The Reliability of Survey Measures
RESULTS Series

QUESTION FORM AND RELIABILITY OF MEASUREMENT

Duane F. Alwin

Pennsylvania State University and the University of Michigan

Paula A. Tufiș

University of Bucharest

Daniel N. Ramírez

University of Wisconsin-Madison

DECEMBER 2023

Suggested citation: Alwin, D.F., Tufiș, P.A. & Ramírez, D. N. (2023). Question Form and Reliability of Measurement. *The Reliability of Survey Measures Results Series.*

Introduction

This document summarizes our results to date on various attributes of survey questions and their potential role in the quality of measurement in responses to survey questions. Isolating the effects associated with the formal attributes of survey questions on reliability is important, but challenging because in practice the attributes of questions are confounded with question content and question context (see Table 1 presented below). However, in many cases critical comparisons of specific forms of questions can be isolated by carefully controlling for other pertinent features of question content, question context and other attributes of question form, e.g. looking only at closed-form questions and analyzing differences in their formal features relative to measurement error. To date, several investigators have examined the formal characteristics of survey questions believed to affect the quality of measurement, using measurement reliability and validity as evaluation criteria. And there is an emerging empirical literature on questionnaire design and measurement errors, e.g., Alwin and Krosnick, 1991; Alwin, 1989, 1992, 2007; Madans, Miller, Maitland and Willis, 2011; Saris and Gallhofer, 2007; Schaeffer and Dykema (2020), and others, a literature which provides the background for this presentation.

Framing the Question of 'Question Form'

One broad distinction concerning question form is whether the questions are in an "open" vs. "closed" form. Open questions do not present the respondent with a set of response options, letting the respondent provide their own categories, e.g. Q: "What kind of work do you do? A: I am a brick molder." This information will then be coded into some type of occupational standing variable (e.g. an occupational prestige score or a socio-economic index (SEI) score. Open-ended questions are the "tried and true" approach to gathering such types of factual content, employed near-universally in qualitative research. On the other hand, the "closed" form provides the respondent with a fixed set of response categories from which to reply, e.g., a Likert-scale approach, "Do you think climate

change is happening as environmental scientists argue? Do you (a) strongly agree, (b) agree, (c) neither disagree nor agree, (d) disagree, or (e) strongly disagree?" The respondent then chooses a response category that reports their subjective assessment of his/her position on the object/issue.

With respect to the background for considering these two broad approaches, it is interesting that, according to Converse (1987), one of the earliest areas of debate about questionnaire design in survey research originated in the U.S. during World War II when surveys of the American public were gaining a foothold. The debate was about whether surveys should use "open" or "closed" question forms. Among those early survey researchers, the "closed question" or "forced choice" approach won out for most purposes, but even in the early 1980s concern that such questions might lead respondents to ignore unconsidered options was sufficient to warrant further study (see, e.g. Schuman and Presser, 1981).

Regarding closed-form questions, there are several attributes of questions that have been studied, that might affect the quality of measurement. Comparisons among closed-form response formats have focused on variations involving certain key attributes. Some such examples studied in this research program include: the number of response options provided in questions that ask for ratings, the use of unipolar versus bipolar response formats, the use of visual aids (aka "showcards"), the use of verbal labeling for response categories, the provision of an explicit Don't Know option, and the length of questions (see Alwin and Krosnick, 1991; Alwin, 1992, 2007; Alwin and Beattie, 2016; Alwin, Beattie, and Baumgartner, 2015; Alwin, Baumgartner, and Beattie, 2018). We address some of this literature here.

One of the reasons for examining the role of question form as a variable affecting survey quality is that some of the effects of question attributes (e.g., type of rating scale) may be masquerading as a context effect. In other words, measurement reliability may have more to do with the formal properties of questions appearing in different types of questionnaire units (stand alone,

series or batteries) than with question context *per se*. If differences in reliability can be linked to formal attributes of questions, we can perhaps pay less attention to the way in which survey questionnaires are organized, [that is, as *stand alone* questions, or as part of a *series* of questions on the same general topic, or as questions in *batteries* that have the exact same response format], and more to the characteristics of the questions themselves. Here we review some of the findings of this project relating to the effects of the formal attributes of survey questions on measurement reliability.

Open- and Closed-form Questions

We begin with an overview of broadest possible attributes of questions across the contemporary landscape of survey questions—the distinction between the open- versus closed-questions. Open-ended questions clearly have many uses, across the spectrum of social science methods, but in modern survey research the "standard" approach formulates questions with a fixed set of response categories. In most instances, carefully pre-tested closed-form questions are preferred to open-ended questions because of their greater specificity (Converse and Presser, 1986, p. 325). Open-ended questions are, however, employed widely, but selectively, and they have proven their mettle. Differences in measurement quality, however, may be due <u>not</u> to question form itself, that is, rather than to the types of content measured using open and closed questions. For example, factual content is often measured using open-ended forms, and coded later, e.g. "what was the year and month of your birth. At the same time, few non-factual questions, especially attitudes and beliefs, are open-ended. Furthermore, closed-form questions almost always appear within topical series and batteries, as opposed to a stand-alone format. Some evidence suggests that information assessed using *open-ended* response formats tends to be more reliable than questions with *closed-form* response formats, but this is likely due in part to question content (Alwin, 2007, pp. 183-185). In any event, such results suggest that researchers might more profitably exploit the open-ended approach, at least when measuring factual content. At the same time, because these two approaches often measure

different types of content, it is difficult to compare them with respect to measurement reliability.

Table 1. Distribution of GSS questions by content, context and form

	Context					Total
		Series		Battery		
Response Format/Content	Stand-alone	with introduction	without introduction	with introduction	without introduction	
Open-ended						
Facts	18		58		---	
Non-facts	3		3		---	
Total	21		61		---	82
						13.5%
Closed-form						
Facts	3	3	15		---	
Non-facts	93	48	103	244	18	
Total	96	51	118	244	18	527
						86.5%
Total		51	179	244	18	
		8.4%	29.4%	40.1%	3.0%	
	117	230		262		609
	19.2%	37.8%		43.0%		100%

Table 1 above presents the distribution of questions in the General Social Surveys (GSS) from the three GSS panels studied here by content, context, and form. We include self- and proxy-reports for non-redundant reports.[1] Like many other surveys, the GSS includes a variety of types of question contexts. We rely on Andrews' (1984) definitions of these question contexts: stand-alone, series and batteries (SASB). In this table we present a breakdown of the GSS (non-redundant self- and proxy-report) questions with respect to the intersection of question content (factual vs. non-factual), questionnaire context (SASB), and question form (open-ended vs. closed-form questions)

[1] Proxy reports include respondent reports about other people (e.g. mother's education). If multiple scalings of information exist in our database (e.g. Duncan SEI and. occupational prestige scores), we include only one of them in our analysis; thus, we include only non-redundant questions. Variables derived from original questions (e.g. church attendance) and synthesized from multiple sources (e.g. the occupation variables) are included in the open-ended category.

to clarify the different questionnaire organization for different content. In this table, we have added one feature to the distinctions used in Andrews' (1984) research, that is, whether the series or battery includes an introduction.

We consider the GSS organization of questions to be representative of many surveys, although not representative of all surveys. In this regard, the GSS panel studies include a substantial number of questions in batteries, especially in the measurement of non-factual content. As displayed in Table 1, nearly 43 percent of GSS questions appear in batteries, another 38 percent are in series, and the remaining 19 percent are stand-alone questions. Like the typical survey, virtually all GSS batteries include an introduction, a major feature of units of organization in survey questionnaires, whereas only approximately 22 percent of series do. These results reinforce the observation that in the GSS most questions about facts appear in series, whereas by contrast, questions involving non-facts appear mostly in series and batteries, and many fewer are presented as stand-alone questions. This table nicely illustrates the confounding of question context and question content, which ultimately makes it difficult to evaluate the effects of question-form sources of variation in measurement error.

Closed-form Questions

The question and response format of "fixed-form" or "closed-form" questions are usually considered as "one package," that it, it is difficult separate question and response attributes (see Dillman, 2007). There is, however, some value in examining the contribution of the response format *per* se to measurement error. Closed questions employ several types of response formats, which may be related to reliability: e.g., Likert-type or "agree-disagree" questions, "forced-choice" questions (with two or three options), "feeling thermometers," and various other kinds of rating scales (see, e.g., Alwin, 2007, pp. 185-191).

The Use of Middle Categories

The simplest and perhaps the most efficient survey questions use a 2-category (i.e., yes or no, agree vs. disagree, fair or unfair, etc.) format. One of the debates in the survey methods literature has focused on whether such questions are adequate and whether it is valuable to include a middle category in what are essentially binary choices, e.g. agree vs. disagree, yes vs. no, etc. (see Schuman and Presser, 1981, and a recent review by Sturgis, Roberts, and Smith, 2014). It may be less of an issue with longer category scales because they allow the separation of weak positives and negatives from the neutral category. Alwin, Baumgartner and Beattie (2018) suggest that comparing 4- and 5-category scales is essentially a comparison involving a scale with or without the middle category. They found there were essentially no differences between the two approaches in bipolar scales. By way of contrast, there may be several reasons why differences in estimated reliability are associated with the use of 2- vs. 3-category questions.

We can examine this issue using the GSS panel surveys, as they include a range of 2- and 3-category questions, measuring both unipolar and bipolar concepts are employed to measure attitudes, beliefs and values. In Table 2 we present a comparison of reliability estimates for 2- vs. 3-category formats from the GSS panels, partitioned by polarity of the concept. This table is limited to the comparison of unipolar and bipolar closed-form questions for self-reports of non-factual content. A set of similar results are presented by Alwin, Baumgartner, and Beattie (2018). Our results here use a slightly different categorization of GSS questions, but previous substantive conclusions are unchanged.

Table 2. Comparison of reliability estimates for unipolar and bipolar closed-form non-fact measures using 2- and 3-category response options

Number of	Polarity			Unipolar vs Bipolar	
Response Categories	Unipolar	Bipolar	Total	F-ratio	p-value
2 Categories	0.773	0.788	0.776	0.522	0.471
	(129)	(36)	(165)		
3 Categories	0.623	0.648	0.638	1.828	0.178
	(60)	(93)	(153)		
Total	0.725	0.687	0.710		
	(189)	(129)	(318)		
Comparisons					
2 vs 3 Categories					
F-ratio	69.518	45.619	121.806		
p-value	0.000	0.000	0.000		

With only a few exceptions, the unipolar versus bipolar difference is not revealed in the comparisons across this variable within 2- and 3-category questions. There does, however, appear to be an important difference here between 2- and 3-category questions within unipolar and within bipolar content. The typical non-factual question involving 2 response categories enjoys a level of reliability almost as high as that involving factual content. Across both unipolar and bipolar content assessments, the estimated reliabilities for 2-category questions is about .77; whereas, as shown in Table 2, reliabilities for 3-category questions are significantly lower. This highly significant result argues in favor of 2-category response scales in the measurement of subjective content, whether the questions are unipolar or bipolar.

Volunteered Middle Categories

Even when 2-category scales are provided, respondents sometimes voluntarily offer a "middle category" response, such as "somewhere in between," "it depends," or "neutral." The question is whether these responses should be ignored, on the assumption that the respondent did not actually

answer the questions in terms of the response categories provided. Or, should this be considered additional information that should be included in the data. Our research has compared the reliability results using 2 versus 3 categories among questions in which the respondent was presented with only 2 categories, but the interviewers were instructed to record "middle category" responses that were volunteered.

The results of this study strongly suggest that the practice of recording volunteered middle category responses when they are not offered as part of the question is not good from the point of view of the reliability of responses. Including the volunteered middle category produces reliabilities of less than .7, whereas the treatment of the volunteered middle category as missing data produces reliabilities in excess of .8, in most cases (see Table 3). These results are replicated across the two different approaches to the treatment of missing data (see Alwin, Baumgartner, and Beattie, 2018, pp. 228-229). These patterns should discourage the use of such volunteered middle category responses in analyses of these data.

Table 3. Comparison of reliability estimates for two-category questions with and without volunteered middle category: GSS panel studies—FIML estimates

				Middle Category		
				Without	With	Difference
Topic	Mnemonic	Triad[1]	Description of Question Content			
0103	AGED	20022	Should older people live with their grown children	0.724	0.587	0.137
0104	COURTS	20110	How harshly do the courts deal with criminals	0.861	0.688	0.173
0102	DIVLAW	20122	Should divorce be easier or more difficult to obtain	0.844	0.692	0.152
0108	FAIR	20162	People try to take advantage or should they try to be fair	0.799	0.719	0.080
0108	HELPFUL	20244	People try to be helpful or are just looking out for themselves	0.736	0.660	0.076
0109	RACOPEN	20907	Favor or oppose open housing laws	0.657	0.619	0.038
0301	RELITEN	20616	Would R call self a strong or not very strong (denomination)	0.916	0.847	0.069
0108	TRUST	20757	People can be trusted or you can't be too careful	0.831	0.783	0.048
			Average	0.796	0.699	0.097

Note: results are averaged over all three GSS panel studies

[1] *GSS 2006 triad numbers listed*

Rating Scales as Closed-form Questions

There is a long history in survey research of using what are called "rating scales" to measure subjective attitudes, beliefs, values and other reports and evaluations (see review by Krosnick and Fabrigar, 1997). One prominent approach to measuring such subjective content was introduced early in the history of modern survey research by Rensis Likert (1932), an approach that ultimately gained the moniker of the "Likert scale." Likert suggested that attitudes could be measured by presenting respondents with a five-category scale that included the measurement of three elements relating to the attitude concept: the *direction* of the attitude (e.g., agree versus disagree, approval versus disapproval, etc.), the *strength* of the attitude (e.g., agree versus strongly agree, or disagree versus strongly disagree), and a *neutral* point (neither agree or disagree) for those respondents who could not choose the alternate poles. Likert did not suggest offering an explicit "Don't know" response to distinguish between those people who had "no opinion" and those who were truly neutral, but this practice has become a well-accepted strategy in modern survey methods and has come to be associated with this approach.

We mention Likert's (1932) approach because it helps to provide a framework for discussing issues concerning the optimal number of response categories, which we consider below. His approach is recognized as one of the most practical strategies for the measurement of attitudes and other subjective variables in surveys. His approach avoided some of the more cumbersome (although perhaps more theoretically elegant) psychophysical scaling techniques introduced earlier by Thurstone and others. The "Likert scale" eventually became the textbook approach to measuring attitudes and other subjective phenomena. Such question forms have been adopted throughout the world, and although the term is used more broadly to refer to any bipolar survey question (regardless of the number of categories) that attempts to assess the direction and strength of attitudes, it should be noted that Likert did not propose rating scales of more than five categories. His initial work

presented three- and five-category scales almost exclusively.

The Likert scale, or a Likert-type scale, may be the most often used approach to measuring subjective phenomena. To distinguish between Likert's original ideas and the approaches that have introduced modifications, we use the term "Likert scale" for survey questions that are composed of an ordered five-category bipolar measure, using a set of agree-disagree (or approve-disapprove) categories, plus a neutral category – all labeled. There are several different forms such questions can take, and we refer to other bipolar measures as "Likert-type" scales. In contrast to these approaches, not all forced-choice rating type questions reflect the measurement of a bipolar concept—some are clearly unipolar in character. Such unipolar scales include a "zero point" at one end of the scale rather than a "neutral" point. Thus, unipolar measurement typically assesses questions of "how many?" "how much?" or "how often?" where the zero point is represented by categories such as "none," "not at all," and "never." The appendix lists the GSS rating questions by categories defined by the number of responses, "true Likert" and "Likert-type" questions, and polarity. For further discussion of these matters, see Alwin, Baumgartner, and Beattie (2018, pp. 214-219), who used a similar categorization of the GSS rating questions in the analysis. We have refined the criteria by which we assign questions to "true Likert" versus "Likert-type" categories here. Alwin, Baumgartner, and Beattie (2018) considered survey questions as "true Likert" if they were ordered five-category bipolar measures, using a set of agree-disagree (or approve-disapprove) categories, plus a neutral category. We add here the condition that all responses are labeled for considering a measure as "true Likert". We replicated the Alwin, Baumgartner, and Beattie (2018) analyses, using this refined definition of what a "true Likert" scale is, and the new categorization does not change the substantive results.

Table 4. Comparison of reliability estimates for unipolar and bipolar Likert and Likert-type closed-form non-fact measures using 4- and 5-category response options

Number of Response Categories	Types of Measures			
	Unipolar	Bipolar		
		Likert	Likert-type	Total Bipolar
4 Categories	0.741 (30)	0.617 (20)	0.530 (12)	0.585 (32)
5 Categories	---	0.581 (18)	0.573 (48)	0.575 (66)
Total	0.741 (30)	0.600 (38)	0.565 (60)	0.578 (98)
Comparisons				
Unipolar vs bipolar (4 Categories)				
F-ratio	30.010			
p-value	0.000			
Bipolar 4 vs 5 Categories				
F-ratio	---	1.002	1.330	0.133
p-value	---	0.324	0.254	0.716
Likert vs Likert type (4 Categories)				
F-ratio		5.428		
p-value		0.027		
Likert vs Likert type (5 Categories)				
F-ratio		0.027		
p-value		0.833		

Four versus Five Category Scales

In keeping with the need to evaluate the effects of number of response categories, while controlling for the bipolar and unipolar distinction, the GSS panels allow us to draw several essential conclusions on the issue of comparing 4 and 5 category scales for several different types of content. Table 4 presents reliability estimates on an array of different types of response scales, including 4- and 5-category questions. These results were obtained from closed-form non-factual questions, and they permit us to evaluate the reliabilities of Likert and Likert-type questions. These results permit

several observations. First, there are significant differences between questions involving unipolar content (e.g., how much, how likely, how satisfied, anchored by the category of "none" or "not at all"), as compared to bipolar questions aimed at measuring both direction and intensity, and in the case of 5-category questions, a neutral point. Unipolar questions are favored in this comparison. Second, an overall finding here is that 5-category bipolar questions that include a neutral, or middle, category are no more or less reliable than are 4-category questions. This goes against our expectations that the middle category would simply add measurement error, but this does not appear to be the case for GSS 5-category scales. For bipolar concepts, we assume the meaning of the middle category is very different for 3- vs. 5-category measures.

It is important to note that in our earlier comparisons of 2- and 3-category questions (see Table 2), there was a different result, namely that those measures with middle categories were significantly less reliable, a result also found in the Alwin, Baumgartner, and Beattie (2018) analyses. Finally, these results reveal no significant differences in measurement reliability among types of bipolar questions, that is, Likert measures versus Likert-type measures (see Alwin, Baumgartner and Beattie, 2018, pages 225-226). There is a slight tendency for Likert measures to have higher reliabilities both among 4-category questions, and among the 5-category measures, but the differences compared to Likert type measures are not statistically significant among the 5-category measures. From his research we conclude that, at least in the GSS data, the largest difference revealed is *not* in the patterns found among the bipolar measures, but between the unipolar and bipolar questions.

Number of Response Options

The above discussion raises a related question: Is there an optimal number of response categories for non-factual or subjective questions? Early research on question form focused on marginal differences, essentially ignoring the question of the number of response options used in survey

measurement. A small literature has emerged relating to the issue of the relationship of the number of response categories and reliability (Andrews, 1984; Alwin and Krosnick, 1991; Alwin, 1992, 2007; Alwin, Baumgartner and Beattie, 2018; Scherpenzeel and Saris, 1997; Saris and Gallhofer, 2007). The initial expectations in this research area were based on information theory, which argues that more categories can carry more information and thereby enhance measurement accuracy (e.g., Shannon and Weaver, 1949; Alwin and Krosnick, 1991; Alwin, 1992, 1997; Andrews, 1984; Krosnick and Fabrigar, 1997). Running counter to such expectations, however, are factors such as respondent cognitive capacity or motivation to consider large numbers of categories (Alwin, 2007, p. 192).

We found that previous results of comparisons of reliability for questions with different numbers of response categories differ depending on the approach used to estimate reliability (Alwin, 1992, 2007). Most past results have treated ordinal variables as if they were interval and used Pearson correlation coefficients. When continuous latent variables are measured using a small number of response categories, this practice attenuates estimated levels of association by comparison with those that would be estimated if an "ideal" continuous variable were measured directly. Because the degree of attenuation produced by crude categorization is directly related to the number of response categories, discussions of the number of response options must consider methods of estimation.

Several conclusions emerge from the most recent examination of these issues, which more adequately handles the statistical estimation issues by using tetrachoric and polychoric correlations to measure association (see Alwin, 2007, pp. 191-196). First, it yields little if any support for the information-theoretic view that more categories produce higher reliability. Estimated reliability does *not* increase monotonically with the number of categories. Indeed, if one were to ignore estimates for nine-category scales, one would conclude that more categories produce systematically less

reliability. Similarly, estimates provide little support for the suggestion that five-category response formats are less reliable than four- and six-category scales, or that seven-category scales are superior to all others.[2]

It is difficult, however, to evaluate the effects of number of response categories independent of polarity, that is, bipolar versus unipolar, as noted above. Specifically, we found that four-category scales appear superior for unipolar concepts (see Table 4). For bipolar measures, two-category scales show the highest reliability levels of measurement reliability, followed by three- and five-category scales (see Tables 2 and 4). It is relatively clear that seven-category scales achieve the poorest results among bipolar measures—data not shown (see Alwin, 2007; Alwin, Baumgartner and Beattie, 2018). Net of the number of response categories, there are few differences in reliability between unipolar and bipolar measures that cannot be attributed to the content measured.

[2] Alwin's (2007) results do uphold one conclusion of previous analyses: that reliability levels for nine- and 11-category scales are superior to seven-point scales.

Table 5. Comparison of reliability estimates for GSS questions using open-ended and closed-form response formats by question content and question context--FIML estimates

	Stand-alone			Series			Battery			
Response Format	Facts	Non-facts	Total	Facts	Non-facts	Total	Facts	Non-facts	Total	Total[3]
Open-ended[1]	0.903	0.728	0.878	0.839	0.910	0.842	---	---	---	0.851
	(18)	(3)	(21)	(58)	(3)	(61)	---	---	---	(82)
Closed-form										
2-categories	---	0.796	0.796	0.854	0.741	0.751	---	0.768	0.768	0.767
	---	(39)	(39)	(6)	(63)	(69)		(63)	(63)	(170)
3-categories	---	0.639	0.639	---	0.637	0.637	---	0.630	0.630	0.632
	---	(15)	(15)	---	(18)	(18)		(120)	(120)	(153)
4-categories	---	0.611	0.639	---	0.613	0.613	---	0.721	0.721	0.635
	---	(27)	(27)	---	(34)	(34)		(16)	(16)	(77)
5-categories	0.825	0.675	0.639	0.832	0.575	0.604	---	0.516	0.516	0.582
	(3)	(9)	(12)	(3)	(24)	(27)		(33)	(33)	(72)
6+categories	---	0.633	0.712	0.850	0.669	0.747	---	0.483	0.483	0.594
	---	(3)	(3)	(9)	(12)	(21)		(30)	(30)	(53)
Total closed form	0.825	0.700	0.704	0.848	0.668	0.687	---	0.637	0.637	0.666
	(3)	(93)	(96)	(18)	(151)	(169)	---	(262)	(262)	(527)
Total	0.892	0.701	0.735	0.841	0.673	0.728	---	0.637	0.637	0.691
Total n	(21)	(96)	(117)	(76)	(154)	(230)	---	(262)	(262)	(609)
F-ratio[2]	1.850	0.130	32.240	0.090	11.830	26.480	---	39.980	39.980	65.750
p-value	0.190	0.717	0.000	0.965	0.000	0.000	---	0.000	0.000	0.000

[1]Derived and synthesized variables are included with the open-ended response format
[2]Test for overall differences between open-ended and closed-form reliabilities
[3]Total reliability estimates for open-ended and closed-form items.
Note: The number of questions on which reliability estimates are based is given in parentheses.

Using the rich set of data from the GSS panels, involving a wide range of content and question forms, we can test several hypotheses regarding the number and kind of response categories relative to measurement reliability. The conclusions summarized above are largely borne out by the patterns of reliability from the GSS panels, as shown in Table 5, which presents the comparison of reliability estimates for GSS questions using open-ended and closed-form response formats, organized by question content, question context, and number of response options. These results show that for non-facts, independent of question context (series vs. batteries) two-category closed-

form questions are without question the most reliable, and in general reliability declines with additional categories. In a unipolar context, more is better, and thus, following a 3-category dip in reliability in all cases, reliability increases with more categories for some unipolar measures but declines with additional categories for bipolar content. This table, however, does not control for polarity, which may be related in opposite directions to the number of categories for some types of content (Alwin, Baumgartner and Beattie, 2018), which we take up in the next section.

Polarity and Number of Response Categories

Closed-form response formats also differ with respect to whether they use "unipolar" or "bipolar" rating scales. Typically, the ends of a bipolar scale are of the same intensity but opposite valence, whereas the ends of a unipolar scale tend to differ in amount or intensity, but not valence. Some types of content, such as attitudes, are usually measured using bipolar scales, whereas others, such as behavioral frequencies, are always measured using unipolar scales. Unipolar scales rarely use more than 5 categories, and the 3-, 4- or 5-category scales used to measure unipolar concepts are quite different from scales that measure bipolar concepts. Also, the meaning of the "middle category" in 3- and 5-category scales is quite different depending upon whether the corresponding concept is unipolar or bipolar (see Alwin, 2007, p. 185-191). In this instance, the confounding of different formal features of questions raises serious issues. One clearly cannot evaluate the effects of bipolar and unipolar scales on data quality, for example, without considering the number of response categories and vice versa.

Here we present an analysis of the number of categories taking polarity into account. In Table 6 we present a series of regression equations that include these predictors that we consider affect levels of reliability, namely polarity and the number of response categories.

Table 6. Regression of GSS reliability estimates on attributes of questions: GSS panel studies

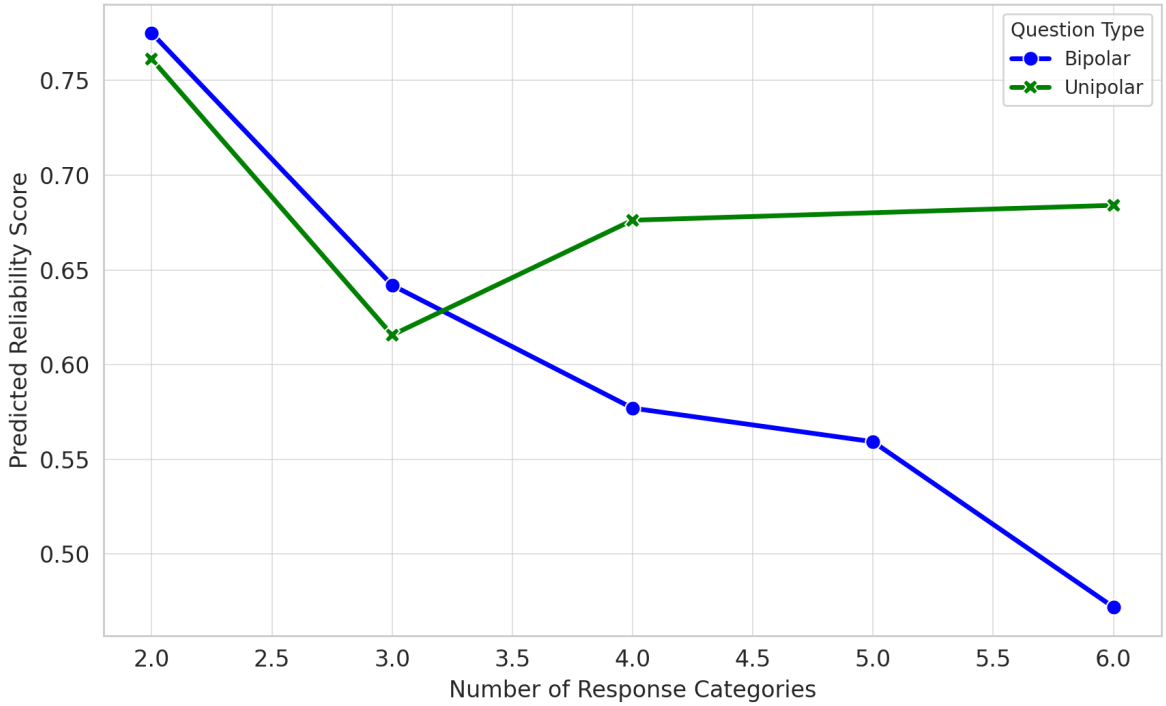
	Model					
Predictors	1		2		3	
Intercept[1]	0.764	***	0.742	***	0.775	***
Polarity (unipolar=1)			0.028	*	-0.014	
Response (3 categories)	-0.133	***	-0.122	***	-0.133	***
Unipolar*Response (3 categories)					-0.013	
Response (4 categories)	-0.129	***	-0.124	***	-0.198	***
Unipolar*Response (4 categories)					0.113	**
Response (5 categories)	-0.205	***	-0.183	***	-0.216	***
Unipolar*Response (5 categories)					---	
Response (6+ categories)	-0.222	***	-0.209	***	-0.303	***
Unipolar*Response (6+ categories)					0.226	***
R^2	0.301		0.307		0.362	
N of cases	506		506		506	

Notes: The models are restricted to non-facts, closed-form only (n = 506)
[1] The omitted categories are: in Model 1, 2 non-missing response categories; in Models 2 and 3, 2 non-missing response categories, bipolar.
*Key: + p < .1, * p < .05, ** p < .01, *** p < .001*

In these models we restrict the analysis to closed-form questions of non-factual content (n = 506), in order to gain a better purchase on the factors that impinge on respondent errors in survey measurement of subjective variables. The results in Table 6 (see model 1) for the number of response categories show that there is an almost monotonic decline in reliability with the addition of response categories – these results suggest fewer categories are better! The two-category question (yes-no, agree-disagree, etc.) is the most reliable for non-factual closed-form questions. These results parallel those presented above regarding the number of response categories, but as we noted earlier, it is important to take into account the potential interaction between the number of categories and the polarity of the underlying concept in these data. As we pointed out, to adequately specify the processes involved, we have to take the polarity of the concept being measured into account.

We include a *polarity* variable in model 2. Controlling for the number of response categories, unipolar questions are more reliable than bipolar questions (see Model 2). Due to the fact that the effects of numbers of response categories depends upon whether the question measures bipolar or unipolar dimensions, we also include interactions with the number of response categories in model 3 (see Alwin, Baumgartner and Beattie, 2018). This model's results indicate that the effect of the number of response categories on reliability depends on whether the questions are unipolar or bipolar. There is a highly significant increment in the coefficient of determination – the R^2 – for this model. The R^2 increased from .307 to .362, that is, an increase of .055. The graphical representation of the interaction effect between the number of response categories and unipolarity is presented in Figure 1. The results in Table 6 and Figure 1 indicate that reliability tends to decrease with the number of response categories for bipolar concepts. The relationship between reliability and the number of response categories is more complex for unipolar concepts, with 2 category unipolar questions being the most reliable, followed by 4 and 6 categories questions, and 3-category questions being the least reliable. There are no cases of 5-category unipolar questions in the GSS panel studies.

Figure 1. Predicted Reliability by Number of Response Categories and Unipolarity, Predicted GSS reliability scores in Model 3



Other Considerations

Concern that random behavior by respondents may introduce measurement unreliability also exists, because many non-factual questions and their response categories are vague. Respondents nonetheless perceive pressure to answer such questions even when they have little knowledge of a subject or have given little thought to it. Converse's (1964) famous critique of attitude measurement conceptualized the problem in terms of the existence of "non-attitudes," that is respondents who have no attitude but nonetheless answer the question. Converse's idea was that respondents may not always have an attitude but feel pressure during the interview to offer opinions to questions even when they have none. Respondents concocting attitude reports when they do not have an attitude produce what are essentially random choices, contributing to unreliability of measurement.

Don't Know Filters

Explicitly offering a "Don't Know" option, either as a filter, or as an option provided in the question,

may forestall such contributions to error. Numerous split-ballot experiments have found that the number of Don't Know responses is significantly greater when respondents are initially asked if they have an opinion, rather than when they must volunteer a no-opinion response (see review by Krosnick, 2002). Studies focusing on how offering such an option affects the quality of measurement do not all agree, but most find no significant difference. Andrews (1984) found that offering the Don't Know increased the reliability of attitude reports. Alwin and Krosnick (1991) found the opposite for seven-point rating scales and no differences for agree-disagree questions. McClendon and Alwin (1993) and Scherpenzeel and Saris (1997) found no differences in measurement error between forms. Alwin (2007, pp. 196-200) compared questions about non-factual content with and without an explicit Don't Know option—within three-, five-, and seven-category bipolar rating scales—and found no significant differences. The Don't Know option is rarely explicitly offered in modern surveys, and we do not know what effect its absence has on the quality of data.

Labeling Response Scales

Labeling response options can reduce ambiguity about how respondents are to translate subjective responses into categories of response scales. Simple notions of communication and information transmission suggest that better labeled response categories may be more reliable. Thus, it is reasonable to expect that the estimated reliability of subjective variables will be greater when more verbal labeling is used. Oddly, Andrews (1984, p. 432) reports below average data quality when all categories are labelled. On the other hand, research has also found that among 7-point scales, fully labeled ones were significantly more reliable (e.g., Alwin and Krosnick, 1991). More recent research suggests a significant difference in reliability between fully- and partially-labeled response categories: measures with fully-labeled categories were more reliable (see Alwin, 2007, pp. 200-202; see also Alwin et al., 2023).

Table 7. Reliability estimates for GSS questions differing in the extent of labelling of response categories--FIML estimates

Labelled Response scales				N of	Sample	Reliability
Topic	Mnemonic	Triad	Description of question content	Items	Size[1]	Estimate
0304	GOD	20203	R's confidence in the existence of god (6 labelled categories)	3	2015	0.846
0303	PRAY	20568	R's frequency of prayer (6 labelled categories)	3	2014	0.853
1501	POLVIEWS	20558	R's self-rating on liberal-conservative dimension (7 labelled categories)	3	1990	0.670
0603	SOCBAR	20704	R's frequency of going to a bar or tavern (7 labelled categories)	3	1500	0.865
0603	SOCFREND	20705	R's frequency of spending a social evening with friends (7 labelled categories)	3	1500	0.541
0603	SOCOMMUN	20706	R's frequency of spending a social evening with neighbors (7 labelled categories)	3	1500	0.583
0603	SOCREL	20707	R's frequency of spending a social evening with relatives (7 labelled categories)	3	1500	0.587
			Average			0.706

Endpoints-only labelled scales				N of	Sample	Reliability
Topic	Mnemonic	Triad	Description of question content	Items	Size	Estimate
0902	EQWLTH	20146	Should government reduce income differences (7 categories)	3	1488	0.633
0102	WLTHWHTS	20811	Group differences--whites tend to be rich or poor (7 categories)	3	1347	0.379
0102	WLTHBLKS	20810	Group differences--blacks tend to be rich or poor (7 categories)	3	1346	0.337
0102	WORKWHTS	20833	Group differences--whites tend to be hard working or lazy (7 categories)	3	1342	0.491
0102	WORKBLKS	20827	Group differences--blacks tend to be hard working or lazy (7 categories)	3	1341	0.365
0102	INTELWHTS	20308	Group differences--whites tend to be unintelligent or intelligent (7 categories)	3	1339	0.307
0102	INTELBLKS	20306	Group differences--blacks tend to be unintelligent or intelligent (7 categories)	3	1338	0.377
0102	CLOSEBLK	20068	How close does R feel to blacks (9 categories)	3	1327	0.662
0102	CLOSEWHT	20069	How close does R feel to whites (9 categories)	3	1329	0.499
			Average			0.450

[1] *Average sample size and average reliability over three GSS panels*

This issue is relevant to the GSS, because these surveys include several unlabeled response scales. The average reliabilities for GSS measures of subjective variables using labelled versus partially-labelled scales are systematically different. Such GSS scales using unlabeled categories are notoriously unreliable, as illustrated in Table 7, which compares the reliability estimates for GSS questions differing in the extent of labelling of response categories. This comparison does not control for content or context, but it nonetheless is suggestive of a pattern. We argue that unlabeled categories contribute to unreliability because if they are unlabeled each respondent uses a potentially different meaning over time. Labels definitely help anchor respondent interpretation in ways that promote consistency/reliability. We also explored this difference using the available 5-point scales in the GSS, but in this case, there were mostly fully-labelled scales. Still, those few unlabeled 5-point scales in the GSS were much less reliable than the fully labelled ones as shown in Table 7. [3]

Question Length

Most researchers (but not all) subscribe to the view that questions should be as short as possible. Payne's (1951) early manual on surveys suggested that questions should rarely number more than 20 words. This has for decades been the mantra for experts in questionnaire design (Schaeffer and Dykema, 2020; see Marquis, Cannell & Laurent, 1972, for an opposing view). Alwin (2007, pp. 202-210) examined the relationship of question length to the reliability of measurement, controlling for question content, question context and length of introductions. He found a consistently negative length-reliability relationship for stand-alone questions and questions in series. Results were the

[3] For example, Hout and Hastings (2016, page 993) report that for the questions dealing with group differences, "treating scores 1-3 as low, 4 as medium, and 5.-7 as high and ignoring variation within high or low substantially increased the estimated reliability … [and] reducing the items to dichotomies 1-4 versus 5-7 further increased [reliabilities]." This of course raises the question of whether researchers might profitably redesign this question as dichotomous, or to use shorter fully-labelled response scales instead.

same for questions in batteries, except for the case of batteries with medium-length introductions. The exception, though not significant, poses an interesting puzzle about possible interaction between question context, question length and reliability of measurement. These results provide a relatively convincing argument that in some cases that levels of reliability decline when questions contain greater numbers of words, further supporting the typical advice that survey questions should be as short as possible (see Alwin and Beattie, 2016).

Discussion and Conclusion

Although we initially framed the issue of question form effects in terms of the broader landscape involving the distinction between open- and closed-form questions, our considerations have mainly focused on the latter. Open-form questions are often found to be somewhat more reliable than those using a fixed- or closed-form approach, but we reason that the possible advantage in reliability of open-form questions may lie in the fact that they are more likely to be used with factual rather than non-factual content, which tend to be vastly more reliable. There has not been a solid experimentally controlled test of this hypothesis, it would likely be difficult to test given the confounding with question content. Both forms of questions are valuable and have their place. In the present research, as noted, we have instead focused attention on question-form effects among closed-form questions.

This topic is important because question-form effects may account for differences observed in other elements of the survey question. For example, many believe that there may be context effects—the placement of a question within the organizational context of the questionnaire—that affect the levels of measurement error (e.g. Andrews, 1984; Alwin, 2007; Schaeffer and Dykema, 2020). Context effects are typically designated in terms of whether the question is presented as a "stand alone" question, or whether it is included in a topical series of questions, or further, whether such a topical series includes exactly the same response formats. This research has raised the issue of whether these differences may in fact be due to the types of question forms are used across these

important aspects of organizational context, specifically series and batteries, and our results support such a conclusion. If differences in reliability can be linked to formal attributes of questions, we can perhaps pay less attention to the way in which survey questionnaires are organized at a more macro-level [that is, as *stand alone* questions, or as part of a *series* of questions on the same general topic, or as questions in *batteries* that have the exact same response format], and focus attention more to the characteristics of the questions themselves, such as those reviewed here.

Thus, another reason for examining the role of question form as a variable affecting survey quality is that some of the effects of the formal properties of questions, for example, the number of response options, may provide a strong basis for preferring one type of question over another. Here we have reviewed some of the findings of this project relating to the effects of the formal attributes of survey questions on measurement reliability. The results of this research contribute to uncovering sources of measurement error in surveys associated with the form of the question, thereby ultimately potentially improving survey data collection methods through the modification of common approaches to measurement. The number of response options is a question feature that has been studied, but there are problems with much of this research. The problems have been remedied by considering a more sophisticated approach, specifically the use of tetrachoric and polychoric correlations for categoric variables, which allow formal comparisons with traditional Pearson correlations computed for interval-level (or continuous) variables.

These adjustments have allowed us to arrive at a stronger empirical basis for conclusions regarding the use of middle categories and the possible salutary effects of offering greater numbers of response categories. One of the most significant set of findings from this project is that more response categories produce better results for unipolar question formats, compared to bipolar formats, which tend to introduce more measurement errors with greater numbers of response options. Thus, for bipolar questions the fewer response options provide more reliable information.

There is therefore some evidence in our present set of results to suggest an interaction of the number of response options with the polarity of the concept measured, indicating that we must be cautions in making general sweeping statements about the overall effects of the number of response categories.

In this regard, our research suggests that the use of middle categories, especially for dichotomous variables, should be avoided. The main takeaway is that bipolar questions that offer middle categories in some cases produce unreliable responses from respondents who find it difficult to distinguish between a "weak positive" (or "weak negative) versus true neutrality. In this case, the use of the middle categories may simply produce ambiguity, or they may be chosen as a way of saying "I don't know." In the case of unipolar questions, the use of 3 categories still produces more unreliable responses compared to the use of 3-categories, but greater numbers of categories beyond 3 may enhance reliability. One takeaway from this research discourages the use of middle categories for truly bipolar dichotomous or binary variables, e.g. yes vs. no, more vs. less, etc. We also find that when middle category responses are "volunteered,' reliability appears to be enhanced if such "volunteered" middle categories are ignored and treated as missing data. Although the use of middle categories is discouraged in the use of binary variables, the difference does not appear to be replicated in this research for the comparison of 4- version 5-category scales. Our comparison of such question forms, contrary to Likert's famous proposal for question form – the "Likert scale" – the use of a middle category provides no apparent advantage.

There are potentially other lessons to be learned from these findings, but further analyses are required to establish confidence in our conclusions. As we noted at the outset, we are critically aware that there is a natural confounding of elements of question form with other features of survey design. Both question content and question context are factors that make it difficult to sort out the factors contributing to sources of measurement error—but this does not mean that it is not necessarily

impossible. Often factual questions are asked in an open-ended form, as "stand alone" questions, whereas non-facts are almost always phrased in a closed-ended form in the context of other questions in series or batteries. This dichotomy makes it hard to compare the two forms – open versus closed forms – and thus, we have focused on variation within forced-choice forms.

This line of research may contribute to substantive conclusions regarding question form that heretofore survey methodologists only speculated about. One dictum observed by survey researchers is that survey questions should be as clear and as short as possible. Longer and more ambiguous questions it is believed are generally less reliable, a principle that resonates with a frequently-given desideratum for survey questions, namely that they should be as short as possible. With the types of evidence we present here, we can potentially add support in some cases to some views of "best practices." For example, confirming the importance of labelling response categories offers a scientific basis for common sense, in that there is substantial evidence suggesting the unlabeled categories leads to substantially more measurement error. Our results strongly reinforce the longstanding survey research practice of labelling all response categories. This result, coupled with earlier findings regarding question length, further reinforces the need for greater clarity and simplicity in the improvement of the accuracy of responses to survey questions.

References

Alwin, Duane F. 1992. Information Transmission in the Survey Interview: Number of Response Categories and the Reliability of Attitude Measurement. In Peter V. Marsden (Ed.), *Sociological Methodology 1992* (Pp. 83-118). Washington D.C.: American Sociological Association.

Alwin, Duane F. 1997. Feeling Thermometers vs. Seven-point Scales: Which are Better? *Sociological Methods and Research*, 25:318-340.

Alwin, Duane F. 2007. *Margins of Error—A Study of Reliability in Survey Measurement*. Hoboken, NJ: John Wiley & Sons, Inc. [Wiley Series in Survey Methodology]

Alwin, Duane F. 2021. Developing Reliable Measures: An Approach to Evaluating the Quality of Survey Measures using Longitudinal Designs. Pp. 113-154 in Alexandru Cernat and Joseph Sakshaug (Eds.), *Measurement Error in Longitudinal Data*. Oxford, UK: Oxford University Press.

Alwin, Duane F. and Brett A. Beattie. 2016. The Kiss Principle in Survey Measurement—Question Length and Data Quality. *Sociological Methodology*, 46:121-152.

Alwin, Duane F. and Jon A. Krosnick. 1991. The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes. *Sociological Methods and Research* 20:139-181.

Alwin, Duane F., Erin M. Baumgartner, and Brett A. Beattie. 2018. Number of Response Categories and Reliability in Attitude Measurement. *Journal of Survey Statistics and Methodology* 6:212-239.

Alwin, Duane F., Brett A. Beattie, and Erin M. Baumgartner. 2015. Assessing the Reliability of Measurement in the General Social Survey: The Content and Context of the GSS Survey Questions. Paper presented at the session on "Measurement Error and Questionnaire Design,"

70th annual meetings of the American Association for Public Opinion Research, Holly wood FL, May 14, 2015.

Alwin, Duane F., Kristina Zeiser, and Don Gensimore. 2014. Reliability of Self-reports of Financial Data in Surveys: Results from the Health and Retirement Study. *Sociological Methods and Research* 43:98-136.

Alwin, Duane F., Paula A. Tufiş, Daniel N. Ramírez, Erin M. Baumgartner, and Brett A. Beattie. 2023. The Devil is in the Details: An Evaluation of the Reliability of Measurement in the General Social Survey. Unpublished paper.

Andrews, Frank M. 1984. Construct Validity and Error Components of Survey Measures: a Structural Modeling Approach. *Public Opinion Quarterly* 46:409-42.

Converse, Philip E. 1964. The Nature of Belief Systems in the Mass Public. In D.E. Apter (Ed.), *Ideology and Discontent* (pp. 206-261). New York: Free Press.

Converse, Jean M. 1987. *Survey Research in the United States: Roots and Emergence, 1890-1960.* Berkeley, CA: University of California Press.

Converse, Jean M. and Stanley Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire.* Beverly Hills, CA: Sage.

Dillman, Don A. 2007. *Mail and Internet Surveys—The Tailored Design Method* (2nd edition). John Wiley & Sons, Inc .

Heise, David R. 1969. Separating Reliability and Stability in Test-retest Correlation. *American Sociological Review* 34:93-191.

Hout, Michael and Orestes P. Hastings. 2016. Reliability of the Core Items in the General Social Survey: Estimates from the Three-Wave Panels, 2006-2014. *Sociological Science* 3:971-1002.

Krosnick, Jon A. 2002. The Causes of No-Opinion Responses to Attitude Measures in Surveys: They Are Rarely What They Appear To Be. Pp. 87-100 in R.M. Groves, D.A. Dillman, J.L.

Eltinge, and R.J.A. Little (Eds.). 2002. *Survey Nonresponse*. New York: John Wiley and Sons.

Krosnick, Jon A., and Leandre R. Fabrigar. 1997. Designing Rating Scales for Effective Measurement in Surveys. Pp. 141-164 in Lars Lyberg and others (Eds.), *Survey Measurement and Process Quality.* New York, NY: John Wiley & Sons, Inc.

Likert, Rensis. 1932. "A Technique for the Measurement of Attitudes," *Archives of General Psychology,* 140:5-55.

Madans, Jennifer, Kristen Miller, Aaron Maitland and Gordon Willis (Eds.). 2011. *Question Evaluation Methods: Contributing to the Science of Data Quality.* Hoboken, NJ: John Wiley & Sibs,

Marquis, Kent H., Charles F. Cannell, and A. Laurent. 1972. Reporting Health Events in Household Interviews: Effects of Reinforcement, Question Length, and Reinterviews. *Vital and Health Statistics*. Series 2, No. 45.

McClendon, McKee .J. and Duane F. Alwin. 1993. No-opinion Filters and Attitude Measurement Reliability. *Sociological Methods and Research* 21:438-464.

Payne, Stanley L. 1951. *The Art of Asking Questions*. Princeton, NJ: Princeton University Press.

Saris, Willem E. and Irmtraud Gallhofer. 2007. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. New York NY: John Wiley & Sons.

Schaeffer, Nora Cate and Jennifer Dykema. 2020. Advances in the Science of Asking Questions. *Annual Review of Sociology*, 46:37-60.

Scherpenzeel, Annette C. and Willem.E. Saris. 1997. The Validity and Reliability of Survey Questions: A Meta-Analysis of MTMM Studies. *Sociological Methods and Research* 25:341-383.

Schuman, Howard, and Stanley Presser. 1981. *Questions and Answers: Experiments in Question Wording, Form and Context*. New York: Academic Press.

Shannon, Claude E. and Warren Weaver. 1949. *The Mathematical Theory of Communication.* Urbana, IL: University of Illinois Press.

Sturgis, Patrick, Caroline Roberts, and Patten Smith. 2014. Middle Alternatives Revisited: How the Neither/Nor Response Acts as a Way of Saying 'I Don't Know'. *Sociological Methods and Research*, 43:15-38.

APPENDIX A—GSS subjective questions

This appendix lists, by GSS mnemonic, the questions used in response categories analysis, including the two-, three-, four-, and five-category "true Likert" and "Likert-type" questions and the four-category unipolar questions, as well as all other questions used from the GSS data. The "code numbers" in parentheses are primarily for internal study purposes only, but they indicate the number of replications for a given question. Code numbers beginning with "20" refer to the 2006-08-10 panel, those beginning with "30" refer to the 2008-10-12 panel, and those beginning with "40" refer to the 2010-12-14 panel.

2-category bipolar (n of items = 36)

AGED2 (20920, 30920, 40920) (good idea, bad idea)
CAPPUN (20057, 30057, 40057) (favor, oppose)
COURTS2 (20921, 30921, 40921) (too harshly, not harshly enough)
DIVLAW2 (20922, 30922, 40922) (easier, more difficult)
FAIR2 (20923, 30923, 40923) (take advantage, try to be fair)
FEPOL (20176, 30176, 40176) (agree, disagree)
GUNLAW (20229, 30229, 40229) (favor, oppose)
HELPFUL2 (20925, 30925, 40925) (try to be helpful, looking out for themselves)
PRAYER (20569, 30569, 40569) (approve, disapprove)
RACOPEN2 (20926, 30926, 40926) (first law, second law)
SEXEDUC (20666, 30666, 40666) (for, against)
TRUST2 (20927, 30927, 40927) (can be trusted, can't be too careful)

2-category unipolar (n of items = 129)

ABANY (20001, 30001, 40001) (yes, no)
ABDEFECT (20002, 30002, 40002) (yes, no)
ABHLTH (20003, 30003, 40003) (yes, no)
ABNOMORE (20004, 30004, 40004) (yes, no)
ABPOOR (20005, 30005, 40005) (yes, no)
ABRAPE (20006, 30006, 40006) (yes, no)
ABSINGLE (20007, 30007, 40007) (yes, no)
COLATH (20071, 30071, 40071) (allowed, not allowed)
COLCOM (20073, 30073, 40073) (fired, not fired)
COLHOMO (20075, 30075, 40075) (allowed, not allowed)
COLMIL (20076, 30076, 40076) (allowed, not allowed)
COLRAC (20078, 30078, 40078) (allowed, not allowed)
FEAR (20169, 30169, 40169) (yes, no)
LETDIE1 (20333, 30333, 40333) (yes, no)

LIBATH (20335, 30335, 40335) (favor, not favor)
LIBCOM (20336, 30336, 40336) (favor, not favor)
LIBHOMO (20337, 30337, 40337) (favor, not favor)
LIBMIL (20338, 30338, 40338) (favor, not favor)
LIBRAC (20340, 30340, 40340) (favor, not favor)
POLABUSE (20536, 30536, 40536) (yes, no)
POLATTAK (20545, 30545, 40545) (yes, no)
POLESCAP (20549, 30549, 40549) (yes, no)
POLHITOK (20905, 30905, 40905) (yes, no)
POLMURDR (20556, 30556, 40556) (yes, no)
POSTLIFE (20563, 30563, 40563) (yes, no)
RACDIF1 (20583, 30583, 40583) (yes, no)
RACDIF2 (20584, 30584, 40584) (yes, no)
RACDIF3 (20585, 30585, 40585) (yes, no)
RACDIF4 (20586, 30586, 40586) (yes, no)
GRASS (20213, 30213, 40213) (should, should not)
REBORN (20908, 30908, 40908) (yes, no)
RELITEN2 (20616, 30616, 40616) (Strong, not very strong)
RICHWORK (20626, 30626, 40626) (continue working, stop working)
SPKATH (20723, 30723, 40723) (yes allowed, not allowed)
SPKCOM (20724, 30724, 40724) (yes allowed, not allowed)
SPKHOMO (20725, 30725, 40725) (yes allowed, not allowed)
SPKMIL (20727, 30727, 40727) (yes allowed, not allowed)
SPKRAC (20728, 30728, 40728) (yes allowed, not allowed)
SUICIDE1 (20740, 30740, 40740) (yes, no)
SUICIDE2 (20741, 30741, 40741) (yes, no)
SUICIDE3 (20744, 30744, 40744) (yes, no)
SUICIDE4 (20745, 30745, 40745) (yes, no)
USWARY (20773, 30773, 40773) (yes, no)

3-category bipolar (n of items = 93)

FINALTER (20179, 30179, 40179) (getting better, getting worse, stayed the same)
GETAHEAD (20197, 30197, 40197) (hard work most important, hard work and luck equally important, luck most important)
NATAID (20408, 30408, 40408) (too much, too little, about right)
NATAIDY (20409, 30409, 40409) (too much, too little, about right)
NATARMS (20410, 30410, 40410) (too much, too little, about right)
NATARMSY (20411, 30411, 40411) (too much, too little, about right)
NATCHLD (20412, 30412, 40412) (too much, too little, about right)
NATCITY (20413, 30413, 40413) (too much, too little, about right)
NATCITYY (20414, 30414, 40414) (too much, too little, about right)
NATCRIME (20415, 30415, 40415) (too much, too little, about right)
NATCRIMY (20416, 30416, 40416) (too much, too little, about right)
NATDRUG (20417, 30417, 40417) (too much, too little, about right)
NATDRUGY (20419, 30418, 40418) (too much, too little, about right)
NATEDUC (20419, 30419, 40419) (too much, too little, about right)

NATEDUCY (20420, 30420, 40420) (too much, too little, about right)
NATENVIR (20422, 30422, 40422) (too much, too little, about right)
NATENVIY (20423, 30423, 40423) (too much, too little, about right)
NATFARE (20424, 30424, 40424) (too much, too little, about right)
NATFAREY (20425, 30425, 40425) (too much, too little, about right)
NATHEAL (20426, 30426, 40426) (too much, too little, about right)
NATHEALY (20427, 30427, 40427) (too much, too little, about right)
NATMASS (20428, 30428, 40428) (too much, too little, about right)
NATPARK (20429, 30429, 40429) (too much, too little, about right)
NATRACE (20430, 30430, 40430) (too much, too little, about right)
NATRACEY (20431, 30431, 40431) (too much, too little, about right)
NATROAD (20432, 30432, 40432) (too much, too little, about right)
NATSCI (20433, 30433, 40433) (too much, too little, about right)
NATSOC (20434, 30434, 40434) (too much, too little, about right)
NATSPAC (20435, 30435, 40435) (too much, too little, about right)
NATSPACY (20436, 30436, 40436) (too much, too little, about right)
TAX (20746, 30746, 40746) (too high, about right, too low)

3-category unipolar (n of items = 60))

BIBLE (20035, 30035, 40035) (Bible word of God, Bible inspired by God, Bible written by man)
CONARMY (20086, 30086, 40086) (a great deal, only some, hardly any at all)
CONBUS (20087, 30087, 40087) (a great deal, only some, hardly any at all)
CONCLERG (20089, 30089, 40089) (a great deal, only some, hardly any at all)
CONEDUC (20093, 30093, 40093) (a great deal, only some, hardly any at all)
CONFINAN (20095, 30095, 40095) (a great deal, only some, hardly any at all)
CONJUDGE (20097, 30097, 40097) (a great deal, only some, hardly any at all)
CONLABOR (20098, 30098, 40098) (a great deal, only some, hardly any at all)
CONLEGIS (20099, 30099, 40099) (a great deal, only some, hardly any at all)
CONMEDIC (20100, 30100, 40100) (a great deal, only some, hardly any at all)
CONPRESS (20101, 30101, 40101) (a great deal, only some, hardly any at all)
CONSCI (20102, 30102, 40102) (a great deal, only some, hardly any at all)
CONTV (20105, 30105, 40105) (a great deal, only some, hardly any at all)
DISCAFF (20119, 30119, 40119) (very likely, somewhat likely, not very likely)
HAPMAR (20235, 30235, 40235) (very happy, pretty happy, not too happy)
HAPPY (20236, 30236, 40236) (very happy, pretty happy, not too happy)
JOBFIND (20317, 30317, 40317) (very easy, somewhat easy, not easy at all)
LIFE (20341, 30341, 40341) (exciting, routine, dull)
PORNLAW (20562, 30562, 40562) (laws against whatever the age, laws agains for persons under 18, no laws forbidding)
SATFIN (20638, 30638, 40638) (pretty well satisfied, more or less satisfied, not satisfied at all)

4-category true Likert (n of items = 20)

BLKWHITE (20049) (agree strongly, agree somewhat, disagree somewhat, disagree strongly)
FECHLD (20170, 30170, 40170) (strongly agree, agree, disagree, strongly disagree)
FEFAM (20171, 30171, 40171) (strongly agree, agree, disagree, strongly disagree)

FEPRESCH (20178, 30178, 40178) (strongly agree, agree, disagree, strongly disagree)
PERMORAL (20530) (agree strongly, agree somewhat, disagree somewhat, disagree strongly)
PILLOK (20535, 30535, 40535) (strongly agree, agree, disagree, strongly disagree)
PUNSIN (20581) (agree strongly, agree somewhat, disagree somewhat, disagree strongly)
RELLIFE (20617) (strongly agree, agree, disagree, strongly disagree)
ROTAPPLE (20631) (agree strongly, agree somewhat, disagree somewhat, disagree strongly)
SPANKING (20714, 30714, 40714) (strongly agree, agree, disagree, strongly disagree)

4-category Likert-type (n of items = 12)

AFIRMACT (20020, 30020, 40020) (strongly support, support, oppose, strongly oppose)
DISCAFFM (20120, 30120, 40120) (very likely, somewhat likely, somewhat unlikely, very unlikely)
DISCAFFW (20121, 30121, 40121) (very likely, somewhat likely, somewhat unlikely, very unlikely)
FEJOBAFF (20174, 30174, 40174) (strongly in favor, favor, oppose, strongly oppose)

Unipolar 4-category (n of items = 30)

HOMOSEX (20262, 30262, 40262) (always wrong, almost always wrong, wrong only sometimes, not wrong at all)
HEALTH (20240, 30240, 40240) (excellent, good, fair, poor)
JOBLOSE (20319, 30319, 40319) (very likely, fairly likely, not too likely, not at all likely)
PREMARSX (20570, 30570, 40570) (always wrong, almost always wrong, wrong only sometimes, not wrong at all)
RELPERSN (20619, 30619, 40619) (very religious, moderately religious, slightly religious, not religious at all)
CLASS (20065, 30065, 40065) (lower class, working class, middle class, upper class)
SATJOB (20639, 30639, 40639) (very satisfied, somewhat satisfied, not too satisfied, not at all satisfied)
SPRTPRS (20731, 30731, 40731) (very spiritual, moderately spiritual, slightly spiritual, not spiritual at all)
TEENSEX (20751, 30751, 40751) (always wrong, almost always wrong, wrong only sometimes, not wrong at all)
XMARSEX (20916, 30916, 40916) (always wrong, almost always wrong, wrong only sometimes, not wrong at all)

5-category true Likert (n of items = 18)

FEHIRE (20172, 30172, 40172) (agree, disagree)
GOODLIFE (20205, 30205, 40205)
MARHOMO (20359, 30359, 40359) (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree)
MEOVRWRK (20391, 30391, 40391) (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree)
WRKWAYUP (20841, 30841, 40841) (agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, disagree strongly)

INCGAP (30315) (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree)
INEQUAL3 (30322) (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree)
INEQUAL5 (30970) (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree)

5-category Likert-type (n of items = 48)

FINRELA (20182, 30182, 40182) (far below average, below average, average, above average, far above average)
HELPBLK (20243, 30243, 40243) (partially labelled: strongly agree government should, agree with both, strongly agree people take care of themselves)
HELPNOT (20247, 30247, 40247) (partially labelled: strongly agree government should, agree with both, strongly agree people take care of themselves)
HELPPOOR (20250, 30250, 40250) (partially labelled: strongly agree government should, agree with both, strongly agree people take care of themselves)
HELPSICK (20251, 30251, 40251) (partially labelled: strongly agree government should, agree with both, strongly agree people take care of themselves)
INCOM16 (20292, 30292, 40292) (far below average, below average, average, above average, far above average)
KIDSOL (20323, 30223, 40223) (much better, somewhat better, about the same, somewhat worse, or much worse than now)
LETIN1 (20334, 30334, 40334) (increased a lot, increased a little, remain the same, reduced a little, reduced a lot)
LIVEBLKS (20344, 30344, 40344) (very much in favor, somewhat in favor, neither in favor nor opposed, somewhat opposed, very much opposed)
LIVEWHTS (20346, 30346, 40346) (very much in favor, somewhat in favor, neither in favor nor opposed, somewhat opposed, very much opposed)
PARSOL (20504, 30504, 40504) (much better, somewhat better, about the same, somewhat worse, much worse)
POPESPKS (20559, 30559, 40559) (certainly true, probably true, uncertain whether true or false, probably false, certainly false)
MARBLK (20901, 30901, 40901) (very in favor, somewhat in favor, neither in favor nor opposed, somewhat opposed, very much opposed)
MARASIAN (20355, 30355, 40355) (very in favor, somewhat in favor, neither in favor nor opposed, somewhat opposed, very much opposed)
MARHISP (20358, 30358, 40358) (very in favor, somewhat in favor, neither in favor nor opposed, somewhat opposed, very much opposed)
MARWHT (20361, 30361, 40361) (very in favor, somewhat in favor, neither in favor nor opposed, somewhat opposed, very much opposed)

Unipolar 6- and 7-category (n of items = 15)

GOD (20203, 30203, 40203) (fully labelled 6 categories: range from don’t believe in God, God exists no doubts)

SOCBAR (20704, 30704, 40704) (fully labelled 7 categories: almost every day, once or twice a week, several times a month, about once a month, several times a year, about once a year, never)
SOCFREND (20705, 30705, 40705) (fully labelled 7 categories: (fully labelled: almost every day, once or twice a week, several times a month, about once a month, several times a year, about once a year, never)
SOCCOMMUN (20706, 30706, 40706) (fully labelled 7 categories: (fully labelled: almost every day, once or twice a week, several times a month, about once a month, several times a year, about once a year, never)
SOCREL (20707, 30707, 40707) (fully labelled: (fully labelled 7 categories: almost every day, once or twice a week, several times a month, about once a month, several times a year, about once a year, never)

Bipolar 7-category (n of items = 24)

EQWLTH (20146, 30146, 40146) (end points only: government should do something, government should not concern itself)
INTLBLKS (20306, 30306, 40306) (end points only: unintelligent to intelligent)
INTLWHTS (20308, 30308, 40308) (end points only: unintelligent to intelligent)
POLVIEWS (20558, 30558, 40558) (fully labelled: extremely liberal, liberal, slightly liberal, moderate (middle of the road), slightly conservative, conservative, extremely conservative)
WLTHBLKS (20810, 30810, 40810) (end points only: rich to poor)
WLTHWHTS (20811, 30811, 40811) (end points only: rich to poor)
WORKBLKS (20827, 30827, 40827) (end points only: hard working to lazy)
WORKWHTS (20833, 30833, 40833) (end points only: hard working to lazy)

Bipolar 9-category (n of items = 6)

CLOSEBLK (20068, 30068, 40068) (partially labelled: end- and middle points only: not at all close, middle: neither one feeling or the other, to very close)
CLOSEWHT (20069, 30069, 40069) (end- and middle points only: not at all close, middle: neither one feeling or the other, to very close)