are grouped together, they (often, but not always) use different response dimensions and response categories. Placement in a battery, the number of questions contained in the battery, and the extent to which the response categories vary across questions for IS questions are likely to impact respondents' cognitive effort and affect data quality. In interviewer-administered instruments, items in batteries are associated with lower reliability.[56] When multiple questions are presented in a grid in self-administered instruments, they may be answered more quickly, more vulnerable to straightlining,[19] and more highly correlated.[57] Higher correlations in a grid presentation may signal higher measurement error due to shared error variance.[9]

Valence and alignment

In order to measure constructs validly and reliably, researchers use multi-item scales that combine respondents' answers to create a single value.[58] Relationships among a construct's valence, the valence of the objects to be evaluated in the questions, and the alignment between the construct and questions gives rise to a complicated set of relationships with implications for measurement error.

Valence refers to the inherently positive, negative, neutral, or ambiguous nature of the construct and the objects asked about in the questions. For example, a construct like trust is inherently more positively valenced, while a construct like racial resentment is more negatively valenced. Valence also varies across questions within a scale. For a scale measuring political efficacy,[2] a question asking "(how much) public officials care about what people think" is positively valenced, while a question about "(how often) politics and governments seem so complicated people can't really tell what's going on" is negatively valenced.

Alignment refers to whether lower- or higher-valued response categories indicate lower or higher values of the construct. Positively aligned items are those for which a higher-valued category (e.g., "strongly agree" for an AD question and "a great deal" for an IS question) indicate higher levels of the construct being measured and negatively aligned items are those for which a higher-valued category indicates lower levels of the construct. For example, the question about public officials caring what people think would be positively aligned because the highest-valued categories ("strongly agree" and "a great deal") indicate the highest level of political efficacy. By contrast, the question about politics and governments would be negatively aligned because the highest-valued categories indicate the lowest level of political efficacy.

For AD questions, a question's valence can lead to undesired response effects due to acquiescence. For positively valenced constructs and questions, acquiescence can make responses and constructs appear more positive than they are in reality; for positively valenced constructs and negatively worded questions, acquiescence can make responses and constructs appear more negative. For more negatively valenced constructs like depression, a tendency to agree with items that are aligned to indicate higher values for the construct (e.g., "I have felt sad and blue"), can lead to overestimates of the construct. Depending on how items are scored, acquiescence can inflate estimates of mean scores, artificially inflate or deflate reliability estimates (particularly for items worded in the same direction), and create spuriously high correlations between AD measures and criterion measures.[59–61]

In order to reduce effects due to acquiescence (and inattention), researchers often recommend creating scales that include both (and often an equal number of) positively and negatively aligned items[62–64] (also called "item reversals"[64] and reverse-worded questions[65]). The logic behind this approach is that it will reduce bias in scale means by placing those who acquiesce in the middle of the response distribution. However, research indicates several problems with this approach. First, writing negatively worded questions that convey the same meaning across all respondents can be difficult (e.g., to measure the opposite of "interesting," a researcher could use "not interesting," "uninteresting," or "boring," but it is unlikely these have the same meaning across respondents and including oppositely worded items will only reduce bias if respondents answer those items as extremely as they would their counterparts[66]). Second, the use of negations like "not," "un-," "non-," and "-less" may decrease comprehensibility and data quality.[67,68] This may be particularly problematic for AD items where the inclusion of a negation in the statement (e.g., "My gender does not affect the way others treat me") requires processing a double negative in order to reject the statement's contents (e.g., by selecting "disagree").[69,70] Third, attempts at balancing scales may create methodological problems including lowering the validity and internal consistency of the measures and adding a method effect by creating an unexpected factor structure for the negatively-aligned items.[71–74]

Concluding comments, recommendations, and future directions

Limitations of experimental studies comparing AD and IS questions

Overall, more studies find IS questions are associated with desirable data quality outcomes (validity, reliability) and AD questions are

associated with undesirable outcomes (acquiescence, response effects, etc.). A number of studies, however, find no differences between the question types, and a few studies find higher levels of undesirable outcomes for IS questions. Several limitations of these comparative studies may account for inconsistent or null findings. First, the number of experimental studies comparing AD and IS questions is relatively small. Our review identified twenty studies. Second, highlighted in our discussion of question characteristics, AD-IS question pairs often vary across a number of characteristics that are usually not controlled for, which may confound the results. Third, studies explore a limited number of topics and the effects of AD and IS questions may vary by topic.

Fourth, studies examine many different data quality outcomes: validity, reliability, acquiescence, straightlining, etc. These outcomes vary in terms of their strength and operationalizations. While estimates of validity and reliability potentially offer more direct measures of data quality, studies evaluate different measures of reliability and validity that vary in their quality. For example, estimates of reliability of items in a scale, such as from Cronbach's alpha, include correlated error variance and do not provide values for individual items. Estimated test-retest reliabilities, over the short intervals that are commonly used, may be too compromised by memory or reliable method effects to provide a strong criterion.[56] It is plausible that a combination of acquiescence, the repetition of the response categories, and the presentation of items in a battery increases correlated method variance among a set of AD items, a reminder that simple correlations are fundamentally an ambiguous indicator of data quality. Because method variance is central to evaluating the relative quality of AD and IS items, methods for estimating reliability and construct validity that can identify method variance are needed.[14]

What the overview of question characteristics tells us

Our analysis of the key question characteristics that vary between AD-IS questions included in AD-IS experiments highlights the fact that in these experiments, the questions being compared often vary on a number of characteristics, complicating our ability to draw conclusions. In one study,[4] researchers noted their AD-IS pairs measuring trust varied based on: offered response dimensions (the AD questions measured intensity while the response dimensions for the IS questions were item-specific by design and measured intensity, frequency, and quantity); the direction of the response categories (the AD response categories were ordered from high to low – "strongly agree" to "strongly disagree" – while the IS categories were ordered from low to high – "not at all" to "a great deal," "never" to "always"); and polarity (the AD questions were bipolar; the IS questions were unipolar). The structural differences between these two response formats have important consequences for respondents' cognitive processing and data quality. To date, no studies feature a design that allows for estimation of all the unique or joint effects of these characteristics. Indeed, only a handful of experiments cross the use of an AD-IS response format with systematic variation in other characteristics that are likely to be important for data quality, such as the number of response categories or scale direction. Findings from such studies may ultimately uncover systematic interactions between AD-IS response formats and other question characteristics.

Challenges of translating AD questions to IS questions

When writing questions to measure subjective evaluations for a new study, the issues presented here recommend using IS questions. Many studies, however, aim to use items from previously administered questionnaires and translating from an AD to IS format can pose a number of challenges. Because AD statements are relatively easy to write, they often include several elements – such as multiple target objects and conditional statements – to be evaluated simultaneously.[42] Consider, the following AD question from the GSS: "Because of past discrimination, employers should make special efforts to hire and promote qualified women." This question asks about several things: beliefs about the causes (e.g., gender) and agents of discrimination (e.g., employers), the responsibility of employers to make amends for past discrimination, and whether hiring and promoting qualified women rectifies past behavior. Agreement or disagreement with this statement could be based on beliefs about any of these components or combinations of them. Translating this question into an IS format underscores the complexity of the item and decisions that must be made about the underlying response dimension: is the question asking about intensity (*how special* efforts should be), quantity (*how much* effort should be made), or frequency (*how often* efforts should be made)?

A related problem with AD questions that likely contributes to their lower data quality is that they are often written in way that leaves their underlying response dimension ambiguous or open to multiple interpretations.[6] Consider the AD question in Table 3, taken from the GSS and included in a scale designed to measure political efficacy. While the threshold word "most" implies a quantity response dimension, the AD statement can easily be translated into IS questions using intensity, quantity, or frequency dimensions, and indeed, two possible quantity dimensions – "how much" and "how many" are possible.

AD questions are widely used because many items can be combined into a battery using the same response categories, even if the items ask about completely different topics. For self-administration, AD questions can be formatted in a grid to minimize space. However, because IS questions use response categories that match the questions' underlying response dimensions, translating a set of items from AD to IS often reveals that the items do not share the same underlying response dimension. For example, while the six AD items in Table 4 use the same response categories, compactly formatted in the grid,[11] their IS counterparts use response dimensions for intensity, quantity, and frequency and require response categories relevant for those dimensions. When combined, the IS items result in a slightly longer grid. While a visually longer grid may be perceived by respondents as more burdensome, because they are more clearly written and easier to understand, the IS questions are likely less burdensome. More research measuring respondents' cognitive effort while answering AD and IS questions and directly linking effort measures to data quality is needed.

Question writers often need to balance revision against replication.[69] Given the wide-spread use of AD questions, researchers may need to weigh disadvantages of not using previously administered questions or "validated" AD scales, including losing trends from time-series data, versus potential gains in data quality to converting IS measures. While

Table 3
Illustrative translation of an AD question on political efficacy into multiple IS questions with variable response dimensions.

Response format	Response dimension	Question wording
AD	Intensity	Most government administrators can be trusted to do what is best for the country. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?
IS	Intensity	How well can we trust government administrators to do what is best for the country: not at all, a little, somewhat, very, or extremely?
IS	Quantity	How much can we trust government administrators to do what is best for the country: not at all, a little, some, quite a bit, or a great deal?
IS	Quantity	How many government administrators can be trusted to do what is best for the country: none, a few, some, many, or all?
IS	Quantity	How many government administrators can be trusted to do what is best for the country: none, less than half, about half, more than half, or all?
IS	Frequency	How often can we trust government administrators to do what is best for the country: never, rarely, sometimes, very often, or always?

Table 4
Translation and presentation of an AD scale into an IS scale for self-administration on paper.

AD. **Next, we have a few questions about people's general views on politics and government. Please tell me how strongly you agree or disagree with each of the following statements.**						
		Strongly agree	**Agree**	**Neither agree nor disagree**	**Disagree**	**Strongly disagree**
a.	I feel that I have a pretty good understanding of the important political issues facing our country.	○	○	○	○	○
b.	I think most people are better informed about politics and government than I am.	○	○	○	○	○
c.	The average citizen has considerable influence on politics.	○	○	○	○	○
d.	People like me don't have any say about what the government does.	○	○	○	○	○
e.	People we elect to Congress try to keep the promises they have made during the election.	○	○	○	○	○
f.	Most government administrators can be trusted to do what is best for the country.	○	○	○	○	○
IS.Next, we have a few questions about people's general views on politics and government.						
		Not at all	**A little**	**Somewhat**	**Very**	**Extremely**
a.	How good is your understanding of the important political issues facing our country?	○	○	○	○	○
b.	Compared to most people, how informed are you about politics?	○	○	○	○	○
		None	**A little**	**Some**	**Quite a bit**	**A great deal**
c.	How much influence does the average citizen have on politics?	○	○	○	○	○
d.	How much say do people like you have about what the government does?	○	○	○	○	○
		Never	**Rarely**	**Sometimes**	**Very often**	**Extremely often**
e.	How often do the people we elect to Congress try to keep the promises they have made during the election	○	○	○	○	○
f.	How often can we trust government administrators to do what is best for the country?	○	○	○	○	○

many issues related to developing a validated instrument[75,76] are beyond the scope of this review, we remind readers that instrument validation is not a binary outcome, but a process.[77] An instrument validated for a specific population for a specific purpose would not – without evidence – extend to a different population or purpose. Further, many "validated" instruments use questions that fall short of evidenced-based standards for writing questions for standardized measurement.[9]

Future research

Although experimental studies directly comparing AD and IS response formats yield some mixed results, given the strong theoretical underpinning and available evidence in support of the IS format, we recommend IS questions over AD questions for most purposes. Our review also points to the need for more experimental research comparing AD and IS questions across a range of substantive topics and with designs that incorporate strong criteria to evaluate data quality. Future work should prioritize the following: 1) Are some constructs or questions with specific characteristics better measured with AD questions? Dykema et al.[4] noted that when asking about a non-salient construct like trust in medical researchers, questions using frequency-based response dimensions, especially when asking about externally-focused actors (e.g., "how hard do medical researchers work to ensure participants in their studies are safe"), were difficult for respondents because they sounded like they were asking respondents about their knowledge of the target object and not for an evaluation.[78] Similar to the statements Likert used in his early work, an agreement response dimension may also be easy to apply to statements of values using "should" (e.g., "Adult children should take care of their parents when the parents become old"). 2) What combinations of characteristics yield the best data outcomes? We encourage future work using multifactorial designs that can provide researchers with the ability to estimate the effects of particular question characteristics and combinations of characteristics in order to determine which combinations yield the highest quality data. 3) To what extent do the measurement properties of AD and IS questions vary across groups based on socio-demographic characteristics such as education, language spoken, and age? Many studies demonstrate that unwanted response effects like acquiescence are higher among respondents with lower education,[38,39] but few studies examine whether an AD or IS format is more likely to protect against such effects. 4) How do AD and IS response formats interact with the mode of administration, which format is optimal for which modes, and which features of implementation within mode have consequences for measurement? A limitation of interviewer-administration is that respondents must encode and recall response categories. While providing showcards for IS items during in-person interviews may reduce respondents' cognitive burden, this solution is not easily applicable to phone interviews and IS scales that include many items with variable response categories may be difficult for respondents. Further, an increasing share of surveys are completed on mobile devices which usually use a responsive design that limits horizontal scrolling by replacing grids with stand-alone questions, rendering any advantages of grids null. Issues related to mode are likely to receive increased scrutiny as surveys that mix modes grow and researchers continue to explore methods to measure and reduce mode effects.[79,80] Although recommendations may change when more and stronger research becomes available, the strongest evidence we currently have suggests that IS items will yield higher quality data and offer researchers considerable flexibility in design.

Declaration of competing interest

None.

Acknowledgements

Support for this work was provided by the University of Wisconsin Survey Center (UWSC) at the University of Wisconsin-Madison, which is supported by the College of Letters & Science. Additional support for this research was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison with funding from the Wisconsin Alumni Research Foundation to Jennifer Dykema, and the facilities of the Social Science Computing Cooperative and the Center for Demography and Ecology (NICHD core grant P2C HD047873). Opinions expressed here are those of the authors and do not necessarily reflect those of the sponsors or related organizations.

References

1. Likert R. A technique for the measurement of attitudes. *Arch Psychol*. 1932;22:5–55.
2. Liu M, Lee S, Conrad FG. Comparing extreme response styles between agree-disagree and item-specific scales. *Publ Opin Q*. 2015;79:952–975.
3. Lelkes Y, Weiss R. Much ado about acquiescence: the relative validity and reliability of construct-specific and agree–disagree questions. *Res Polit*. 2015:1–8.
4. Dykema J, Garbarski D, Wall IF, Edwards DF. Measuring trust in medical researchers: adding insights from cognitive interviews to examine agree-disagree and construct-specific survey questions. *J Off Stat*. 2019;35:353–386.
5. Willits FK, Theodori GL, Luloff AE. Another look at Likert scales. *J Rural Soc Sci*. 2016;31:126–139.
6. Fowler J, Floyd J, Cosenza C. Design and evaluation of survey questions. In: Bickman L, Rog DJ, eds. *The Sage Handbook of Applied Social Research Methods*. Thousand Oaks, CA: Sage; 2009:375–412.
7. Krosnick JA, Presser S. Question and questionnaire design. In: Marsden PV, Wright JD, eds. *Handbook of Survey Research*. second ed. Bingley, UK: Emerald Group Publishing Limited; 2010:263–313.
8. Saris WE, Revilla M, Krosnick JA, Shaeffer EM. Comparing questions with agree/disagree response options to questions with item-specific response options. *Surv Res Methods*. 2010;4:61–79.
9. Schaeffer NC, Dykema J. Advances in the science of asking questions. *Annu Rev Sociol*. 2020;46:37–60.
10. Cibelli KL, Callegaro M. Assessing the Measurement Quality of Agree/Disagree Items versus Item-specific Answer Scales. In: *Paper presented at the annual meeting of the Midwest Association for Public Opinion Research, Chicago, IL*. 2011.
11. Dykema J, Schaeffer NC, Garbarski D. Effects of Agree-Disagree versus Construct-specific Items on Reliability, Validity, and Interviewer-Respondent Interaction. In: *Paper presented at the annual meeting of the American Association for Public Opinion Research Conference, Orlando, FL*. 2012.
12. McIntyre J, Gehlbach H. *The Cost of Agree-Disagree: Satisficing and Sacrificing Reliability*. Society for Research on Educational Effectiveness; 2014.
13. Wilson DC, Davis DW, Dykema J, Schaeffer NC. Response scales and the measurement of racial attitudes: agree-disagree versus item specific formats. In: *Paper presented at the annual meeting of the American Political Science Association, Chicago, IL*. 2013.
14. Kuru O, Pasek J. Improving social media measurement in surveys: avoiding acquiescence bias in Facebook research. *Comput Hum Behav*. 2016;57:82–92.
15. Revilla M, Ochoa C. Quality of different scales in an online survey in Mexico and Colombia. *J Polit Lat Am*. 2015;7:157–177.
16. Hanson T. Comparing agreement and item-specific response scales: results from an experiment. *Soc Res Pract*. 2015;1:17–25.
17. Gehlbach H. Agree-Disagree: A "Strongly Disagreeable" Response Scale. In: *Paper presented at the annual meeting of the American Educational Research Association, New York, NY*. 2008.
18. Schuman H, Presser S. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Orlando, FL: Academic Press; 1981.
19. Kim Y, Dykema J, Stevenson J, Black P, Moberg DP. Straightlining: overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Soc Sci Comput Rev*. 2019;37:214–233.
20. Lewis JR. Comparison of item formats: agreement vs. item-specific endpoints. *J Usability Stud*. 2018;14:48–60.
21. Höhne JK, Revilla M, Lenzner T. Comparing the performance of agree/disagree and item-specific questions across PCs and smartphones. *Methodology*. 2018;14:109–118.
22. Kunz T. Evaluation of Agree-Disagree versus Construct-specific Scales in a Multi-Device Web Survey. In: *Paper presented at the 19th General Online Research Conference, Berlin, Germany*. 2017.
23. Höhne JK, Schlosser S, Krebs D. Investigating cognitive effort and response quality of question formats in web surveys using paradata. *Field Methods*. 2017;29:365–382.
24. Schaeffer NC, Dykema J. Questions for surveys: current trends and future directions. *Publ Opin Q*. 2011;75:909–961.
25. Dykema J, Schaeffer NC, Garbarski D, Hout M. The role of question characteristics in designing and evaluating survey questions. In: Beatty P, Collins D, Kaye L, Padilla J, Willis G, Wilmot A, eds. *Advances in Questionnaire Design, Development, Evaluation, and Testing*. Hoboken, NJ: Wiley; 2020:449–470.
26. Tourangeau R, Rips LJ, Rasinski K. *The Psychology of Survey Response*. New York, NY: Cambridge University Press; 2000.
27. Carpenter PA, Just MA. Sentence comprehension: a psycholinguistic processing model of verification. *Psychol Rev*. 1975;82:45–73.
28. Höhne JK, Lenzner T. New insights on the cognitive processing of agree/disagree and item-specific questions. *J Surv Stat Methodol*. 2018;6:401–417.
29. Cliff N. Adverbs as multipliers. *Psychol Rev*. 1959;66:27–44.
30. Dobson KS, Mothersill KJ. Equidistant categorical labels for construction of Likert-type scales. *Percept Mot Skills*. 1979;49:575–580.
31. Höhne JK, Krebs D. Scale direction effects in agree/disagree and item-specific questions: a comparison of question formats. *Int J Soc Res Methodol*. 2018;21:91–103.
32. Höhne JK, Lenzner T. Investigating response order effects in web surveys using eye tracking. *Psihologija*. 2015;48:361–377.
33. Dykema J, Garbarski D, Schaeffer NC, Anadon I, Edwards DF. Correlates of differences in interactional patterns among black and white respondents. In: Brenner PS, ed. *Understanding Survey Methodology: Sociological Theory and Applications*. Cham: Springer; 2020:277–304.
34. Höhne JK. Eye-tracking methodology: exploring the processing of question formats in web surveys. *Int J Soc Res Methodol*. 2019;22:199–206.
35. Schaeffer NC, Dykema J. Response 1 to Fowler's chapter: coding the behavior of interviewers and respondents to evaluate survey questions. In: Madans J, Miller K, Maitland A, Willis G, eds. *Question Evaluation Methods: Contributing to the Science of Data Quality*. Hoboken, NJ: John Wiley & Sons, Inc.; 2011:23–39.
36. Ehlen P, Schober MF, Conrad FG. Modeling speech disfluency to predict conceptual misalignment in speech survey interfaces. *Discourse Process*. 2007;44:245–265.
37. Petty RE, Rennier GA, Cacioppo JT. Assertion versus interrogation format in opinion surveys: questions enhance thoughtful responding. *Publ Opin Q*. 1987;51:481–494.
38. Narayan S, Krosnick JA. Education moderates some response effects in attitude measurement. *Publ Opin Q*. 1996;60:58–88.
39. Warnecke RB, Johnson TP, Chávez N, et al. Improving question wording in surveys of culturally diverse populations. *Ann Epidemiol*. 1997;7:334–342.
40. Schaeffer NC. Conversation with a purpose–or conversation? Interaction in the standardized interview. In: Biemer PP, Groves RM, Lyberg LE, Mathiowetz NA, Sudman S, eds. *Measurement Errors in Surveys*. New York: John Wiley & Sons; 1991:367–392.
41. Witry MJ. Medication adherence beliefs of U.S. community pharmacists. *Res Social Adm Pharm*. 2017;14:471–478. https://doi.org/10.1016/j.sapharm.2017.06.006.
42. Fowler J, Floyd J. *Improving Survey Questions: Design and Evaluation*. Thousand Oaks, CA: Sage; 1995.
43. Alwin DF, Baumgartner EM, Beattie BA. Number of response categories and reliability in attitude measurement. *J Surv Stat Methodol*. 2018;6:212–239.
44. O'Muircheartaigh C, Gaskell G, Wright DB. Weighing anchors: verbal and numeric labels for response scales. *J Off Stat*. 1995;11:295–307.
45. Solomon S. Measuring dispositional and situational attributions. *Pers Soc Psychol Bull*. 1978;4:589–594.
46. Mazaheri M, Theuns P. Structural equation modeling (SEM) for satisfaction and dissatisfaction ratings; multiple group invariance analysis across scales with different response format. *Soc Indicat Res*. 2009;90:203–221.
47. Revilla MA, Saris WE, Krosnick JA. Choosing the number of categories in agree–disagree scales. *Socio Methods Res*. 2014;43:73–97.
48. Converse JM, Presser S. *Survey Questions: Handcrafting the Standardized Questionnaire*. Thousand Oaks, CA: Sage; 1986.
49. Saris WE, Gallhofer I. Estimation of the effects of measurement characteristics on the quality of survey questions. *Surv Res Methods*. 2007;1:29–43.
50. Wang R, Krosnick JA. Middle alternatives and measurement validity: a recommendation for survey researchers. *Int J Soc Res Methodol*. 2019:1–16.
51. Weijters B, Cabooter E, Schillewaert N. The effect of rating scale format on response styles: the number of response categories and response category labels. *Int J Res Market*. 2010;27:236–247.
52. Sturgis P, Roberts C, Smith P. Middle alternatives revisited: how the neither/nor response acts as a way of saying "I don't know"? *Socio Methods Res*. 2014;43:15–38.
53. Blasius J, Thiessen V. The use of neutral responses in survey questions: an application of multiple correspondence analysis. *J Off Stat*. 2001;17:351–367.
54. Baka AS, Figgou L, Triga V. 'Neither agree, nor disagree': a critical analysis of the middle answer category in Voting Advice Applications. *Int J Electron Govern*. 2012;5:244–263.
55. Nadler JT, Weston R, Voyles EC. Stuck in the middle: the use and interpretation of mid-points in items on questionnaires. *J Gen Psychol*. 2015;142:71–89.
56. Alwin DF. *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2007.
57. Tourangeau R, Couper MP, Conrad FG. Spacing, position, and order: interpretive heuristics for visual features of survey questions. *Publ Opin Q*. 2004;68:368–393.
58. DeVellis RF. *Scale Development: Theory and Applications*. fourth ed. NC: SAGE Publications, Inc; 2017. Chapel Hill.
59. Billiet J, McClendon MJ. Modeling acquiescence in measurement models for two balanced sets of items. *Struct Equ Model*. 2000;7:608–628.
60. Berkowitz NH, Wolkon GH. A forced choice form of the F scale-free of acquiescent response set. *Sociometry*. 1964;27:54–65.
61. Ross CK, Steward CA, Sinacore JM. A comparative study of seven measures of patient satisfaction. *Med Care*. 1995;33:392–406.
62. Churchill GA. A paradigm for developing better measures of marketing constructs. *J Market Res*. 1979;16:64–73.
63. Furr RM. *Scale Construction and Psychometrics for Social and Personality Psychology*. Thousand Oaks, CA: SAGE Publications Inc; 2011.
64. Paulhus DL. Measurement and control of response bias. In: Robinson JP, Shaver PR, Wrightsman LS, eds. *Measures of Personality and Social Psychological Attitudes*. San Diego, CA: Academic Press; 1991:17–59.
65. Baker-Prewitt M, Miller J. *Mobile Research Risk: What Happens to Data Quality when Respondents Use a Mobile Device for a Survey Designed for a PC*. San Francisco, CA: CASRO Online Research Conference; 2013.
66. Weijters B, Baumgartner H. Misresponse to reversed and negated items in surveys: a review. *J Market Res*. 2012;49:737–747.
67. Eifermann RR. Negation: a linguistic variable. *Acta Psychol*. 1961;18:269–278.
68. Fraenkel T, Schul Y. The meaning of negated adjectives. *Intercult Pragmat*. 2008;5:517–540.
69. Gehlbach H. Seven survey sins. *J Early Adolesc*. 2015;35:883–897.
70. Hodge DR, Gillespie D. Phrase completions: an alternative to Likert scales. *Soc Work Res*. 2003;27:45–55.
71. Babakus E, Boller GW. An empirical assessment of the SERVQUAL scale. *J Bus Res*. 1992;24:253–268.
72. Rubright JD, Cary MS, Karlawish JH, Kim SYH. Measuring how people view biomedical research: reliability and validity analysis of the research attitudes questionnaire. *J Empir Res Human Res Ethics*. 2011;6:63–68.
73. Schriesheim CA, Hill KD. Controlling acquiescence response bias by item reversals: the effect on questionnaire validity. *Educ Psychol Meas*. 1981;41:1101–1114.

74. Zhang X, Savalei V. Improving the factor structure of psychological scales:the expanded format as an alternative to the Likert scale format. *Educ Psychol Meas*. 2016;76:357–386.
75. Tsang S, Royse CF, Terkawi AS. Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi J Anaesth*. 2017;11:S80–S89.
76. Juniper EF, Guyatt GH, Jaeschke R. How to develop and validate a new health-related quality of life instrument. In: Spilker B, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials*. second ed. Lippincott Williams & Wilkins; 1996:49–56.
77. Messick S. Test validity and the ethics of assessment. *Am Psychol*. 1980;35:1012–1027.
78. Scherpenzeel AC, Saris WE. The validity and reliablity of survey questions: a meta-analysis of MTMM studies. *Socio Methods Res*. 1997;25:341–383.
79. de Leeuw E, Berzelak N. Survey mode or survey modes? In: Wolf C, Joye D, Smith TW, Fu Y, eds. *The SAGE Handbook of Survey Methodology*. Los Angeles, CA: SAGE Publications Ltd; 2016:142–156.
80. Christian LM, Dillman DA, Smyth JD. The effects of mode and format on answers to scalar questions in telephone and web surveys. In: Lepkowski JM, Tucker C, Brick JM, et al., eds. *Advances in Telephone Survey Methdology*. Hoboken, NJ: John Wiley & Sons, Inc; 2008:250–275.