Journal of Official Statistics, Vol. 13, No. 2, 1997, pp. 181–199

Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality

Bärbel Knäuper[1], *Robert F. Belli*[1], *Daniel H. Hill*[1] *and A. Regula Herzog*[1,2]

There is increasing evidence that reductions in cognitive functioning can be negatively related to the quality of survey data. Research also indicates that the accuracy and completeness of reports decrease with increasing question difficulty. In the present article the *interaction* of question difficulty with respondents' cognitive ability is investigated. It was expected that respondents with relatively low information processing ability are less able than those with higher ability to provide complete and accurate answers when responding to particularly difficult questions. The number of "don't know" responses in a survey conducted with people over 70 years of age was used as an indicator of reduced data quality. The data was analyzed as a function of question difficulty and respondents' cognitive ability. The findings demonstrate the expected interaction: Respondents lower in cognitive ability were particularly likely to answer "don't know" to difficult questions. Respondents higher in cognitive ability were less affected by variations in question difficulty. The selective loss of data due to failures to respond may bias survey findings because it leads to an under-representation of data from lower cognitive ability respondents in the data. Implications for conducting surveys with lower cognitive ability respondents are discussed.

Key words: Elderly; survey methodology; don't know responses; response bias.

1. Introduction

Question answering involves a series of cognitive tasks that respondents have to solve to provide high-quality data. These tasks include understanding and interpreting the meaning of the question, conducting a memory search, making judgments, and formatting the response onto a given response scale (e.g., Cannell, Miller, and Oksenberg 1981; Tourangeau 1984). An optimal performance of these tasks may be difficult for respondents low in cognitive ability. There is some evidence to suggest that reductions in cognitive functioning have a negative effect on the quality of survey data. So far, however, this has mostly

[1] Institute for Social Research, Survey Research Center, University of Michigan, Ann Arbor MI, 48106-1248, U.S.A. Address correspondence to the first author who is now at the Freie Universität Berlin, Abt. Gesundheitspsychologie, Habelschwerdter Allee 45, D-14195 Berlin (Germany), e-mail: bknauper@zedat.fu-berlin.de
[2] Institute of Gerontology and Department of Psychology, University of Michigan, Ann Arbor MI, 48106, U.S.A.

Acknowledgments: Support for the present research came primarily from a data analysis grant of the National Institute of Aging (AHEAD Wave I Early Results Workshop). Preparation of this article was also supported by a Research Fellowship from the German Research Council and a Career Development Award Fellowship from the Max Planck Society to the first author and a grant from the National Institute on Aging (AG02038) to the fourth author. We thank Steve Blixt, Lynn Dielman, Lucia Juarez, Jim Lepkowski, Dan Mroczek, and Trivellore Raghunathan for valuable assistance with data analysis. For very helpful comments on an earlier draft of this article we would like to thank Peter Muehlberger, Willard Rodgers, Norbert Schwarz, and Hans-Ulrich Wittchen.

been concluded from studies in which respondents' age was taken as a proxy variable for reduced cognitive ability (e.g., Andrews and Herzog 1986; Colsher and Wallace 1989; Gergen and Beck 1966; Herzog and Dielman 1985; Rodgers, Herzog, and Andrews 1988). In these studies it is implicitly assumed that the reductions in data quality are due to reductions in the respondents' cognitive ability (as indicated by the respondent age). Andrews and Herzog (1986), for example, found that the percentage of true score variance in survey measures tends to decline as respondent age increases and that the percentage of random and correlated error variance increases. Colsher and Wallace (1989) found that older respondents are more likely to refuse to answer specific questions. Other authors found older respondents more likely to show response biases such as social desirability tendencies (Campbell, Converse, and Rogers 1976) and acquiescence biases (Kogan 1961). It was also suggested that older respondents tend to use response scales in a more stereotypical way (Andrews and Herzog 1986) and show larger response order effects (Knäuper 1997). Although cognitive aging research has provided convincing evidence that aging is correlated with decreasing cognitive ability (see Salthouse, Babcock, and Shaw 1991, for a review) the above mentioned studies fail to provide direct evidence for a negative effect of reduced cognitive ability on data quality because cognitive ability was not directly assessed. The present article investigates directly if reductions in cognitive ability and memory capacity among older people are associated with the completeness and accuracy of responses to survey questions.

It is well known that the accuracy and completeness of reports also decline with increasing *question difficulty*. According to several authors (e.g., Cannell, Miller, and Oksenberg 1981; Schwarz 1990; Tourangeau 1984; Tourangeau and Rasinski 1988; Willis, Royston, and Bercini 1991), providing optimal answers to survey questions requires the performance of a number of cognitive tasks. First, the question's meaning has to be understood. Then, a memory search has to be conducted to retrieve information relevant for answering the question. Once information is retrieved, there may need to be an integration of the information to construct a summary judgment. Finally, the retrieved or constructed answer has to be formatted onto a given response scale. The difficulty of these cognitive tasks varies depending on, for example, the content or topic of the question, the time frame it covers, or the response options it provides (for an overview see Belson 1981; Clark and Schober 1992).

As Krosnick (1991) points out, difficult questions may lead respondents to provide a satisfactory answer instead of an "optimal" one, which he terms as "satisficing." He assumes that satisficing can present itself in either incomplete or biased information reports, or in no report at all (e.g., answering "don't know"). Krosnick suggests that satisficing is a function of three factors and their interactions: (a) the difficulty of the task (question difficulty), (b) the ability of performing the task, and (c) the motivation to perform the task. In a related vein, John and Cole (1986) point out that information-processing deficits in older persons are pronounced in situations where (a) large amounts of information are presented, (b) the information is not presented in an optimal format, and (c) when the task requires the use of difficult response formats. Applying these notions to the survey situation, we expect to find respondents with low cognitive ability more likely than those with higher ability to give incomplete, biased, or no responses to particularly difficult questions (cf., Kaldenberg, Koenig, and Becker 1994).

The interaction of question difficulty with individual differences in cognitive ability and its effect on data quality so far has not been empirically investigated. Survey results can be systematically biased if respondents lower in cognitive ability show a selective, non-random failure to respond to difficult questions or if they answer them in a systematically biased way. Support for this has been found in recent secondary and laboratory studies which revealed a response bias among older respondents low in cognitive ability for difficult questions about mental health issues (Knäuper and Wittchen 1994). The findings demonstrate that respondents who performed poorly in working memory tasks showed an acquiescence bias for questions with a high degree of complexity. Similarly, lower cognitive ability respondents may entirely fail to give a response to particularly difficult questions. Accordingly, responses from this subsample of respondents would be underrepresented for topics that are assessed by more difficult questions.

The relationship between content and question difficulty probably is not random. Complex matters are more likely to be phrased in difficult questions. Thus, reduced data quality due to low cognitive ability can be expected to be more of a problem for complex issues. False conclusions are a danger since low cognitive ability has been shown to be associated with a number of other variables, such as poor physical health (cf., Colsher and Wallace 1989). The following question (which is taken from the survey on which the data from the present article are based) may illustrate this possible bias: ''Not counting costs covered by insurance, about how much did you end up paying for any part of hospital and doctor bills and any other medical or dental expenses in the last twelve months, since [MONTH of (1992/1993]?''. Respondents low in cognitive ability may be less likely to answer complex questions such like this one. If, as indicated by other research (e.g., Colsher and Wallace 1989), low cognitive ability is associated with poor physical health, the aggregate amount of expenses for hospital and doctor bills would be underestimated.

For the purpose of this article, the number of ''don't know'' responses (DKs) was chosen as an indicator of reduced data quality in investigating the interaction between question difficulty and individual differences in cognitive ability. Krosnik (1991) and others have pointed to the role of DKs as indicators of comprehension difficulties or satisficing response behavior. Choosing the number of DKs as a data quality indicator facilitated the investigation of interactions between variables on the question level (indicators of question difficulty) and variables on the respondent level (cognitive ability) because every question provided us with a data point for this dependent variable. Other indicators of reduced data quality such as acquiescence biases or biases in using response scales would only have been applicable for selected questions. In addition, the number of refusals were analyzed in the same manner as the number of DKs. The results were basically the same for both indicators of reduced data quality. Because of the low overall number of refusals, however, these findings are not reported here.

We used data from the Assets and Health Dynamics Among the Oldest Old (AHEAD) survey for the analysis. Overall, it is expected that in comparison to persons with higher levels of cognitive ability, persons with lower levels will have more DKs in responding to questions generally, but have an even greater tendency to respond DK with difficult questions. For persons with higher levels of cognitive ability, the number of DKs is predicted to be less dependent on question difficulty.

2. Method

2.1. Sample and measures of cognitive ability

Sample. AHEAD is a national panel survey of community residents born in 1923 or earlier living in the community. The primary goal of AHEAD is to provide data to understand the health and resource dynamics in advanced old age. The data collection on which this article is based was the first wave of the survey conducted in 1993 and 1994. In 1992 households with residents born in 1923 or before were identified as part of the sampling screening for the Health and Retirement Survey (HRS; Juster and Suzman 1995) from a multi-stage area probability sample. If more than one eligible person was in the household, one was randomly selected. In cases where the selected individual was married at the time of the baseline interview in 1993–94, the spouse was also asked to complete the survey, regardless of his or her age. The response rate was 80.0%. More than ninety per cent (90.6%) of the respondents were older than 70 years of age at the time of the interview. The median age was 76 years (Mean = 76.48 years). 36.7% of the respondents were male, 63.3% female.

Interview mode. The intention was to seek telephone interviews with those under age 80, and face to face interviews with those over age 80. In practice, about 75% of the interviews with those in each age group were done using the ''intended'' mode, the rest was done in the mode that was preferred by the individual respondent. As expected, respondents interviewed by telephone were younger, more highly educated, and healthier than those interviewed face to face. They also performed significantly better on all cognitive tests (for details see Herzog and Wallace 1995).

Mode and its interactions with our independent variables would have been an additional interesting variable to examine. However, as just described, there are problems in terms of the confound of mode with cognitive ability. Because of this self-selection bias and due to the complexity of the analysis we concluded that the present data would not be an optimal means by which to examine mode effects directly. Therefore, no analysis by mode will be reported in this article.

Cognitive tasks. As part of the AHEAD survey, a series of cognitive functioning tests was administered to the respondent. These tests were derived from or modified on the basis of well-validated measures developed in psychological research on intelligence and cognition as well as in geriatric and neurological research on cognitive impairment and dementia. A detailed description of the measures and their psychometric properties is given in Herzog and Wallace (1995). Basically, the cognitive measures capture the dimensions of fluid intelligence, crystallized intelligence, and memory. Many of the tests were drawn or adapted from the Telephone Interview for Cognitive Status (TICS) by Brandt, Spencer, and Folstein (1988), adapted for use over the telephone from the Mini-Mental State Exam (Folstein, Folstein, and McHugh 1975). Out of the entire pool of conducted cognitive tests, those were selected for the purpose of the present article that have a high emphasis on verbal learning, reasoning, and attention abilities (Herzog and Wallace 1995). This was done because these abilities reflect most closely the cognitive tasks involved in

answering survey questions (cf., Krosnick 1991). Specifically, the following measures were selected:

(1) Immediate free recall. This task tests memory: Ten short, concrete, high-frequency nouns were read to the respondent, who was then asked to recall as many of them as possible.
(2) Serial 7s test: This task assesses working memory: The respondent was asked to start at 100 and subtract by increments of seven for five trials. One point was given for each correct subtraction for a maximum of five points.
(3) Measures of knowledge, language, and orientation: They include counting backwards from 20 for 10 continuous numbers; naming the day of the week and the date; naming of the objects that "people usually use to cut paper" and the "kind of prickly plant that grows in the desert"; naming the president and vice president of the U.S.

Out of these measures, a combined "cognitive score" was built. The cognitive score is gained by summing up the raw scores for the three different test domains. It ranges from 1 to 25.

We chose to treat cognitive ability as a dichotomy because we needed to keep the number of covariate classes at a computationally manageable level.[3] The alternative would have been to maintain the original interval scaling in which case we would have had to evaluate the likelihood function separately for each of the 1,104,427 person-question pairs in the sample. This would be feasible for a single analysis, but our study requires testing several alternative specifications of interactions and the computation of complex sampling errors – each involving many repetitions of the estimation. We therefore dichotomized the cognitive ability measure on the basis of the distribution of the cognitive ability score. For $N = 6,933$ cognitive ability data were available. The distribution of the cognitive ability score is highly skewed, with most respondents receiving rather high scores. Based on a median split, the sample was divided into a from now on called "lower cognitive ability group" (scores 1–17, $n = 3,271$) and a so called "higher cognitive ability group" (scores 18–25, $n = 3,662$). Because of the age and potential frailty of the AHEAD population, an interview with a proxy respondent was conducted for respondents who could not be interviewed themselves. No data on cognitive measures are available for these respondents and they are therefore not included in the presented analysis.

2.2. *Indicators of question difficulty*

Nine question characteristics that are likely to be related to difficulty were derived from the literature (e.g., Belson 1981; see Table 1 for an overview). One or more question characteristics were identified for each cognitive step associated with the question answering process as described above (understanding and interpreting, retrieval and judgment, and response formatting). Below, we discuss each of the question difficulty indicators with respect to the demands they most likely place on cognitive processing (although this may, of course, vary with question topic). With regard to these cognitive demands, indicators that signify an increase in cognitive load are expected to be especially troublesome

[3] For an explanation of the computational advantages of grouping and covariate classes, see McCullagh and Nelder (1989), Chapter 4.

Table 1. Question characteristics: description, coding, and interrater-agreement

Characteristics	Description	Code	Agreement per cent	Kappa* (*SE*)
Understanding				
Question length	Question length indicated by number of words.	# of words	$r = -.94.2$†	–
Question complexity	Whether or not the Q is (mainly syntactical) complex.	0 = no, 1 = yes	76.0	.36 (.14)
Instructions	Whether or not the Q contains one or more instructions.	0 = no, 1 = yes	97.2	.54 (.11)
Introductory phrase	Whether or not the Q contains an introductory phrase.	0 = no, 1 = yes	98.6	.93 (.07)
Ambiguous terms	Whether or not the Q includes ambiguous or unfamiliar terms.	0 = no, 1 = yes	76.4	.53 (.10)
Retrieval/judgment				
Retrospective report	Whether or not the Q asks for a retrospective report.	0 = no, 1 = yes	81.9	.54 (.11)
Frequency report	Whether or not the Q asks for a behavioral frequency report.	0 = no, 1 = yes	97.2	.65 (.23)
Quantity report	Whether or not the Q asks for a report of numerical quantity.	0 = no, 1 = yes	93.1	.85 (.07)
Response formatting				
Response scale	Whether or not the Q provides a response scale.	0 = no, 1 = yes	94.4	.80 (.10)

*Interrater-agreement is based on independent ratings of $N = 72$ questions. †Pearson Correlation Coefficient.

for respondents lower in cognitive ability in comparison to those higher in cognitive ability, and indicators that signify a decrease in cognitive load are expected to be especially beneficial to respondents lower in cognitive ability.

2.2.1. Understanding and interpretation of the question meaning

1. Question length. Increased question length *by itself* (i.e., controlling for all other question difficulty characteristics) has been shown to make a question easier to answer (Cannell, Marquis, and Laurent 1977; Cannell et al. 1981). That is, a long question should be easier to understand than a shorter one if it contains redundant information, but not if new terms are introduced or if the question gets syntactically complex due to the length. Cannell and colleagues demonstrated that long questions providing redundant information can lead to increased responding because they give respondents more time to think about and formulate an answer.

2. Question complexity. Complex syntactical structures will tax the ability to apply the appropriate parsing and inference rules that are necessary to comprehend and understand the meaning of the question. Questions that contained complex syntactical structures, such as, for example, embedded sentences or inverted sentences, were coded as being complex.

3. Instructions. Questions which instruct subjects to include or exclude experiences or examples from consideration in their answer should be more difficult to answer, because they involve additional cognitive "computing" and may require comprehending the association among many details. (Example: "Not counting any money or assets that you may have given to children or others, did you use up any of your investments or savings during 1992 to pay for expenses?").

4. Introductory phrases. Questions which are introduced with a phrase, for example to soften the question's directness or to introduce complex terms, can provide difficulties because they introduce a number of details that are required to be integrated. However, in cases in which complex or ambiguous terms are explained in the introductory phrase it can be a help in understanding the questions, thus, *decreasing* the difficulty of the question. (Example: "The next question might not be easy to talk about, but it is very important for research on health and aging. During the last twelve months, have you lost any amount of urine beyond your control?").

5. Ambiguous terms. Some concepts may be difficult to translate into easy to understand questions. Comprehension problems can arise due to ambiguous or abstract terms, requiring the respondent to derive their meaning. For example, in the question, "In general, do you have less than one drink a day, one or two drinks a day, three or four drinks a day, or five or more drinks a day?") it might be unclear what a "drink" refers to.

2.2.2. Retrieval and judgment

6. Retrospective reports. Questions about the present are presumably easier to answer than questions that ask about some time in the past. Retrieving information from the past can pose considerable memory challenges for respondents. All questions asking about the past were coded as "retrospective reports."

7. Frequency reports. Questions asking for frequency reports of engaging in certain behaviors are taxing as they require an exhaustive memory search and counting (or the

computation of an estimate) of events or experiences. Examples are questions about the frequency of hospitalization, doctor visits, etc.

8. Quantitative reports. Questions asking for numerical quantity may also require an extensive memory search and counting or estimation processes (Example: "What about the value of the trusts? If you sold all the assets held in trust(s), about how much would you have?").

2.2.3. Response formatting

9. Response scales. Although, for example, frequency and quantity reports are expected to increase cognitive demands, providing responses scales should *decrease* cognitive demand because respondents can use the information given in the response scale to simplify the memory search and computational processes that they otherwise would use in formatting their response (cf., Schwarz 1990, 1994).

2.3. Coding

All questions from the sections on health care and costs, housing, job status, expectations, income, net worth, and insurance issues of the AHEAD survey were coded for question characteristics. Some questions ("unfolding questions") were asked only in cases in which a "don't know" response occurred to a previous question, and these unfolding questions were treated and coded as separate questions. Altogether, 723 AHEAD questions were coded by the first author (BK) with regard to the nine question characteristics. After consultation with the first author regarding coding criteria, the second author (RB) independently coded a random selection of 72 questions (10%) for purposes of estimating reliability. The next to the last column in Table 1 presents the final agreement (in per cent) between the two coders for the different difficulty characteristics (Pearson Correlation Coefficient for question length) on the 72 questions that were coded by both. Additionally, Kappa values of agreement are presented in the last column of Table 1 to provide a measure that accounts for agreement by chance. According to Fleiss (1981), values of Kappa above .75 represent excellent, and values from .40 to .75 represent fair to good, agreement beyond chance. The level of agreement for the question characteristic "complexity" was unsatisfactorily low (76%) and the chance adjusted Kappa is also very low for this measure ($K = .36$). This code was therefore removed from subsequent analysis. The Kappas for the remaining characteristics had good or excellent reliabilities with Kappa values ranging from $K = .53$ (ambiguity) to $K = .93$ (introductory phrase). In all cases, the codes assigned by the first author were maintained for the analysis reported below. Table 1 provides an overview about the coding criteria and reliabilities of the various question characteristics. The following question example may serve to illustrate the coding procedure for some of the question characteristics:

J54: "We have already asked how much income you [and your (husband/wife/partner)] received in (OCT93:1992/JAN94:1993). About how much total income did others in your household receive in (OCT93:1992/JAN94:1993) from Social Security, pensions, welfare, interest, gifts, jobs, or anything else? Do not include your own [or your (husband's/wife's/partner's)] income in this answer."

The length of this question indicated by the number of words is 43. Information that is not

asked to every respondent (as indicated by brackets) was not included into the counting of the number of words. This particular question contains an instruction (''Do not include your own income in this answer.'') and an introductory phrase (''We have already asked how much income ...''). The question asks for a retrospective report and more specifically a quantity report (''... how much total income ...?''). It does not provide a response scale.

2.4. *Probit analysis*

Our aim in the analysis was to assess the degree of influence of each of the question characteristics in combination with cognitive ability as predictors in leading to DKs. Statistical control for the collinearity among the question characteristic predictors was accomplished by conducting a multiple regression analysis. Because of the limited nature of the dependent variable, we chose to use probit analysis as the form of multiple regression. Symbolically, this statistical model can be derived by assuming that each individual has an underlying propensity to provide DK responses to each particular question. This propensity varies depending on the characteristics of the question and on the respondent's cognitive ability. For each of the cognitive ability groups ($c = 1, 2$) we can represent this propensity to give DK responses as:

$$D^*_{cqi} = \boldsymbol{\beta}^c \mathbf{X}_{qi} + \varepsilon_{cqi} \tag{1}$$

where $\mathbf{X}_{qi}$ is a vector of characteristics i of question q, $\boldsymbol{\beta}^c$ is a vector of parameters relating these characteristics to the DK propensity and ε_{cqi} is a random disturbance term which captures the net effect of all excluded factors on the propensity.

The propensity cannot, of course, be observed directly, but only its sign. That is, it is observed:

$$\begin{aligned} D_{cqi} &= 1 \quad \textit{iff } D^*_{cqi} > 0 \\ D_{cqi} &= 0 \quad \textit{iff } D^*_{cqi} \leq 0 \end{aligned} \tag{2}$$

where D is the empirically observable analogue of D^*. Substituting (1) into (2) and rearranging terms yield the following condition for obtaining a DK response:

$$\varepsilon_{cqi} > -\beta^c \mathbf{X}_{qi} \tag{3}$$

Assuming the ε's are distributed normally, Equation (3) implies a predicted probability of obtaining a DK response of:

$$Pr(\varepsilon_{cqi} > -\boldsymbol{\beta}^c \mathbf{X}_{qi}) = 1 - \Phi(-\boldsymbol{\beta}^c \mathbf{X}_{qi}) = \Phi(\boldsymbol{\beta}^c \mathbf{X}_{qi}) \equiv Pr_{cqi} \tag{4}$$

where Φ is the normal distribution function and the last equality holds by virtue of the symmetry of the normal distribution. The corresponding predicted probability of *not* obtaining a DK response is simply $1 - \Phi(\boldsymbol{\beta}^c \mathbf{X}_{qi})$.

Assuming that the ε are identically independently distributed the likelihood function for the model is obtained from these predicted probabilities via:

$$L(\boldsymbol{\beta}^c_{qi} | D_{ci}, \mathbf{X}_{qi}) = \prod_{i=1}^{N} \prod_{q=1}^{n_{qi}} Pr_{cqi}^{D_{cqi}} (1 - Pr_{cqi})^{(1-D_{cqi})} \tag{5}$$

Unbiased and fully efficient estimates of the β's can be obtained by maximizing this likelihood function with respect to them.

It is important to note that there are two major problems with this formulation. First, if estimated in its present form it would be a gargantuan numerical problem. The reason is that there are more than a million respondent-question combinations (exactly $N = 1,104,427$ combinations). Running probit models, such as this, with such a large sample is a major numerical undertaking. The solution to this problem is to collect common terms within the likelihood function – which amounts to analyzing the data at the question level and forming the dependent variable from the counts of DKs observed for each question (separately for higher and lower cognitive ability respondents). Each question's contribution to the likelihood function is effectively weighted by number of respondents falling into the two response categories (''DK'' and ''not DK'') with the sample stratified by cognitive ability score.

The second problem with the approach is more difficult to remedy satisfactorily. This problem relates to the independence assumption of the ε's. Because these error terms contain the effects of omitted factors on the DK propensity and some of these factors are at the individual rather than at individual-question level, there is bound to be some correlation across questions for a given respondent. Unless adjustments are made, these correlations will tend to bias the variance estimates for the β's downward – leading us to think that the results are more significant than they actually are. Ordinarily this situation would be remedied by means of some form of re-sampling scheme such as jackknifing or bootstrapping in which the true variances are teased out from the observed variability of repeated estimations on subsamples. This however is not feasible given the size of the estimation problem.

Instead, we adjusted the variance estimates by using a ''design effect'' based on the jackknifing variance estimates of the D_{qi} rather than the β's directly. In other words, the variance estimates (standard errors) were inflated by the (square root of the) design effect obtained for the proportion of each respondent's questions which were answered DK[4]. This design effect was, in turn, obtained using standard jackknife techniques and includes adjustments for other departures from simple random sampling inherent in the design (i.e., clustering and oversampling within some strata). A design effect of 15 was obtained and used for the current analysis.

To test the assumptions regarding the effects of question difficulty, cognitive ability, and the interaction between these predictors, one single probit model was built in which all question characteristics and interactions were entered simultaneously. In addition to the two-way interactions between cognitive ability and each question difficulty indicator, we were interested in examining two three-way interactions: (a) quantity report, response scale, and cognitive ability, and (b) ambiguity, introductory phrase, and cognitive ability. Specifically, we expected that quantity reports in general increase cognitive demands, but that providing response scales *decreases* cognitive demand because respondents can use the information given in the response scale to simplify the memory search and computational processes that they otherwise would use in formatting their response. This should be

[4] In general, parameters such as β are *less* seriously affected by departures from simple random sampling than are univariate parameters such as the mean proportion of DK responses. This means our correction procedure may result in over-inflated variance estimates.

a particular help for respondents lower in cognitive ability. Thus it is expected that the difference in DK responses between quantity reports with, in comparison to quantity reports without, response scales should be larger among respondents lower in comparison to respondents higher in cognitive ability. The expectations for the interaction between ambiguity and introductory phrase is based on similar theoretical assumptions: Introductory phrases are often used to clarify or introduce ambiguous terms. In these cases, the tendency to respond with DK should be attenuated, particularly among respondents lower in cognitive ability. The interactions between these specific question characteristics (quantity report/response scale, and ambiguity/introductory phrase) were entered simultaneously into the same model.

3. Results

Tables 2 and 3 present the results of the probit model. Table 2 includes the examination of the two-way interactions between all of the question characteristics, and Table 3 includes the results of two three-way interactions included in the model as discussed above. The tables present the adjusted mean percentages of DKs, the multivariate B-coefficients and design-corrected standard errors of the coefficients, and the *t*-statistics (the standard errors and *t*-statistics were computed while including the design-effect adjustment factor of 15). The adjusted means are obtained by averaging (across the entire sample) the predicted mean percentage of DKs obtained when each predictor is independently set to arbitrary values of zero and one (except for length, as discussed below). They are interpretable – as are adjusted means in traditional regression or MCA analysis – as the predicted value of the dependent variable association with the value of the independent variable in question, controlling for the distribution of all other independent variables. Question length scores of 5, 9, and 14 words were chosen to represent short, middle, and long questions, respectively, because they represent the number of words that were observed at the 25th, 50th (median), and 75th percentiles. The first and second columns present the results for respondents higher and lower in cognitive ability, respectively. The last column presents the test statistics for the interaction between cognitive ability and the different question characteristics. An alpha level of $p < .001$ was used to control for the possibility of Type I errors due to the large number of *t*-tests that were conducted.

3.1. Main effects of question difficulty

For higher, as well as for lower cognitive ability respondents, all question difficulty indicators represented in Table 2 are significant predictors of the propensity to say DK. This may in part be a reflection of the high power that resulted from the very large number of respondent-question pairs ($N = 1,104,427$) that contributed to the analysis. Our discussion of the findings will concentrate on the more substantive effects. Consistent with the predictions, all of the question characteristics led to more frequent DKs when they were present in comparison to when they were absent. The most sizable question difficulty predictor of DK was whether or not the question asks for a quantity report. Respondents lower and higher in cognitive ability responded "DK" to a considerable portion of quantity report questions (lower cognitive ability: 9.3%, $B = 1.17$, $t = 466.9$; higher cognitive ability: 7.4%, $B = 1.11$, $t = 367.5$). Another important predictor of the propensity to say

Table 2. Summary of probit analysis for variables predicting the mean percentage of "Don't know" responses

Variable	Lower cognitive ability			Higher cognitive ability			Interaction: low–high	
	M(per cent)	*B (SE B)*	*t*	*M* (per cent)	*B (SE B)*	*t*	*B (SE B)*	*t*
Long: 5 words	3.6	−0.07 (0.001)	−58.2*	2.5	−0.06 (0.012)	−48.2*	−0.01 (0.002)	−3.3
Long: 9 words	3.4			2.3				
Long: 14 words	3.3			2.2				
Instruction: no	2.9	0.26 (0.002)	117.4*	1.9	0.33 (0.003)	126.5*	−0.06 (0.004)	−18.8*
Instruction: yes	4.7			3.6				
Introduction: no	3.1	0.21 (0.004)	51.0*	2.0	0.15 (0.008)	21.5*	0.06 (0.008)	6.8*
Introduction: yes	3.8			2.6				
Ambiguity: no	2.4	0.24 (0.002)	141.2*	1.6	0.21 (0.002)	86.8*	0.04 (0.003)	12.6*
Ambiguity: yes	3.8			2.5				
Retrospective report: no	2.6	0.23 (0.002)	98.2*	1.9	0.11 (0.003)	41.3*	0.12 (0.004)	34.2*
Retrospective report: yes	4.0			2.4				
Frequency report: no	3.1	0.11 (0.02)	5.4*	2.1	0.19 (0.023)	7.7*	−0.08 (0.031)	−2.5
Frequency report: yes	3.8			3.0				
Quantity report: no	1.6	1.17 (0.003)	466.9*	0.9	1.11 (0.003)	367.5*	0.05 (0.004)	13.6*
Quantity report: yes	9.3			7.4				
Scale: no	2.4	1.20 (0.003)	371.9*	1.9	0.75 (0.004)	183.7*	0.44 (0.004)	85.2*
Scale: yes	9.1			3.5				

*$p < .001$.

Table 3. Summary of probit analysis: three-way interactions for particular question difficulty indicators

Variables	Lower cognitive ability			Higher cognitive ability			Interaction: low–high	
	M (per cent)	*B (SE B)*	*t*	*M* (per cent)	*B (SE B)*	*t*	*B (SE B)*	*t*
Interaction quant./scale							−0.27 (0.023)	−10.7*
no quant/no scale	0.7	1.20 (0.003)	371.9*	0.6	0.75 (0.004)	183.6*	0.45 (0.005)	85.2*
no quant/scale	10.4			3.9				
Quant/no scale	9.9	−0.62 (0.016)	39.4*	8.0	−0.77 (0.02)	−41.4*	0.15 (0.25)	6.0*
Quant/scale	3.1			1.5				
Interaction ambiguity/intro							−0.12 (0.012)	−11.9*
not ambiguous/no intro	2.3	0.21 (0.004)	51.2*	1.6	0.15 (0.007)	21.5*	0.06 (0.006)	9.2*
not ambiguous/intro	3.5			2.2				
ambiguous/no intro	3.7	0.06 (0.004)	12.8*	2.4	0.11 (0.005)	23.9*	−0.05 (0.008)	6.7*
ambiguous/intro	4.1			3.1				

*$p < .001$.

DK is whether or not the question requires instructions to be followed (lower ability: $B = 0.26$, $t = 117.4$; higher ability: $B = 0.33$, $t = 126.5$). As indicated by the means, questions that require instructions to be followed are more often answered DK than questions that do not contain instructions (lower ability: 2.9 vs 4.7%; higher ability: 1.9 vs 3.6%). Furthermore, questions containing ambiguous terms are more likely to be answered DK than those questions that are not ambiguous (lower ability, $B = 0.24$, $t = 141.2$; higher ability, $B = 0.21$, $t = 86.8$).

Whether or not a response scale is provided is also an important predictor of the propensity to give DK responses (lower ability: $B = 1.20$, $t = 371.9$; higher ability: $B = 0.75$, $t = 183.7$). Here the adjusted mean percentages show that questions that provide response scales are more likely to be answered DK for both respondents lower and higher in cognitive ability. However, as expected, this difference is qualified by the interaction between "quantity report" and "response scale." Table 3 illustrates the relationship between these two question difficulty indicators separately for respondents lower and higher in cognitive ability. A higher number of DK responses were given for quantity report questions when no response scale is provided. In this case, respondents lower in cognitive ability respond with DK to 9.9% of the questions, and higher ability respondents respond with DK 8.0% of the time. However, when quantity report questions provide a response scale, the percentage of DKs drops to 3.1% for lower and 1.5% for respondents higher in cognitive ability. Additionally, we have expected that introductory phrases would reduce the number of DKs when they serve to clarify ambiguous terms. As can be seen in Table 3, the data do not support this assumption: Questions with ambiguous terms are more likely to be answered DK when they are introduced by a phrase.

3.2. *Interaction between question difficulty and cognitive ability*

A further main purpose of the article is to investigate if respondents lower and higher in cognitive ability are differentially affected by question difficulty. The findings support this assumption. The last section of Table 2 depicts the interaction between cognitive ability and the different question characteristics. Of the eight question characteristics, six reveal significant interactions with cognitive ability. Of these six, five (introductory phrase, ambiguity, retrospective report, quantity report, and scale) are in the predicted direction by showing that lower cognitive ability people are significantly more strongly affected by question difficulty than higher cognitive ability people.[5] For example, questions that provide a response scale evoke a significantly larger increase in the propensity of saying DK among respondents lower in cognitive ability than among respondents higher in cognitive ability (differences in adjusted means of 6.7% for lower ability respondents vs 1.6% for higher; $B = 0.44$, $t = 85.2$). The difference between the mean percentage of DKs for retrospective report questions in comparison to other questions also is more pronounced among respondents lower than among respondents higher in cognitive ability (1.4% vs

[5] The exception is Instruction, which had a larger effect upon the high ($B = 0.33$) than low ($B = 0.26$) cognitive ability respondents. Although the interaction effect is statistically reliable, the effect is small, and the difference in adjusted means is actually larger for the low than high cognitive respondents. The results for this predictor illustrate that the adjusted means have to be interpreted with care. Since the probit model is nonlinear, an examination of the adjusted means alone can provide a misleading indication of effect sizes, whereas the Beta statistics can be consistently interpreted as an accurate reflection of effect sizes.

0.5%; $B = 0.12$, $t = 34.2$). Similarly, for quantity report questions versus other questions, the difference in the mean percentage of DK responses is larger among lower in comparison to higher cognitive ability respondents (7.7% vs 6.5%; $B = 0.05$, $t = 13.6$). The same pattern is found with the absence or presence of an introductory phrase (0.7% vs 0.6%; $B = 0.06$, $t = 6.8$) and ambiguous terms (1.4% vs 0.9%, $B = 0.04$, $t = 12.6$). In sum, these findings support the assumption that respondents lower in cognitive ability are more affected by the difficulty of questions than respondents higher in cognitive ability.

The three-way interaction tests reported in Table 3 present a more equivocal picture concerning the influence of question characteristics on lower and higher cognitive ability respondents. For both three-way interactions, the omnibus inferential tests are significant (Quantity × Scale × Cognitive Ability $B = -0.27$, $t = -10.7$; Ambiguity × Introductory Phrase × Cognitive Ability $B = -0.12$, $t = -11.9$). The remaining interaction tests reported in Table 3 focus on two-way interactions that examine, for respondents lower and higher in cognitive ability, the influence of having a scale or not for questions that do not require quantity reports, and for questions that do require quantity reports, and the influence of having introductory phrases or not for questions that do not have ambiguous (or unfamiliar) terms, and for questions that do have ambiguous terms. As Table 3 reveals, lower cognitive ability respondents were more influenced by question characteristics than higher cognitive ability respondents for two of the four comparisons. Clearly, the strongest of these findings ($B = 0.45$, $t = 85.2$) involves lower cognitive ability respondents, in comparison to higher ability respondents, having been more influenced to increase the propensity of DKs when a scale was used with non-quantity report questions (differences in adjusted means of 9.7% for lower ability respondents vs 3.3% for higher). Lower cognitive ability respondents were also more strongly affected to report DK when introductory phrases were used with questions that were not ambiguous. On the other hand, there was a slight tendency for higher cognitive ability respondents, in comparison to those lower in ability, to less often respond DK when scales were used for quantity report questions, and to more often respond DK when introductory phrases were used for questions that had ambiguous terms.

4. Discussion

This article examined how the quality of the survey data (indicated by the number of DKs) varies with question difficulty and older respondents' cognitive ability. Overall, the analysis illustrates that respondents' cognitive ability moderates the effect of question difficulty. Although the overall number of DKs was small, it varied systematically with cognitive ability and question difficulty: Respondents lower in cognitive ability gave more DK responses to difficult than to easy questions. For respondents higher in cognitive ability, the differences between difficult and easy questions were less pronounced. The interaction between question difficulty and cognitive ability is observed for indicators representing all stages of the question answering process. Specifically, respondents lower in cognitive ability are more affected than respondents higher in cognitive ability by introductory phrases, ambiguous questions, questions requiring retrospective or quantity reports, and questions that do provide response scales. As predicted, although not statistically significant, lower cognitive ability respondents also profit more from

longer questions (controlling for variations in all other question characteristics) than higher cognitive ability respondents. Altogether, the findings are consistent with our theoretical assumptions.

4.1. *Unexpected findings*

Although the general trends of our results agree with our hypotheses, a few unexpected findings emerged. Consistent with the predictions, the use of scales did reduce the tendency for DKs with quantity report questions, that is, quantity report questions were more difficult when they had to be answered without the aid of the additional information that a scale provides (cf., Schwarz 1990). However, contrary to prediction, lower cognitive ability respondents did not benefit more from the scales in comparison to higher ability respondents (in fact, there was a slight tendency in the opposite direction). But response scales increased the propensity for DKs for questions that did not ask for a quantity report particularly among respondents lower in cognitive ability. It can be speculated that this latter effect is driven by the use of scales in questions that require difficult judgments or which may even exacerbate formatting problems. For example, in one particular section of the survey, respondents are asked for a number between 0 and 100 to gather estimates on respondent's expectations concerning the likelihood of the occurrence of future hypothetical events. These kinds of judgments may be difficult in general. Moreover, the provided response scales (numbers from 0 to 100) do not provide any information respondents could use to simplify the answering process, and can actually increase formatting demands. Finally, the findings also did not support the expected benefit from what we are calling ''introductory phrases'' in reducing DKs when ambiguous terms are present. Any interpretation here is difficult. On the one hand, introductory phrases may simply be doing more harm than good. On the other hand, they may have been primarily used by the survey researchers to soften particularly difficult questions, and thus their benefits may be masked in that they may be reducing a propensity for DKs that would be more pronounced without their use.

4.2. *Limitations*

Limitations of our analysis compromise the interpretation of expected and unexpected findings alike. Since questions could not be randomly assigned to respondents lower and higher in cognitive ability, there are potential confounds of question difficulty with cognitive ability. The possibility of self-selection effects cannot be ruled out. That is, since cognitive ability is associated with other factors, skip patterns in the survey may have resulted in biases of more difficult questions associated with particular question characteristics being asked of those respondents lower in cognitive ability, or vice versa. As examples, confounding through skip patterns may have led to more difficult retrospective report questions being asked to lower cognitive ability respondents, and with quantity report questions, those that included scales may have been of a more difficult variety when asked to higher cognitive ability respondents.

Additionally, there may be an interaction between mode of the interview (face to face vs phone) and the observed effects. For example, in face to face interviews it is easier for interviewer and interviewee to communicate and resolve (verbally and nonverbally)

comprehension difficulties (Herzog and Rodgers 1988). However, as described earlier, a direct examination of interactions with mode was not feasible because a considerable portion of the AHEAD respondents self-selected themselves into their preferred mode of administration. By not examining mode we recognize that if face to face interviewing did assist those lower in cognitive ability, that any effect associated with cognitive ability would emerge from a fairly conservative test, leading us to be even more confident in the reliability of our results. Experiments, using more controlled conditions, would help to substantiate our findings regarding the interaction between particular question characteristics and respondents' cognitive abilities, and could further explore any effects associated with the mode of the interview.

4.3. Implications for questionnaire design

With the discussed limitations in mind, some implications of the findings can be directly used for recommendations on redesigning and simplifying questions. Specifically, there is indication that quantity reports can be assisted by providing response scales that serve as recall and formatting aids. Results also indicate that there may be dangers in the use of instructions. Such questions could be broken down into several questions to make them easier to answer. In designing questions, researchers should be aware to select terms and phrases which can easily be understood by both higher and lower cognitive ability subjects. Some of the problems might also be remedied if interviewers were trained to give lower cognitive ability respondents enough time to provide answers to complex survey questions. Lower cognitive ability individuals are probably under-represented in pretests conducted in the developmental stages of a survey interview. However, to identify possibly difficult questions, the pretesting of survey questions should be deliberately done with both *lower and higher* cognitive ability subjects. This is particularly recommended if the target population has a high probability of a large number of individuals with lower cognitive ability (e.g., surveys of the elderly).

4.4. Conclusions

The reported analysis suggest that particular questions elicit a considerably high number of DKs among a subsample of respondents: Older respondents lower in cognitive ability are more affected by question difficulty than older respondents higher in cognitive ability. This effect can lead to biased survey findings because responses from lower cognitive ability respondents would be underrepresented for topics that are assessed by more complex questions. These, however, are often exactly the topics for which unbiased information from the mainly affected subpopulation is needed (e.g., information about medical conditions and expenditures of those in poor health).

5. References

Andrews, F.M. and Herzog, A.R. (1986). The Quality of Survey Data as Related to Age of Respondent. Journal of the American Statistical Association, 81, 403–410.

Belson, W.A. (1981). The Design and Understanding of Survey Questions. London: Gower.

Brandt, J., Spencer, M., and Folstein, M. (1988). The Telephone Interview for Cognitive Status. Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 1, 111–117.

Campbell, A., Converse, P.E., and Rogers, W.L. (1976). The Quality of American Life. New York: Sage.

Cannell, C.F., Marquis, K.H., and Laurent, A. (1977). A Summary of Studies of Interviewing Methodology. Vital and Health Statistics, Series 2, No. 69.

Cannell, C.F., Miller, P.V., and Oksenberg. L. (1981). Research on Interviewing Techniques. In Sociological Methodology, S. Leinhardt (ed.), San Francisco: Jossey-Bass.

Clark, H.H. and Schober, M. (1992). Asking Questions and Influencing Answers. In Questions about Questions: Inquiries Into the Cognitive Bases of Surveys, J.M. Tanur (ed.). New York: Russell Sage Foundation.

Colsher, P.L. and Wallace, R.B. (1989). Data Quality and Age, Health and Psychobehavioral Correlates of Item Nonresponse and Inconsistent Responses. Journal of Gerontology, Psychological Sciences, 44, P45–52.

Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions. New York: John Wiley.

Folstein, M.F., Folstein, S.E., and McHugh, P.R. (1975). Mini-Mental State: A Practical Method for Grading the Cognitive State of Patients for the Clinician. Journal of Psychiatry Research, 12, 189–198.

Gergen, K.J. and Beck, K.W. (1966). Communication in the Interview and the Disengaged Respondent. Public Opinion Quarterly, 44, 223.

Herzog, A.R. and Dielman, L. (1985). Age Differences in Response Accuracy for Factual Survey Questions. Journal of Gerontology, 40, 350–357.

Herzog, A.R. and Rodgers, W.L. (1988). Interviewing Older Adults. Mode Comparison Using Data from a Face to Face Survey and a Telephone Survey. Public Opinion Quarterly, 52, 84–99.

Herzog, A.R. and Wallace, R.B. (1995). Measures of Cognitive Functioning in the AHEAD Study. Journal of Gerontology, Series B, Psychological and Social Sciences, 52B (Special Issue), 37–48.

Juster, T. and Suzman, R. (1995). An Overview of the Health and Retirement Study. Journal of Human Resources, 30, S7–S56.

John, D.R. and Cole, C.A. (1986). Age Differences in Information Processing: Understanding Deficits in Young and Elderly Consumers. Journal of Consumer Research, 13, 27–35.

Kaldenberg, D.O., Koenig, H.F., and Becker, B.W. (1994). Mail Survey Response Rate Patterns in a Population of the Elderly: Does Response Deteriorate with Age? Public Opinion Quarterly, 58, 68–76.

Knäuper, B. (1997). The Impact of Age on Response Order Effects in Attitude Measurement. (Manuscript under review).

Knäuper, B. and Wittchen, H.-U. (1994). Diagnosing Major Depression in the Elderly: Evidence for Response Bias in Standardized Diagnostic Interviews? Journal of Psychiatric Research, 28, 147–164.

Kogan, N. (1961). Attitudes Towards Old People in an Older Sample. Journal of Abnormal and Social Psychology, 62, 616.

Krosnick, J.A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. Applied Cognitive Psychology, 5, 213–236.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, 2nd Edition. Monographs on Statistics and Applied Probability 37, 98–148. New York: Chapman and Hall.

Rodgers, W.L., Herzog, A.R., and Andrews, F.M. (1988). Interviewing Older Adults, Validity of Self-Reports of Satisfaction. Psychology and Aging, 3, 264–272.

Salthouse, T.A., Babcock, R.L., and Shaw, R.J. (1991). Effects of Adult Age on Structural and Operational Capacities in Working Memory. Psychology and Aging, 6, 118–127.

Schwarz, N. (1990). Assessing Frequency Reports of Mundane Behaviors: Contributions of Cognitive Psychology to Questionnaire Construction. In Research Methods in Personality and Social Psychology (Review of Personality and Social Psychology, Vol. 11), C. Hendrick and M.S. Clark (eds.), Beverly Hills, CA: Sage.

Schwarz, N. (1994). Judgment in a Social Context: Biases, Shortcomings, and the Logic of Conversation. Advances in Experimental Social Psychology (Vol. 26). San Diego, CA, Academic Press.

Tourangeau, R. (1984). Cognitive Sciences and Survey Methods. In Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines, T. Jabine, M. Straf, J. Tanur, and R. Tourangeau (eds.). Washington, DC: National Academy Press.

Tourangeau, T. and Rasinski, K.A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. Psychological Bulletin, 103, 299–314.

Willis, G.B., Royston, P., and Bercini, D. (1991). The Use of Verbal Report Methods in the Development and Testing of Survey Questionnaires. Applied Cognitive Psychology, 5, 251–267.

Received October 1995
Revised August 1996

Printed in September 1997