The Reliability of Survey Measures
RESULTS Series

QUESTION CONTEXT AND RELIABILITY OF MEASUREMENT

Duane F. Alwin

Pennsylvania State University and the University of Michigan

Paula A. Tufiș

University of Bucharest

December 2023

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

Introduction

Researchers who construct questionnaires for surveys typically suggest that the organization of questions into subunits larger than the question affects the quality of data, which is why, according to lore, for example, questions on the same topic should be placed in particular sections of the questionnaire, or that some questions should be placed earlier or later in an interview, or that questions on the same topic should be placed in batteries, in an effort to enhance data quality. The organization of questions within questionnaires is a topic that has rarely been studied with respect to the effects on measurement errors, but question context is beginning to be a matter of concern to those who design questionnaires (e.g., see Schaeffer and Dykema, 2020). The work that has been done on this topic (e.g., Andrews, 1984; Alwin, 2007; Scherpenzeel and Saris, 1997; Saris and Gallhofer, 2007) suggests that location in the questionnaire is unrelated to reliability of measurement; however, the context of the question within the questionnaire is a factor that potentially affects measurement precision.

The context of a question, as defined here refers to the placement of the question within the questionnaire in varying topical and format arrangements.[1] Specifically, whether the question is a *stand alone* question, unrelated to the content of adjacent questions, or in a *series* of questions pertaining to the same specific topic, or in a series of questions that not only cover the same topic, but also use the exact same response format. The latter are called *batteries* and are assumed to be relevant to the study of reliability and validity of measurement (see Andrews, 1984; Scherpenzeel

[1] In other places, the term 'question context' is used to refer to question order effects (see Schuman and Presser, 1981). This is not the way we use the term here.

and Saris, 1997). The use of batteries is sometimes thought to lower the level of reliability of measurement (Alwin, 2007; Schaeffer and Dykema, 2020).

In this document we first review the key results from prior research that address the question of measurement error and question context, focusing specifically on the differences in reliability of questions in a stand-alone, versus series, versus battery question contexts (Andrews, 1986). We then examine these issues in the GSS panel data with the aim of replicating the main findings from prior research. These analyses are supplemented by an examination of a broader set of issues identified in prior research, such as the role of introductions in series and batteries, and the placement of questions within a series or a battery. And finally, we investigate some of the sources of the differences in reliability of measurement of questions in different question contexts, and we analyze context differences while controlling for key features of question content and question form.

Questionnaire Architecture – Prior Research

Our past research found that a question's location in a questionnaire or its position within a series or battery has little or no effect on reliability (see, e.g., Alwin, 2007, pp. 172-177). Question context and questionnaire position interact to a slight degree, in that questions in batteries located later in a questionnaire are somewhat less reliable than those appearing earlier. Our analysis of these issues in the GSS (results not shown, but available on request) suggests that the length of a series or battery, and the position of a question within a unit has little bearing on estimated reliability.

Questions in series with long introductions (16+ words) appear to have lower reliability, whereas those in batteries having *any* introduction appear to have lower reliability (Alwin, 2007, pp. 177-179). Also, Andrews (1984, pp. 430-431) found higher measurement error for questions in longer batteries than those in shorter ones, although analyses by Scherpenzeel and Saris (1997) and Alwin (2007) did not support this finding. With respect to measurement reliability, our tentative conclusion is that *whether* a question is in a battery or topical series affects measurement reliability, rather than the length of a series/battery or the position of the question within a series/battery. Our analyses of the GSS data in part bear out these conclusions, but more research is necessary to ferret out the relationship between context and precision in measurement. In the GSS panels, however, those batteries with longer introductions have seemingly lower reliability levels, but as we will show later in this document, these particular batteries may also have other characteristics that contribute to this effect.

With regard to the differences in question context, Andrews' early work using a multitrait-multimethod (MTMM) design established a set of key findings that favored series relative to batteries with respect to reliability (Andrews, 1984). Results from our prior research provide some evidence that questions in a "topical series" are less reliable than "stand alone" questions (at least for factual material) and that non-factual questions in "batteries" are the least reliable (Alwin, 2007, pp. 171-172). We reproduce the relevant tables from that study here (see Tables 1 and 2). Table 1 presents the reliability estimates by categories of the cross-classification of question context and content using a collapsed set of categories for both variables. For purposes of this presentation, we arrange the data on question *content* into facts versus non-facts and arrange the data for question *context* into three categories: stand-alone questions, questions in series, and questions in batteries (ignoring for the moment the presence or absence of an introduction). The results here indicate that there are significant differences in the reliabilities of facts measured using stand-alone versus series formats. And among non-facts there are significant differences among the three formats, with those in batteries showing the lowest estimated reliability. These results suggest that net of question content, *stand alone* questions have the greatest level of reliability, followed by questions in series and batteries, with questions in batteries having the lowest level of reliability. These results are completely consistent with what was reported by Andrews (1984), wherein he found that, net of content, questions not in batteries had the lowest levels of measurement errors and questions in batteries containing 5 or more questions had the highest levels of measurement error. We return to our interpretation of these results after we consider a comparison of questions in series and questions in batteries.

Table 1. Comparison of reliability estimates by question content and question context

	Question Context					
Question Content	Alone	Series	Battery	Total	F-ratio[1]	p-value
Respondent self-report and proxy facts	**0.91** **(21)**	**0.77** **(56)**	0.76 (2)	0.81 (79)	16.26	0.000
Respondent self-report non-facts	0.66 (9)	**0.67** **(121)**	**0.61** **(217)**	0.63 (347)	12.96	0.000
Total	0.84	0.70	0.61	0.67		
Total n	(30)	(177)	(219)	(426)		
F-ratio[2]	27.18	15.00				
p-value	0.000	0.000				

[1] Test within facts excludes 2 battery fact triads; test within non-facts excludes the 9 stand-alone items.
[2] Test within batteries was not carried out.
Note: The number of questions on which reliability estimates are based is given in parentheses.
Source: Alwin (2007, p. 170).

Questions in Series versus Questions in Batteries

To examine the origin of the difference observed in estimated reliability among non-facts we also compare the differences between the reliabilities of questions in series and in batteries within categories of non-factual questions: beliefs, values, attitudes, self-assessments and self-perceptions. Results indicate that with few exceptions, the conclusion reached above can be generalized across categories of content of non-factual questions. Except for self-reported values, where we have relatively few measures in our sample of content, questions in batteries have significantly lower levels of reliability, controlling for the content of the question.

Table 2. Comparison of reliability estimates for questions in batteries and questions in series by type of non-factual questions

	Question Context				
Question Content	Series	Battery	Total	F-ratio	p-value
Beliefs	0.67	0.58	0.61	7.73	0.006
	(30)	(84)	(114)		
Values	0.62	0.69	0.67	2.30	0.137
	(12)	(29)	(41)		
Attitudes	0.74	0.65	0.66	3.38	0.070
	(11)	(64)	(75)		
Self-assessments	0.66	0.59	0.63	1.81	0.193
	(14)	(10)	(24)		
Self-perceptions	0.68	0.54	0.63	12.23	0.001
	(54)	(30)	(84)		
Total	0.67	0.61	0.63	12.96	0.000
Total n	(121)	(217)	(338)		

Note: The number of questions on which reliability estimates are based is given in parentheses.
Source: Alwin (2007, p. 171).

Our tentative conclusion regarding the effects of question context on measurement reliability, based on the *Margins of Error* study, then, is that net of content of questions (i.e. facts versus non-facts) stand-alone questions produce the highest level of reliability, questions in series have somewhat less reliability, and questions in batteries have the lowest relative reliability (see Alwin, 2007, pp. 167-180). These results cannot be completely generalized across content, in that factual questions are hardly ever included in batteries, and non-factual questions are relatively less often asked as stand-alone questions. Within the limitations of the data, however, it appears that as one adds *contextual similarity* to questions, reliability decreases. Moving, for example, from stand-alone questions to series where questions are homogeneous with respect to content more measurement errors appear to be produced, and as one moves to the situation of questions in batteries, where questions are homogeneous not only with respect to content but to response format as well, the estimated reliability is lowest. Thus, while placing questions in series and batteries increases the efficiency of questionnaire construction, it may reduce the quality of the data. If true, this has serious implications for the ways in which survey questionnaires are organized.

Regarding the production of measurement errors in batteries, perhaps the best explanation for these differences is one that Andrews (1984) provided, that the contextual similarity motivating researchers to group questions together also promotes measurement errors. The similarity of question content and response format may distract respondents from fully considering what information is being requested, making them less attentive to the specificity of questions. Thus, the "efficiency" features of the questionnaire architecture may generate measurement errors. In the case of batteries of questions, it appears that respondents may be more likely to "streamline" their answers when investigators "streamline" questionnaires (see Andrews, 1984, p. 431).

Finally, we already noted the finding that a question's location in a questionnaire or its position within a series or battery has little or no effect on reliability (results not presented here, but see Alwin, 2007, pp. 172-177). Question context and questionnaire position interact to a slight degree: questions in batteries located later in a questionnaire are somewhat less reliable than those appearing earlier. In addition, the length of introductions to both series and batteries seems to affect the reliability of questions – although somewhat differently in the two cases. Questions in series with long introductions (16+ words) appear to have lower reliability, whereas those in batteries having *any* introduction appear to have lower reliability (Alwin, 2007, pp. 177-179). Also, Andrews (1984, pp. 430-431) found higher measurement error for questions in longer batteries than those in shorter ones, although analyses by Scherpenzeel and Saris (1997) and Alwin (2007) did not support these findings. It thus appears that *whether* a question is in a battery or topical series affects measurement reliability, rather than the length of a series/battery or the position of the question within a series/battery. Further research is necessary to ferret out the relationship between question context and precisions in measurement. We consider these issues here.

Introductions to Series and Batteries

We find that in typical surveys, when batteries of questions are used, virtually all batteries include an introduction, whereas only one-third of series do. Our research indicates that the typical series introduction in these surveys was relatively short, around 16 words, whereas the typical battery introduction was somewhat more than twice that, at about 42 words. Introductions to series of questions tend to entail transitional sentences, e.g., "Now I have some questions about your education …" getting from one topic to another, whereas introductions to batteries tend to be introductions to the purpose of the task (Alwin et al., 2015). The optimal length of introductions to series and batteries has been a topic of research, which we do not address further here (but see Alwin, 2007; Andrews, 1984; Scherpenzeel and Saris, 1997; Saris and Gallhofer, 2007).

Question Context in the General Social Surveys

As noted in the foregoing, one of the great difficulties involved in assessing the effects of question context on reliability of measurement is the fact that it is not independent of question content. Virtually all factual questions appear either in a series or as stand-alone questions, and when they appear in a topical series, the series is rarely provided an introduction. The exact reverse pattern is apparent from these results for questions aimed at non-factual content, that is, most non-factual questions appear in batteries that have introductions (Alwin, Beattie, and Baumgartner, 2015). Because of this confounding of question content with question context, it is difficult to assign independent effects of the two sets of factors.

The GSS is an ideal study for the investigation of context effects because, not only are reliability estimates possible due to the longitudinal design, but there is a wide variety of content, and a large number of batteries, as well as plenty of series and stand-alone questions. The prior discussion raises the question of whether GSS survey questions are more or less reliable depending on whether they are introduced within the context of a topical series, including those series with identical response categories (i.e. batteries), or presented as a "stand alone" question, with no necessary topical similarity to questions before or after?

Question context is increasingly mentioned in recent studies as a factor in questionnaire construction (see, e.g., Schaefer and Dykema, 2020). This body of research provides some evidence that questions in a "topical series" are less reliable than "stand alone" questions in the measurement of factual material, and that for nonfactual questions series are more effective in terms of reliability than batteries. The literature on this subject to date indicates that questions in batteries tend to be less reliable than questions in series (Schaeffer and Dykema, 2020; Alwin, 2007, pp. 171-172; Andrews, 1984). This result reflects a convergence of findings between longitudinal research methods and MTMM models for the estimation of components of measurement error. Here we examine these issues using the GSS panels, which given the architecture of the GSS questionnaires presents a great opportunity for examining this set of issues.

Table 3. Comparison of reliability estimates by question content and context, by GSS panel

	Stand Alone	Series	Battery	Total	F Ratio	p-value
2006 GSS panel study						
Fact	0.890	0.834	---	0.845	2.397	0.131
	(7)	(28)		(35)		
Non-Fact	0.714	0.678	0.633	0.662	4.324	0.015
	(32)	(52)	(89)	(173)		
Total	0.746	0.733	0.633	0.692	13.193	0.000
	(39)	(80)	(89)	(208)		
F Ratio	13.229	31.273	---	52.092		
p-value	0.001	0.000		0.000		
2008 GSS panel study						
Fact	0.891	0.841	---	0.852	1.208	0.281
	(7)	(24)		(31)		
Non-Fact	0.688	0.660	0.633	0.651	1.727	0.181
	(32)	(51)	(88)	(171)		
Total	0.724	0.718	0.633	0.682	7.821	0.001
	(39)	(75)	(88)	(202)		
F Ratio	13.438	32.879	---	51.281		
p-value	0.001	0.000		0.000		
2010 GSS panel study						
Fact	0.896	0.849	---	0.860	1.535	0.225
	(7)	(24)		(31)		
Non-Fact	0.701	0.680	0.647	0.667	1.863	0.158
	(32)	(51)	(85)	(168)		
Total	0.736	0.734	0.647	0.697	8.258	0.000
	(39)	(75)	(85)	(199)		
F Ratio	13.454	27.706	---	49.595		
p-value	0.001	0.000		0.000		

Source: adapted from Alwin (2021)

Table 4. Comparison of reliability estimates by question content and context, combined GSS panels

	Stand Alone	Series	Battery	Total	F Ratio	p-value
Combined GSS panels						
Fact	0.892	0.841	---	0.852	5.183	0.025
	(21)	(76)		(97)		
Non-Fact	0.701	0.673	0.637	0.660	7.573	0.001
	(96)	(154)	(262)	(512)		
Total	0.735	0.728	0.637	0.691	28.951	0.000
	(117)	(230)	(262)	(609)		
F Ratio	41.185	92.901	---	153.607		
p-value	0.000	0.000		0.000		

In examining the effects of series and/or batteries on the quality of measurement, due to the high level of confounding between content and context, an effort to separate the effects of context and content is an important goal. As previously mentioned, facts are measured either as stand alone or in series; there are no facts measured using batteries. Non-facts on the other hand are measured in all three forms: stand alone, series and batteries. In the typical GSS questionnaire, there are 12 topical series in each panel study, and another 17 batteries per panel, containing collectively 262 total items in batteries in the three GSS panels, compared to 230 questions in series. In the GSS there is a preponderance of long batteries, but in fact series are on average longer. Thus, virtually all factual questions appear either in a series or as stand-alone questions, and when they appear in a topical series, the series is rarely provided an introduction. A reverse pattern is apparent for GSS questions aimed at non-factual content, that is, most non-factual questions appear in batteries that have introductions. Because of this confounding of question content with question context, we examine the effects of contexts on facts and non-facts separately in an effort to assign independent effects of the two sets of factors. And because of the confounding of question context with question form – open versus closed-form questions – we also control statistically for an independent role for question context, net of question form as well.

Context Effects Explained

With respect to survey context, our assessment of the GSS panel data to date is that we tend to observe a difference between the reliabilities of stand-alone factual questions and those appearing in series and batteries (see Tables 3 and 4 above). By contrast, for non-factual questions, there is a significant pattern that favors stand-alone questions, followed in level of reliability by questions in series, and then questions in batteries. These results replicate the prior findings of Andrews (1984) and Alwin (2007), although the magnitudes of the differences are very small—nine percent of the variance in reliability. In addition, we found that there were no differences in the reliabilities of questions in series and batteries of differing lengths, no differences in the reliabilities of questions in differing positions within a series or a battery, and no differences in the reliability of questions in series and batteries that used introductions of differing lengths (Alwin et al., 2015).

Using the rich set of data from the GSS panels, Table 5 presents an examination of the hypotheses advanced in previous research that such context effects can be explained away by considering the question content and question form correlates of question context. In other words, it is possible to examine the context effects identified above, net of question content and question form, allowing us to assess the extent to which the context effect is an artifact of these correlates. Table 5 presents a series of regression equations that include predictors that we consider affecting levels of reliability. The first equation presents the regression of reliability on the question context, as operationalized here as a set of dummies representing stand-alone, series and battery context (series is the omitted category). These factors account for nearly nine percent of the variation in reliability.

These results (Table 5) reproduce the findings articulated above that favor stand-alone questions over questions in series, and an advantage of the latter over questions that are a part of a battery. The second equation includes the distinction between series and batteries with and without introductions (note the omitted category in this regression is ‘series without an introduction’) and reveals that batteries with introductions have a slight advantage over those without. In this second equation, there appears to be no difference between series (either with or without an introduction) and stand-alone questions with respect to reliability of measurement.

Earlier we reported that there were significant differences among the three types of contextual formats, especially when we control for content by selection (i.e., fact vs. non-fact). Providing attention to this issue, the third equation in this series of regressions (Table 5) includes a measure of question content, i.e., facts vs. non-facts, to attempt to explain away the context effect. By adding question content to the model, the prediction of variance in reliability more than doubles—increasing the R^2 from .089 to .233—reinforcing the conclusion reached in most studies that facts can be more reliably measured than non-facts. By controlling for question content, as model 3 demonstrates, there remains a context effect, net of question content, but it is lessened. These results indicate a salutary effect of introductions of series on reliability. Series with introductions presumably foster greater measurement precision. By controlling for question content (fact vs. non-fact), then, reveals an advantage to stand-alone and questions in series with introductions and no differences for batteries of either type. In other words, we have shown that control for question content eliminates the apparent disadvantages of questions in batteries.

As one additional consideration, we speculated that there might be an enhancement for factual questions if they were not in series but were stand-alone questions. In equation 4 we consider this issue by allowing for an interaction effect between content and context (i.e. stand alone * facts). This factor does not attain statistical significance, and we drop it from further consideration. The stand alone effect is apparently not accounted for by the preponderance of facts in this contextual category.

Finally, we consider whether the context effects are further reduced, or changed, if in addition to question content, we also control for question form, operationalized here as the distinction between open versus closed-form questions. The equation in model 5 of Table 5 includes this factor, which shows a (non-significant) effect of question form, with open-ended questions having the advantage with respect to measurement reliability. Due to the high correlation between question content and form, the fact versus non-fact difference is reduced slightly, but question content continues to be one of the most important factors in the prediction of variation in reliability. As noted earlier, it is worthy of mention that the components of questions – content and context – account for nearly one-quarter of the variance in reliability—but this is changed little by the further consideration of question form[2].

[2] The consideration of an interaction between content and form (i.e. facts * open-ended form) does not enhance our ability to predict variation in reliability.

Table 5. Regression of GSS reliability estimates on attributes of questions: GSS panel studies

	Model									
Predictors	1		2		3		4		5	
Intercept[1]	0.728	***	0.731	***	0.656	***	0.657	***	0.655	***
Stand alone (SA)	0.007		0.004		0.046	**	0.044	*	0.046	**
Series with an introduction			-0.015		0.049	*	0.048	*	0.053	*
Battery	-0.091	***								
Battery with an introduction			-0.091	***	-0.016		-0.017		-0.015	
Battery without an introduction			-0.128	**	-0.053		-0.054		-0.051	
Facts vs nonfact (fact = 1)					0.184	***	0.182	***	0.149	***
SA*Facts							0.010			
Open-ended question									0.047	
Open-ended * Facts										
R^2	0.087		0.089		0.233		0.233		0.237	
N of cases	609		609		609		609		609	

[1] Omitted categories are: in Model 1, series; in Model 2, series without an introduction; in Models 3 - 4, series without an introduction, non-facts; in Model 5, series without an introduction, non facts, closed questions.

Key: + p < .1, * p < .05, ** p < .01, *** p < .001

Conclusions

We began with the issue of whether the questionnaire organizational context in which the question is placed can potentially affect the reliability of measurement. Prior literature identified an apparent hierarchy of question contexts—stand-alone questions are the most reliable, questions in topical series are somewhat less reliable, and questions placed in batteries have the lowest question-specific reliabilities. In this document we reviewed the key results from prior research that address the question of measurement error and question context, focusing specifically on the differences in reliability of questions in a stand-alone, versus series, versus battery (SASB) question contexts. We then examined these issues in the GSS panel data with the aim of replicating the main findings from prior research. These analyses are supplemented by an examination of a broader set of issues identified in prior research, such as the role of introductions in batteries, and the placement of questions within a series or a battery. And finally, we investigated some of the sources of the differences in reliability of measurement of questions in different question contexts.

Past researchers have reasoned that batteries are the least reliable because they streamline the questionnaire, and due to this configuration, respondents tend to streamline their answers (Andrews, 1986). Past research also suggested that context effects of the type identified may be explained by the association between context and question content, and/or by the association between context and question form. In the above presentation we were able to show that the original findings re-appear in the GSS panels, and that when we control for question content and question form, the initial effects of question context are all but removed. The advantages of stand-alone questions and series with introductions are registered by slightly higher levels of reliability, but we find that the deficits in measurement reliability associated with placement in batteries is accounted for by a consideration of the content and form of the questions. These results lend support for the view that context effects are an artifact of the content of the questions and the nature of the question forms involved, in this case whether the questions are open-ended or closed form. At the same time, we believe the exercise of care in the construction of batteries, so that respondents consider each question independently of others, appears to be highly desirable.

Finally, although we have accounted for some of the previously documented differences in reliability of questions in different contexts, due to correlates with question content and form, there are additional factors that may be relevant. Future attention to other aspects of question form, specifically the number of response categories for closed-form questions, is strongly warranted. And, due to the fact that the effects of numbers of response categories depends upon whether the question measures bipolar or unipolar dimensions, the *polarity* of the response scale (unipolar vs. bipolar) and its interactions with the number of response categories is also relevant. In a related paper, we examine these aspects of question form and reliability and whether the context effects are further removed, once these aspects of question form are considered. Once these attributes of question form – number of response categories offered and polarity of the scale – are controlled, there are no longer effects of question context, as we have hypothesized here. [3]

[3] See RSM Results Series, Duane F. Alwin and Paula A. Tufiş, "Question Form and Reliability of Measurement."

References

Alwin, Duane F. 1989. Problems in the Estimation and Interpretation of the Reliability of Survey Data. *Quality and Quantity* 23:277-331.

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., 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., 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., 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.

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

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