An experimental test of the effectiveness of cognitive interviewing in pretesting questionnaires

Lenzner, Timo; Hadler, Patricia; Neuert, Cornelia

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 491156185 / Funded by the German Research Foundation (DFG) - Project number 491156185

Empfohlene Zitierung / Suggested Citation:
Lenzner, T., Hadler, P., & Neuert, C. (2022). An experimental test of the effectiveness of cognitive interviewing in pretesting questionnaires. *Quality & Quantity*, *57*(3), 3199-3217. https://doi.org/10.1007/s11135-022-01489-4

Quality & Quantity (2023) 57:3199–3217
https://doi.org/10.1007/s11135-022-01489-4

An experimental test of the effectiveness of cognitive interviewing in pretesting questionnaires

Timo Lenzner[1] · **Patricia Hadler**[1] · **Cornelia Neuert**[1]

Accepted: 20 May 2022 / Published online: 27 August 2022

Abstract
Pretesting survey questions via cognitive interviewing is based on the assumptions that the problems identified by the method truly exist in a later survey and that question revisions based on cognitive interviewing findings produce higher-quality data than the original questions. In this study, we empirically tested these assumptions in a web survey experiment ($n=2{,}200$). Respondents received one of two versions of a question on self-reported financial knowledge: either the original draft version, which was pretested in ten cognitive interviews, or a revised version, which was modified based on the results of the cognitive interviews. We examined whether the cognitive interviewing findings predicted problems encountered in the web survey and whether the revised question version was associated with higher content-related and criterion-related validity than the draft version. The results show that cognitive interviewing is effective in identifying real question problems, but not necessarily in fixing survey questions and improving data quality. Overall, our findings point to the importance of using iterative pretesting designs, that is, carrying out multiple rounds of cognitive interviews and also testing the revisions to ensure that they are indeed of higher quality than the draft questions.

Keywords Cognitive interviewing · Criterion-related validity · Data quality · Experiment · Pretesting

✉ Timo Lenzner
timo.lenzner@gesis.org

Patricia Hadler
patricia.hadler@gesis.org

Cornelia Neuert
cornelia.neuert@gesis.org

[1] GESIS – Leibniz Institute for the Social Sciences, Survey Design and Methodology, P.O. Box 12 21 55, 68072 Mannheim, Germany

1 Introduction

Surveys are one of the main methods for gathering information about people's beliefs, values, attitudes, behaviors, and states of affairs (e.g., Schuman and Presser 1981). Many important decisions are made based on survey data; hence it is essential that they are of high quality (i.e., valid, reliable, and unbiased). To accomplish this, survey designers must develop questions that respondents will easily and consistently understand in the intended way (Collins 2003; Fowler 1995). For one, questions that are easily understood and do not challenge respondents' cognitive capacities introduce less measurement error than questions that are difficult to understand (Groves et al. 2004). Secondly, ambiguous questions, which are understood differently by different respondents or interpreted differently than intended, do not measure the concepts they are supposed to measure, or they do not measure them properly (Fowler 1992). For example, if two groups of respondents interpret a question differently, the observed differences in the data are not true differences but an artifact of the varying interpretations. To avoid this situation, survey researchers must pretest their questions before collecting data and make sure that they function as intended. An appraisal of a questionnaire by researchers at the desk is generally not sufficient to ensure high-quality questions. As Sudman and Bradburn (1982, p. 283) point out: "[e]ven after years of experience, no expert can write a perfect questionnaire." Empirical pretests are needed to verify that the questions yield reliable and valid responses.

A variety of methods are available for pretesting survey questions, one of which is cognitive interviewing. The goal of cognitive interviewing is to gain insights into the cognitive processes underlying survey responding (Tourangeau et al. 2000): (1) how do respondents interpret questions?, (2) how do they retrieve relevant information for answering questions from memory?, (3) how do they arrive at a judgement about what to answer?, and (4) how do they map their "internally" determined answer to the response format provided? This information is used to determine whether survey questions measure the concepts they are supposed to measure and whether respondents have difficulties understanding or answering them. Moreover, cognitive interview data are supposed to provide information about the causes of question problems and point out ways in which these can be remedied (Beatty and Willis 2007; Miller 2011).

Even though cognitive interviewing is an established questionnaire pretesting method, there is limited empirical work examining its effectiveness in identifying real question problems (i.e., that indeed undermine data quality) and helping to improve the quality of survey questions. The purpose of this article is to extend the current state of research on the effectiveness of cognitive interviewing in pretesting questionnaires. After shortly reviewing contemporary cognitive interviewing practices and discussing previous research on the method's effectiveness, we report on an experimental study that examined whether question problems identified by cognitive interviewing are observable in a later survey and whether a revision suggested by cognitive interviewing findings produces higher-quality data than the original question. We close with a discussion of the practical implications of our study and suggest perspectives for future research.

2 Background

2.1 Cognitive interviewing practice

Classically, cognitive interviews are carried out face-to-face with small purposive or quota samples of five to 30 respondents (Willis 2005). Interviewers usually conduct these using (semi-)standardized interview protocols that include the questions to be tested, the aims of testing, and the specific cognitive techniques to be used. The techniques most commonly used in cognitive interviews are *thinking aloud* and *probing*. The think-aloud technique involves asking respondents to verbalize their thought processes as they answer a question. Probing refers to asking follow-up questions (probes), either immediately after administering the individual questions (concurrent probing) or after respondents completed the whole questionnaire (retrospective probing). Depending on the aims of testing, probes can, for instance, focus on specific cognitive processes (e.g., comprehension probe: "What do you understand by a 'representative democracy' in this question?"; information retrieval probe: "How did you remember that you purchased 10 books in the past 6 months?") or ask participants to elaborate on their answer (category-selection probe: "Can you explain why you chose this answer?"; Prüfer and Rexroth 2005; Willis 2005).

Even though it seems that practitioners agree about the basic principles and aims of cognitive interviewing (Boeije and Willis 2013; Collins 2015; Lenzner et al. 2016), there is no such thing as 'the' cognitive interview, but considerable variation in practitioners' approaches (Beatty and Willis 2007; Conrad and Blair 2004; Willis 2005). For example, current practices differ regarding the standardization of the interview protocol used in the interviews (scripted vs. semi-scripted vs. improvised), the number of interviews being conducted (five vs. 30 and more), the number of testing rounds carried out (one round vs. iterative testing in multiple rounds), the number of cognitive interviewers (one conducting all interviews vs. several interviewers) and their experience in conducting cognitive interviews ("unskilled data collector" vs. "expert investigator"; Beatty and Willis 2007), the selection of participants (e.g., recruiting monolingual vs. bilingual speakers when testing translated questionnaires; Park et al. 2016), and the ways in which the verbal data are analyzed (based on interview transcripts vs. on interviewer notes). What is common to all approaches is the assumption that cognitive interviewing yields insights into how respondents understand and answer survey questions and that these insights help to detect question problems and hint at possible solutions to these problems.

2.2 Sources of error in cognitive interviews

While cognitive interviewing may indeed step up to these expectations, researchers have pointed out that the assumption that problems revealed by cognitive interviews necessarily undermine data quality may be too optimistic (Conrad and Blair 2009; Willis 2005; Yan et al. 2012). There are several possibilities for error in cognitive interviews. First, they could detect problems which are not actually present (*false alarms*). This might be due to *reactivity bias*, that is, respondents reporting different thought processes just by virtue of being asked to articulate them (Conrad and Blair 2009). For example, if an interviewer probes for the interpretation of a specific term in a question, this could lead the respondent to think about potential ways in which the question could be (mis)interpreted, even though he or

she might have perceived it as clear and unambiguous. And, due to the additional cognitive demands related to answering probes and explaining answers, respondents in cognitive interviews may answer questions differently than later survey respondents (Willis 2015). Second, cognitive interviews could fail to detect problems that really exist (*misses*), for example because too few interviews were being conducted to uncover all problems (Blair and Conrad 2011). Third, in interpreting the verbal data generated during cognitive interviewing, different researchers may come to different conclusions about whether a problem exists or not (Rothgeb et al. 2007), particularly if respondents provide vague reports or have problems expressing themselves verbally (Conrad and Blair 2009). All these issues raise concerns about the method's consistency and effectiveness.

2.3 Previous research on the effectiveness of cognitive interviewing

To date, few studies have been conducted on whether problems identified by cognitive interviewing truly exist in a later survey and whether question revisions based on cognitive interviewing findings produce data of superior quality in comparison to the original questions (e.g., Forsyth et al. 2004; Willis and Schechter 1997). For example, Willis and Schechter (1997) revised five questions that had been tested in cognitive interviews and made differential predictions on the response distributions produced by the original and revised versions in later surveys. For four of the five questions, the predictions based on cognitive interviewing were supported in three different survey settings. While this study lent support to the notion that cognitive interviewing results are also observable in later surveys, the data did not allow for demonstrating that the question revisions produce higher-quality data than the original ones.

Forsyth and colleagues (2004) conducted a similar study, in which they additionally looked at improvements in data quality resulting from pretesting questions. In their study, a set of 12 survey questions was pretested by means of expert reviews, cognitive appraisal systems, and cognitive interviewing and revised based on the results. Afterwards, both the original questions and their revised counterparts were included in a telephone survey using a split-sample experiment. Using item nonresponse rates, behavior coding results, and interviewer ratings as indicators of question quality, the authors found that the original versions identified as very problematic during pretesting were also associated with problems in the telephone survey (e.g., questions classified as having recall and sensitivity problems had higher nonresponse rates). With regard to the revised questions, they obtained mixed results: on the one hand, the revisions were associated with significantly fewer respondent problems (as rated by the interviewers) as well as nonsignificant reductions in item nonresponse and problematic behavior codes. On the other hand, the interviewers rated the revisions as having *more* interviewer problems (e.g., reading problems) than the original questions. Moreover, given that different pretesting methods were applied to evaluate the questions, it is difficult to assess what contribution cognitive interviewing made to the revisions. And finally, the data quality indicators used in this study were relatively indirect and (on part of the interviewer ratings) subjective measures of data quality. All in all, there is limited empirical evidence on the effectiveness of cognitive interviewing in identifying real question problems and improving the quality of survey questions.

One could argue that expecting cognitive interviewing to directly improve survey questions is overly demanding of the method. As Willis (2005, p. 214) points out, "cognitive test-

ing doesn't improve survey questions, questionnaire designers improve survey questions." Thus, it might be more appropriate to focus solely on the method's ability to detect question problems when evaluating its effectiveness. However, we believe that a questionnaire pretesting method that identifies question problems but fails to provide guidance on how to fix them is of limited value. Ultimately, one major function of pretesting is to optimize questions before they are used in a later survey (cf. Forsyth et al. 2004). Even though cognitive interviewing cannot be expected to automatically repair questions, it should provide clear indications of how they can be improved.

3 Research questions

Following Willis and Schechter (1997) and Forsyth et al. (2004), we believe that two things must be shown to demonstrate the effectiveness of cognitive interviewing as a pretesting method. First, the problems diagnosed by cognitive interviewing must also be identified in a later survey, and second, questions that are revised based on cognitive interviewing results should yield higher-quality data than the original questions. Hence, we address the following two research questions in the current study:

> Research question 1: Are the problems identified by cognitive interviewing observable in a later survey?
> Research question 2: Are question revisions based on cognitive interviewing findings of higher quality than the original draft questions?

To address these questions, we selected a survey question on self-reported financial knowledge that had been evaluated via cognitive interviewing at GESIS - Leibniz Institute for the Social Sciences, Germany (Lenzner et al. 2020) and tested the original and revised versions of the question in a web survey experiment. In contrast to the earlier studies described above, we examined the content-related and criterion-related validity (see Kane 2006) of both question versions as more direct measures of data quality.

4 Method

Both the cognitive interviewing study and the web survey experiment were conducted in German. The question formulations presented below are English translations of the original German wordings, which are documented in the Appendix.

In the following, we first describe the procedures used in the cognitive interviewing study and present the original and revised version of the self-reported financial knowledge question implemented in the later web survey experiment. We then outline the design of the experiment and present our hypotheses. Finally, we describe our analytical approach to assessing the content-related and criterion-related validity of both question versions and outline the web survey data collection procedures.

4.1 Cognitive interviewing study

4.1.1 Procedures

In preparation for wave 9 of the German sub-study of the Survey of Health, Ageing and Retirement in Europe (SHARE), we evaluated 19 newly developed items on financial decision making, successful ageing, and other topics in a cognitive interviewing pretest (Lenzner et al. 2020). A total of ten cognitive interviews were carried out in January and February 2020 at GESIS - Leibniz Institute for the Social Sciences, Germany. The participants were recruited from the respondent pool maintained by the institute and paid 30 € for participating. Given that the SHARE surveys are targeted at respondents aged 50 years or older, only participants of these ages were recruited for the cognitive interviews. Half of the participants were female, and similarly, half had a university-entrance degree ($n=5$, respectively). Three were 50 to 59 years old, four were 60 to 69 years old, and another three were 70 years or older.

The interviews were conducted by three experienced cognitive interviewers based on an interview protocol containing pre-scripted probes, such as "How did you arrive at your answer?" or "What does the term 'financial knowledge' mean to you in this question?" (see Appendix in Lenzner et al. 2020 for the interview protocol.). The probes were administered concurrently, that is, directly after respondents had answered the survey questions. The interviewers were encouraged to apply additional probing if they deemed it necessary and respondents frequently commented spontaneously on the items prior to the administration of any probe.

All participants gave their written consent for the video recording of the interviews. The interviews lasted between 43 and 69 min with an average interview length of 55 min ($SD=8.3$). Prior to the analysis, the interviews were written up using an interview notes template, which included the participants' answers to the questions tested, spontaneous comments made by the participants (without being probed for), their answers to the probes, and observations or remarks by the interviewers (see D'Ardenne and Collins 2015). The data used in the current study (i.e., the answers to the probes on the self-reported financial knowledge question) were analyzed by two researchers working independently as follows: first, they openly coded respondents' answers to the probes with regard to the kinds of information they provided. Second, they organized these codes into larger categories and specified the core themes and types of problems that emerged from the analysis. Finally, they developed draft revisions for the question. The researchers then met with two other researchers to discuss the findings and to make a final decision about the recommendations for revision.

4.1.2 Question tested

The question on respondents' self-reported financial knowledge that was tested in the cognitive interviewing study read:

> "On a scale from 1 to 7, where 1 means very low and 7 means very high, how would you assess your overall financial knowledge?" (original version)

The main aim of the cognitive interviews was to examine how respondents interpreted the term 'financial knowledge' in this question. We found that its interpretation varied between respondents and that the respective interpretation had a systematic effect on the answers: respondents who associated the term mainly with knowledge about complex financial processes (e.g., stock market performance) selected lower scale values (and thus estimated their financial knowledge lower) than respondents who associated the term mainly with knowledge about managing one's personal finances (e.g., budgeting). If this was also the case in the later SHARE surveys, the observed differences in the data would not be true differences but an artifact of the different question interpretations. Therefore, we recommended clarifying the construct the question is intended to measure by adding a definition of the term 'financial knowledge'. According to the question designers, the term was meant to encompass both knowledge about (complex) financial issues and, in particular, knowledge about managing one's personal finances. Incorporating the question designers' definition of financial knowledge into the question, our suggested revision read as follows:

> "On a scale from 1 to 7, where 1 means very low and 7 means very high, how would you assess your financial knowledge? Financial knowledge here means the understanding of financial issues and the ability to make appropriate and informed decisions about personal finances, such as budgeting, investments, insurance, real estate, debt management or tax planning." (revised version).

4.2 Web survey experiment

We conducted an online experiment to examine whether we could replicate the findings from the ten cognitive interviews in a larger survey sample and whether the revised question version yielded higher-quality data than the original version.

4.2.1 Design and hypotheses

Respondents were randomly assigned to one of two versions of the target question on self-reported financial knowledge: (1) the original draft version or (2) the revision based on the ten cognitive interviews. The questions were individually presented with an endpoint-labeled, seven-point, horizontally aligned rating scale ranging from "1 – very low" to "7 – very high". The rating scale additionally offered a "don't know" response option.

Based on the findings from the ten cognitive interviews described above, we propose two hypotheses. First, respondents receiving the original version vary in their interpretation of the term 'financial knowledge', associating it either (primarily) with knowledge about complex financial processes, (primarily) with knowledge about handling one's own finances, or with knowledge about both facets (H1). Second, in the original version condition, respondents interpreting financial knowledge to refer (primarily) to knowledge about complex financial processes will rate their financial knowledge lower than those understanding it to refer (primarily) to knowledge about managing one's personal finances, with those thinking in equal parts of both aspects in between (H2).

Regarding our second research question, the comparison between the original and the revised version of the question, we propose four additional hypotheses. First, more respon-

dents in the revision than in the original condition interpret the question as intended, that is, as referring to both facets of financial knowledge, because the revision includes an analogues definition of financial knowledge (H3). As the definition gives considerable weight to the facet of handling personal finances (in terms of the number of words devoted to it), more respondents in the revision than in the original condition interpret the question as referring to this aspect of financial knowledge, even though the intended interpretation would be to include both facets (H4). According to our second hypothesis, this should in turn lead respondents in the revision condition to rate their financial knowledge higher than respondents in the original version (H5). Finally, we expect the revision to be the better measure of subjective financial knowledge, and thus to be associated with higher criterion-related validity than the original version (H6). Specifically, we expect the revised version to correlate more strongly with two conceptually related criterion variables than the original version (H6a and H6b). The reasons underlying the selection of the criterion measures and the associations we expected between them and the two financial knowledge question versions are explained in more detail in the next section. A summary of our research hypotheses is provided in Table 1.

4.2.2 Probing and criterion questions

To examine what aspects of finances respondents had in mind when answering the target question on financial knowledge in the web survey experiment, and thereby to determine the content-related validity of both question versions, we asked a semi-open probing question immediately after the target question. The semi-open probe read:

> "The previous question was about how you assess your financial knowledge. What did you consider when answering the question? (1) How knowledgeable I am about financial topics (stock markets, capital markets, mutual fonds, etc.), (2) How knowledgeable I am about topics that affect my personal finances (budgeting, investing, managing debt, etc.), (3) Both of the above-mentioned topics, (4) Something different, namely: [open text field])".

The response options were derived from the cognitive interview findings, that is, these were the different aspects participants in the cognitive interviews had mentioned when being probed on their interpretation of the term 'financial knowledge'. A total of 82 respondents selected the fourth answer category, of which 44 answers were uninterpretable ("don't know", "xX", "none of your business") and 38 answers related to one of the other three answer options. Prior to the analyses, these answers were either defined as missing values or recoded into one of the three other answer options, respectively.

To gather criterion-related validity evidence for both question versions, we adopted a method that has been used in several previous publications to compare different question versions (e.g., Chang and Krosnick 2003; Höhne and Yan 2020; Shaeffer et al. 2005; Yeager and Krosnick 2012). To apply this method, the web survey included two variables (see Appendix) that were conceptually related to financial knowledge, namely financial behavior and life satisfaction. Both criterion variables correlated significantly with the experimentally manipulated target question in the full sample (financial behavior score: $r = .25$; $p < .001$; life satisfaction: $r = .28$; $p < .001$). Criterion-related validity evidence was determined by inves-

Table 1 Research questions and hypotheses.

Research questions (RQ)	Hypotheses (H)
RQ1: Are the problems identified by cognitive interviewing observable in a later survey?	H1: Respondents receiving the original version interpret the term 'financial knowledge' in different ways (either referring to knowledge about complex financial processes, about managing personal finances, or about both). H2: In the original version condition, respondents interpreting financial knowledge to refer to complex financial processes rate their knowledge lower than those understanding it to refer to managing one's personal finances, with those thinking about both aspects in between.
RQ2: Are question revisions based on cognitive interviewing findings of higher quality than the original draft questions?	H3: More respondents in the revision than in the original condition interpret the question as intended (i.e., as referring to both knowledge about managing personal finances and about complex financial processes). H4: More respondents in the revision than in the original condition interpret the term 'financial knowledge' to refer to managing one's personal finances. H5: Respondents in the revision condition rate their financial knowledge higher than those in the original condition. H6: The revision is associated with higher criterion-related validity than the original version. H6a: The positive correlation between financial knowledge and financial behavior is stronger in the revision condition. H6b: The positive correlation between financial knowledge and life satisfaction is stronger in the revision condition.

tigating which question version correlated more strongly with the two criterion measures. This approach is based on the notion that measurement error weakens associations between target and criterion questions, and hence, if one question version elicits stronger associations than the other, this would indicate that the first version results in more valid measurements than the second (see Shaeffer et al. 2005). The selection of the criterion variables was based on the following considerations:

Financial behavior Research has shown that people with higher levels of subjective financial knowledge display more sound financial behaviors than people with lower financial knowledge (Lind et al. 2020; Robb and Woodyard 2011). We implemented a three-item measure of financial behavior in the web survey, which was adapted from the 2016 OECD/INFE International Survey of Adult Financial Literacy Competencies (OECD 2016). Prior to the analyses, the items were recoded so that higher values indicated higher agreement with the items. To create a score for financial behavior, a principal components factor analysis was conducted on the three items with oblique rotation (Direct Oblimin). The Kaiser-Meyer-Olkin measure of sampling adequacy (Kaiser 1970) was 0.62, exceeding the recommended value of 0.6, and Bartlett's test of sphericity (Bartlett 1954) reached statistical significance ($\chi 2(3) = 837.87$, $p < .001$). The analysis identified one factor with an Eigenvalue greater than 1, explaining 57.78% of the variance. As shown in Table 2, all items loaded above 0.70. Consequently, we used Bartlett factor scores as a measure of financial behavior in the criterion-related validity analyses. Assuming that the revised question version is the better measure of subjective financial knowledge, we expect it to correlate positively and higher with the financial behavior score than the original version (H6a).

Life satisfaction High financial knowledge, or more broadly, high financial literacy (Remund 2010) is associated with a wide range of positive outcomes, in particular higher

Table 2 Principal components factor analysis results of the financial behavior items

Item	Factor Loadings	*M*	*SD*
1 Before I buy something, I carefully consider whether I can afford it.	0.74	4.22	0.86
2 I keep a close personal watch on my financial affairs.	0.82	4.20	0.87
3 I set long term financial goals and strive to achieve them.	0.72	3.65	1.05
Eigenvalue	1.73		
Variance explained (%)	57.78		

Note: $N=2159$. Recoded answer scale: 1 = strongly disagree, 2 = rather disagree, 3 = neither/nor, 4 = rather agree, 5 = strongly agree

financial satisfaction and lower levels of anxiety about life (Goyal and Kumar 2021). Recent studies have examined the relationship between financial literacy and broader measures of subjective well-being, such as life satisfaction, finding that increased financial satisfaction is associated with higher life satisfaction (Falahati et al. 2012). To measure life satisfaction, we used the short scale L-1 developed by Beierlein and colleagues (2015). The L-1 contains only one item with 11 answer categories ranging from "1 – not at all satisfied" to "11 – extremely satisfied" (see Appendix). Assuming that the revised question version is the better measure of subjective financial knowledge, we expect it to correlate positively and higher with respondents' life satisfaction than the original version (H6b).

4.2.3 Sample and data collection

The experiment we report in this article was implemented in a web survey that was fielded in late November and early December 2020. It was part of a larger study with several unrelated experiments, which were independently randomized to reduce the possibility of systematic carry over effects. Average completion time for the web survey was 18.5 min and the current experiment was implemented in the middle of the survey.

Respondents were recruited from a German nonprobability online panel using quotas for gender, age, and education. Of the 2,441 who started the survey, 241 broke off before completing it, leaving 2,200 respondents for statistical analyses ($n=1{,}110$ in the original condition, $n=1{,}090$ in the revision condition). These were between 18 and 82 years of age with a mean age of 45 ($SD=14.9$). 49% were female and 26.5% had graduated from a lower secondary school, 31.0% from an intermediate secondary school, and 40.8% from a college preparatory secondary school. Further, 1.4% still attended school or had finished without a diploma and 0.3% did not report their highest level of education. To evaluate the effectiveness of random assignment and the sample composition between the two experimental groups, we conducted chi-square tests. The results showed no significant differences regarding age, gender, and educational attainment.

5 Results

In our analyses, we first look at how well the cognitive interviewing results predict respondents' interpretations of and their responses to the target question in the web survey (RQ1). To this end, we examine how respondents in the original condition interpreted the term 'financial knowledge' and whether different interpretations led to different responses. We

then turn to the revised question version and investigate its quality in comparison to the original version (RQ2). To do so, we examine if respondents in the revision condition are more likely to interpret the term as intended (i.e., as referring to both knowledge about complex financial processes and about managing one's personal finances), whether they answer the target question differently than those receiving the original version, and whether the revised version is associated with higher criterion-related validity than the original one.

5.1 Research question 1: predictions based on cognitive interviewing results

Supporting our first hypothesis and replicating the cognitive interviewing results, respondents answering the original question version interpreted its meaning differently. While 26.2% (95% CI [23.6, 28.8]) of the respondents understood it to refer (primarily) to knowledge about complex financial processes, 37.8% [34.9, 40.7] interpreted the question as referring (primarily) to knowledge about managing one's personal finances, and 33.9% [31.1, 36.7] as referring to both facets of financial knowledge.

Next, we examined whether the different interpretations had a systematic effect on respondents' answers. In the original condition, respondents who thought primarily about complex financial processes rated their financial knowledge lower ($M=3.74$, $SD=1.45$, $N=271$) than those who thought about personal finances ($M=4.52$, $SD=1.21$, $N=412$) or both of these aspects ($M=4.50$, $SD=1.45$, $N=354$). An ANOVA with Bonferroni corrected post-hoc tests revealed that the differences between the first group and the other two groups were statistically significant, respectively (Welch's $F(2, 619.30)=29.65$, $p<.001$). This finding is in line with our second hypothesis, suggesting that respondents' interpretation of the term 'financial knowledge' influenced how they rated their self-reported financial knowledge.

5.2 Research question 2: quality of both question versions

Regarding the comparison of the original and the revised question versions, we first examined whether the interpretation of the target question varied between the two experimental conditions. As shown in the top panel of Table 3, the answers to the closed probe differed significantly between both conditions ($\chi^2(2)=8.44$, $p=.015$). However, Bonferroni corrected post-hoc tests revealed that this was not due to differences in the share of respondents interpreting the question as intended (i.e., reporting having thought about both facets of financial knowledge; $\chi^2(1)=0.40$, $p>.05$). Hence, we found no support for our third hypothesis. Instead, significantly more respondents in the revision condition claimed to have thought (primarily) about how well they manage their personal finances than did respondents in the original condition ($\chi^2(1)=6.99$, $p=.008$). This finding is in line with our fourth hypothesis.

Next, we examined whether the original and revised versions of the question resulted in different response distributions. These are shown in the lower panel of Table 3. A chi-square test revealed that the response distributions differed significantly from each other ($\chi^2(6)=13.74$, $p=.033$). Supporting our fifth hypothesis, an inspection of the mean ratings showed that respondents receiving the revised version rated their financial knowledge significantly higher ($M=4.42$, $SD=1.46$; $t(2080) = -2.16$, $p=.031$) than those receiving the original version ($M=4.28$, $SD=1.41$). However, these differences were very small with only 0.14 points on a seven-point scale.

Table 3 Response distribution of the closed probe and the target question on self-reported financial knowledge by experimental condition (original or revision).

Response distribution	Original		Revision		Significance level
	n	%	n	%	
Closed probe					
Complex processes	291	26.2	240	22.0	$\chi^2(2)=8.44$, $p=.015$
Personal finances	420	37.8	473	43.4	
Both	376	33.9	356	32.7	
Missing	*23*	*2.1*	*21*	*19*	
Target question					
7 high	43	3.9	53	4.9	$\chi^2(6)=13.74$, $p=.033$
6	150	13.5	201	18.4	
5	321	28.9	284	26.1	
4	264	23.8	229	21.0	
3	147	13.2	138	12.7	
2	89	8.0	86	7.9	
1 low	42	3.8	35	3.2	
DK	*54*	*4.9*	*64*	*5.9*	

Note: $N=2{,}200$ (original condition: $n=1{,}110$; revision condition: $n=1{,}090$).

Table 4 Means, standard deviations, and intercorrelations for criterion questions and predictor variables

	M	*SD*	1	2
1 Financial behavior (Bartlett score)	0.00	1.00	–	
2 Life satisfaction	7.10	2.29	0.12***	–
3 Financial knowledge (original question version)	4.28	1.41	0.23***	0.24***
4 Financial knowledge (revised question version)	4.42	1.46	0.28***	0.32***
5 Financial knowledge (both question versions)	4.35	1.44	0.25***	0.28***
6 Question version[a]	0.50	0.50	0.02	−0.01

Note: All coefficients are Pearson correlations, [a]0 = Original question, 1 = Revised question. ***p < .001

Finally, we investigated the criterion-related validity of both question versions by comparing their associations with the two criterion variables on financial behavior and life satisfaction, respectively. To do so, we conducted four bivariate regressions predicting responses to the criterion measures from responses to the two versions of the financial knowledge question. The descriptive statistics of the criterion and the predictor variables, as well as the intercorrelations between these variables, are shown in Table 4.

As shown in Table 5, the coefficients differed between the two experimental groups and the differences were in the expected directions (i.e., the associations were stronger in the revision condition). To determine the significance of the difference between these associations, we ran two further multiple regressions predicting responses to the criterion questions using responses to the financial knowledge question, a dummy variable representing the question version (coded 0 for individuals who received the original version and 1 for individuals who received the revised version), and the interaction of responses to the financial knowledge question and question version (see Shaeffer et al. 2005 for a similar approach). As shown in Table 6, neither of the two interactions were significant indicating that the two

question versions did not differ in terms of criterion-related validity. Hence, we found no support for our sixth hypothesis.

6 Discussion and conclusion

Pretesting survey questions via cognitive interviewing rests on the assumptions that the method uncovers question problems which – if undetected – undermine data quality and that it provides insights into how survey questions can be improved. The purpose of this study was to empirically test these assumptions in a web survey experiment employing a question on self-reported financial knowledge that had been pretested via cognitive interviewing. Specifically, we examined whether the problems identified by cognitive interviewing were also observable in a web survey experiment and whether a question revision based on cognitive interviewing findings exhibited higher content-related and criterion-related validity than the original draft question.

As predicted by the cognitive interviews, there was no homogenous understanding of the term 'financial knowledge' among respondents in the control condition of the web survey and the individual interpretations had the expected effect on respondents' ratings of their subjective financial knowledge. Respondents interpreting the term as primarily referring to knowledge about complex financial processes rated their financial knowledge lower than those who understood it to refer primarily to the ability of handling one's personal finances. These findings suggest that the problems uncovered by cognitive interviewing were indeed real problems, a result which is similar to that of earlier studies on the effectiveness of cognitive interviewing (Forsyth et al. 2004; Willis and Schechter 1997).

At the same time, we were less successful in improving the question on financial knowledge based on the cognitive interviewing results. We obtained some evidence of higher content-related validity for the revised question version, but this was not pervasive. On the one hand, more respondents in the revision condition interpreted the question as referring to managing one's personal finances. This aspect was explicitly highlighted in the definition in the question and, as expected, it led respondents to rate their financial knowledge higher than respondents in the original condition. On the other hand, the differences in response distributions to the closed probe and the target question were only marginal and arguably of limited practical significance. Moreover, respondents receiving the revised question should ideally have interpreted it as referring both to knowledge about more complex financial processes and about managing one's personal finances (as clarified by the definition). However, the proportion of respondents thinking about both aspects was similar between both conditions.

There are several possible explanations for this finding. First, given that the definition placed more emphasis on handling personal finances (as measured by the number of words used to explain this aspect in comparison to the other aspect), it is possible that it drew respondents' focus primarily on managing personal finances, rather than emphasizing both aspects equally. Second, it is conceivable that respondents perceived the question text as too lengthy and therefore skipped reading the definition or just skimmed over it. Or else, respondents may tend to weigh one facet of financial knowledge more than the other, for example, if instances of one facet come to mind easier than instances of the other. If this was the case, the question should rather be split into two questions (one asking for knowledge about

Table 5 Bivariate regressions predicting financial behavior and life satisfaction with self-reported financial knowledge

Dependent variables	Original coefficients (SE)	$F(df1=1)$	df2	R^2	N	Revision coefficients (SE)	$F(df1=1)$	df2	R^2	N	Differences between coefficients
Financial behavior	0.16*** (0.02)	57.68***	1034	0.05	1036	0.18*** (0.02)	83.32***	1005	0.08	1007	–0.02 (0.03)
Life satisfaction	0.39*** (0.05)	62.66***	1047	0.06	1049	0.48*** (0.05)	113.48***	1016	0.10	1018	–0.09 (0.07)

Note: Unstandardized regression coefficients are shown. Standard errors (SE) in parentheses

***p < .001

Table 6 Multiple regressions testing the difference between question versions in terms of criterion-related validity

	Predictors						
Dependent variables	Financial knowledge (SE)	Question version (SE)	Financial knowledge × Question version (SE)	$F(df1 = 3)$	df2	R^2	N
Financial behavior	0.16*** (0.02)	–0.08 (0.14)	0.02 (0.03)	47.36***	2039	0.07	2043
Life satisfaction	0.39*** (0.05)	–0.50 (0.30)	0.09 (0.07)	57.53***	2063	0.08	2067

Note: Unstandardized regression coefficients are shown. Standard errors (SE) in parentheses

***p < .001

complex financial processes and one asking for knowledge about managing one's personal finances) instead of being supplemented by a definition. A fourth explanation is related to the design of the probe that we used to gather information on how they had interpreted the target question. In contrast to asking open-ended probes, it remains somewhat unclear whether the responses to closed probes really reflect the cognitive processes that respondents are going through while answering the survey questions or whether the response options remind them of possible interpretations they might not have thought of otherwise (Neuert et al. 2021; Schwarz 1999). Finally, it is possible that the probe was flawed because it asked what respondents had thought about when answering the question rather than what they had factored into their response. Some respondents might have thought about something, rejected it as not relevant to the question, and then answered on a different basis. Such participants might still respond to the probe indicating that they thought about something that did not influence their answer to the question. Of course, these are only ad hoc explanations which require additional research.

We obtained no evidence for the revised version to be associated with higher criterion-related validity than the original version. Even though we found the expected differences in correlations between the financial knowledge question and the two criterion questions, these were not statistically significant. Again, this could be due to respondents in the revision condition not reading the definition of 'financial knowledge' thoroughly or – despite doing so – giving more weight to one of the two central facets, and hence answering basically the same question as those in the original condition. Another explanation is that the criterion questions were not particularly well suited for gauging criterion-related validity, an assertion which is supported by the relatively weak (albeit statistically significant) correlation between the target question and the criterion measure of financial behavior. It would be desirable if future research replicated our study with a different set of criterion measures.

All in all, our results indicate that cognitive interviewing identifies problems that truly exist in a real survey setting, but that it is not necessarily effective in repairing these problems and producing a question revision that is of higher quality than the original draft version. As we mentioned in the background section, one could argue that the latter expectation is overly demanding of the method. Cognitive interviewing results may be clear as to whether question problems exist, and ideally, they can also point out the causes of these problems. However, finding a solution to eliminate the problems (without introducing new ones) is ultimately the task of the questionnaire developers. It is possible that the revised

question version would have been associated with higher criterion-related validity if we as survey designers had come to different conclusions on how the draft question could be fixed (e.g., splitting it into two questions instead of adding a definition). The results of our experiment suggest that it might be insufficient to carry out only one round of pretesting and revising questions based on these data only. The revisions themselves need to be evaluated as well, ideally in a field test, to make sure that they are indeed of higher quality than the earlier draft versions. Even though this form of iterative pretesting is often referred to as best practice in textbooks and research articles on cognitive interviewing (e.g., Beatty and Willis 2007; Collins 2015; Willis 2005), we do not know how common this practice is among practitioners given time and resource constraints. Based on this study's results, we clearly recommend this practice being adopted at least to some degree in questionnaire pretesting studies (i.e., conducting at least two rounds of testing), irrespective of the pretesting method being applied.

There are two limitations to this study pointing at additional avenues for future research. First, our experiment included only one survey question, which clearly restricts the generalizability of our findings. It would be useful if future research tested the effectiveness of cognitive interviewing with a larger set of questions on different topics and of different types (e.g., attitudinal, factual, behavioral, and knowledge questions). For some of these questions, other types of analyses are conceivable to examine the effectiveness of question revisions. For instance, for some factual or behavioral questions, the validity of survey results could be examined using external data sources (such as registry data for demographic questions). For attitude questions, the impact of item revisions on multi-item measures could by examined by testing for measurement invariance. Secondly, organizations and researchers vary in the ways they conduct cognitive interviews, and thus, there is no such thing as 'the' cognitive interview. The practices adopted in this study might be different to other cognitive interviewing practices, which again limits the generalizability of our results. Notwithstanding these limitations, we believe that even small-scale studies such as the one reported here are important in expanding our understanding of what cognitive interviewing can and cannot do and how much confidence we can place in it as a method for pretesting questionnaires.

7 Appendix: Question wordings (Original German formulations and English translations)

1. Question on self-reported financial knowledge:

a. Original version: Auf einer Skala von 1 bis 7, wobei 1 ‚sehr gering' und 7 ‚sehr groß' bedeutet: Wie würden Sie Ihr Finanzwissen einschätzen?

(On a scale from 1 to 7, where 1 means very low and 7 means very high, how would you assess your overall financial knowledge?)

b. Revised version: Auf einer Skala von 1 bis 7, wobei 1 ‚sehr gering' und 7 ‚sehr groß' bedeutet: Wie würden Sie Ihr Finanzwissen einschätzen? Unter Finanzwissen verstehen wir hier das Verständnis von Themen im Finanzbereich sowie die Fähigkeit, angemessene und informierte Entscheidungen über persönliche Finanzen zu treffen, wie z.B. Haushaltsplanung, Geldanlagen, Versicherungen, Immobilien, Schuldenverwaltung oder Steuerplanung.

(On a scale from 1 to 7, where 1 means very low and 7 means very high, how would you assess your financial knowledge? Financial knowledge here means the understanding of

financial issues and the ability to make appropriate and informed decisions about personal finances, such as budgeting, investments, insurance, real estate, debt management or tax planning.)

2. Semi-open probing question:

Bei der vorherigen Frage ging es darum, wie Sie Ihr Finanzwissen einschätzen. Woran haben Sie beim Beantworten der Frage gedacht?

(1) Daran, wie gut ich mich mit finanzwirtschaftlichen Themen auskenne (Börsen, Kapitalmarkt, Investmentfonds usw.), (2) Daran, wie gut ich mich mit Themen auskenne, die meine persönlichen Finanzen betreffen (Haushaltsplanung, Geldanlagen, Schuldenverwaltung usw.), (3) An beide der oben genannten Themen, (4) An etwas anderes, und zwar: [open text field]

(The previous question was about how you assess your financial knowledge. What did you consider when answering the question?

(1) How knowledgeable I am about financial topics (stock markets, capital markets, mutual fonds, etc.), (2) How knowledgeable I am about topics that affect my personal finances (budgeting, investing, managing debt, etc.), (3) Both of the above-mentioned topics, (4) Something different, namely: [open text field])

3. Criterion questions:

a. Financial behavior:

i. Bevor ich etwas kaufe, überlege ich mir genau, ob ich es mir leisten kann.
ii. Ich habe meine finanziellen Angelegenheiten sehr genau im Blick.
iii. Ich stecke mir langfristige finanzielle Ziele und versuche, diese auch zu erreichen.

Answer options: (1) Stimme voll und ganz zu, (2) Stimme eher zu, (3) Weder noch, (4) Stimme eher nicht zu, (5) Stimme überhaupt nicht zu.

(i. Before I buy something, I carefully consider whether I can afford it.)

(ii. I keep a close personal watch on my financial affairs.)

(iii. I set long term financial goals and strive to achieve them.)

Answer options: (1) Strongly agree, (2) *Rather agree*, (3) *Neither/nor*, (4) *Rather disagree*, (5) *Strongly disagree*.

b. Life satisfaction:

Wie zufrieden sind Sie gegenwärtig, alles in allem, mit Ihrem Leben?

Answer options (1) Überhaupt nicht zufrieden – (11) Völlig zufrieden.

(All things considered, how satisfied are you with your life at present?

Answer options: (1) Not at all satisfied – (11) Extremely satisfied)

Funding Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 491156185.

Open Access funding enabled and organized by Projekt DEAL.

Data Availability Data are available from the corresponding author upon request.

Code Availability Code for data cleaning and analysis is available from the corresponding author upon request.

Conflicts of interest/Competing interests The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Bartlett, M.S.: A note on the multiplying factors for various chi square approximations. J. R Stat. Soc. **16**(2), 296–298 (1954)

Beatty, P.C., Willis, G.B.: Research synthesis: The practice of cognitive interviewing. Public. Opin. Q. **71**(2), 287–311 (2011). https://doi.org/10.1093/poq/nfm006

Beierlein, C., Kovaleva, A., László, Z., Kemper, C.J., Rammsted, B.: Kurzskala zur Erfassung der Allgemeinen Lebenszufriedenheit (L-1). Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS) (2015). https://doi.org/10.6102/zis229

Blair, J., Conrad, F.G.: Sample size for cognitive interview pretesting. Public. Opin. Q. **75**(4), 636–658 (2011). https://doi.org/10.1093/poq/nfr035

Boeije, H., Willis, G.: The cognitive interviewing reporting framework (CIRF): Towards the harmonization of cognitive testing reports. Methodology. **9**(3), 87–95 (2013). https://doi.org/10.1027/1614-2241/a000075

Chang, L., Krosnick, J.A.: Measuring the frequency of regular behaviors: Comparing the 'typical week' to the 'past week'. Sociol. Methodol. **33**(1), 55–80 (2003). https://doi.org/10.1111%2Fj.0081-1750.2003.t01-1-00127.x

Collins, D.: Pretesting survey instruments: An overview of cognitive methods. Qual. Life Res. **12**(3), 229–238 (2003). https://doi.org/10.1023/A:1023254226592

Collins, D.: Cognitive interviewing practice. Sage, Thousand Oaks (2015)

Conrad, F.G., Blair, J.: Data quality in cognitive interviews: The case of verbal reports. In: Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Singer, E. (eds.) Methods for testing and evaluating survey questionnaires, pp. 67–88. Wiley, New York (2004). https://doi.org/10.1002/0471654728.ch4

Conrad, F.G., Blair, J.: Sources of error in cognitive interviews. Public. Opin. Q. **73**(2), 32–55 (2009). https://doi.org/10.1093/poq/nfp013

D'Ardenne, J., Collins, D.: Data management. In: Collins, D. (ed.) Cognitive interviewing practice, pp. 142–161. Sage, Thousand Oaks (2015)

Falahati, L., Sabri, M.F., Paim, L.H.: Assessment a model of financial satisfaction predictors: Examining the mediate effect of financial behavior and financial strain. World Appl. Sci. J. **20**(2), 190–197 (2012). https://doi.org/10.5829/idosi.wasj.2012.20.02.1832

Forsyth, B.H., Rothgeb, J.M., Willis, G.B.: Does questionnaire pretesting make a difference? An empirical test. In: Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Singer, E. (eds.) Methods for testing and evaluating survey questionnaires, pp. 525–546. Wiley, New York (2004). https://doi.org/10.1002/0471654728.ch25

Fowler, F.J.: How unclear terms affect survey data. Public. Opin. Q. **56**(2), 218–231 (1992). https://doi.org/10.1086/269312

Fowler, F.J.: Improving survey questions. Sage, Thousand Oaks (1995)

Goyal, K., Kumar, S.: Financial literacy: A systematic review and bibliometric analysis. Int. J. of Consum. Stud. **45**(1), 80–105 (2021). https://doi.org/10.1111/ijcs.12605

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R.: Survey Methodology. Wiley, Hoboken (2004)

Höhne, J.K., Yan, T.: Investigating the impact of violations of the "left and top means first" heuristic on response behavior and data quality. Int. J. Soc. Res. Methodol. **23**(3), 347–353 (2020). https://doi.org/10.1080/13645579.2019.1696087

Kaiser, H.F.: A second generation little jiffy. Psychometrika. **35**(4), 401–415 (1970). https://doi.org/10.1007/BF02291817

Kane, M.T.: Validation. In: Brennan, R.L. (ed.) Educational measurement, 4th edn., pp. 17–64. American Council on Education and Praeger, Westport (2006)

Lenzner, T., Hadler, P., Nießen, D., Quint, F., Steins, P., Neuert, C.: SHARE Wave 9 – New items on financial decision making, successful ageing, eating habits, sleep, long-term care insurance, and long-term care expectations (English Version): Cognitive Pretest. GESIS Project Reports (2020). https://doi.org/10.17173/pretest81

Lenzner, T., Neuert, C., Otto, W.: Cognitive Pretesting. GESIS Survey Guidelines (2016). https://doi.org/10.15465/gesis-sg_en_010

Lind, T., Ahmed, A., Skagerlund, K., Strömbäck, C., Västfjäll, D., Tinghög, G.: Competence, confidence, and gender: The role of objective and subjective financial knowledge in household finance. J. Fam Econ. Issues. **41**(4), 626–638 (2020). https://doi.org/10.1007/s10834-020-09678-9

Miller, K.: Cognitive interviewing. In: Madans, J., Miller, K., Maitland, A., Willis, G.B. (eds.) Question evaluation methods: Contributing to the science of data quality, pp. 51–75. Wiley, New York (2011)

Neuert, C., Meitinger, K., Behr, D.: Open-ended versus closed probes: Assessing different formats of web probing. Sociol. Methods Res. (2021). https://doi.org/10.1177/00491241211031271

OECD: : OECD/INFE international survey of adult financial literacy competencies. OECD. (2016). https://www.oecd.org/finance/OECD-INFE-International-Survey-of-Adult-Financial-Literacy-Competencies.pdf Accessed 4 April 2022

Park, H., Sha, M.M., Willis, G.: Influence of English-language proficiency on the cognitive processing of survey questions. Field Methods. **28**(4), 415–430 (2016). https://doi.org/10.1177/1525822X16630262

Prüfer, P., Rexroth, M.: Kognitive Interviews [cognitive interviews]. ZUMA How-to-Reihe 15. (2005). https://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/How_to15PP_MR.pdf?download¼true Accessed 12 June 2021

Remund, D.L.: Financial literacy explicated: The case for a clearer definition in an increasingly complex economy. J. Consum. Aff. **44**(2), 276–295 (2010). https://doi.org/10.1111/j.1745-6606.2010.01169.x

Robb, C.A., Woodyard, A.S.: Financial knowledge and best practice behavior. J. Financ Couns. Plan. **22**(1), 60–70 (2011)

Rothgeb, J.M., Willis, G.B., Forsyth, B.H.: Questionnaire pretesting methods: Do different techniques and different organizations produce similar results? Bull. Methodol. Sociol. **96**(1), 5–31 (2007). https://doi.org/10.1177/075910630709600103

Schuman, H., Presser, S.: Questions and answers in attitude surveys. Academic Press, New York (1981)

Schwarz, N.: Self-reports: How the questions shape the answers. Am. Psychol. **54**(2), 93–105 (1999). https://doi.org/10.1037/0003-066X.54.2.93

Shaeffer, E.M., Krosnick, J.A., Langer, G.E., Merkle, D.M.: Comparing the quality of data obtained by minimally balanced and fully balanced attitude questions. Public. Opin. Q. **69**(3), 417–428 (2005). https://doi.org/10.1093/poq/nfi028

Sudman, S., Bradburn, N.M.: Asking questions: A practical guide to questionnaire design. Jossey-Bass, San Francisco (1982)

Tourangeau, R., Rips, L.J., Rasinski, K.: The psychology of survey response. Cambridge University Press, Cambridge (2000)

Willis, G.B.: Cognitive interviewing: A tool for improving questionnaire design Sage, London (2005)

Willis, G.B.: Analysis of the cognitive interview in questionnaire design. University Press, Oxford (2015)

Willis, G.B., Schechter, S.: Evaluation of cognitive interviewing techniques: Do the results generalize to the field? Bull. Methodol. Sociol. **55**(1), 40–66 (1997). https://doi.org/10.1177/075910639705500105

Yan, T., Kreuter, F., Tourangeau, R.: Evaluating survey questions: A comparison of methods. J. Off Stat. **28**(4), 503–529 (2012)

Yeager, D.S., Krosnick, J.A.: Does mentioning "some people" and "other people" in an opinion question improve measurement quality? Public. Opin. Q. **76**(1), 131–141 (2012). https://doi.org/10.1093/poq/nfr066