RELIABLITY OF RESPONSES IN QUESTIONNAIRE RESEARCH WITH CHILDREN

('Coding scheme: a technical report' : appended)

N.Borgers[1]
University of Amsterdam
J.J. Hox
Utrecht University

Children are no longer neglected as respondents in large-scale surveys. Researchers are convinced that information about perspectives, attitudes, and behaviors of children should be collected from the children themselves. Regarding adults, empirical evidence shows that respondent characteristics as well as question characteristics affect response quality. Because children are in process of developing their cognitive and social skills, it is expected that answering questions in surveys brings along additional problems concerning response quality. So far however, methodological expertise on surveying children is scarce, and researchers rely on ad-hoc knowledge from fields such as child psychiatry and educational testing, or on methodological knowledge on surveying adults.
In this article we report results of a secondary analysis on several data sets. The multi level meta-analysis was directed on the (interaction) effects of child characteristics and question characteristics on the reliability of responses.

Key words: special population, question characteristics, respondent characteristics; response quality, survey resaerch

[1] Corresponding author: Natacha Borgers
Faculty of Social and Behavioral Sciences, University of Amsterdam
Wibautstraat 4
NL-1091 CM Amsterdam
The Netherlands
tel. +31 20 5251526
fax +31 20 5251200
email: nborgers@educ.uva.nl

1 INTRODUCTION

For general surveys, procedures to enhance response quality are well documented (Biemer, Groves, Lyberg, Mathiowetz, & Sudman, 1991; Groves, 1989; Lyberg et al., 1997). However, surveying an adult population is far from simple. Even slight variations in question wording affects responses. There is an increasing body of empirical evidence that both respondent characteristics and question characteristics affect the reliability of responses in surveys (Krosnick, 1991; Krosnick & Fabrigar, 1997; Schwarz & Hippler, 1995). However, there is still a lack of methodological knowledge on procedures to enhance response quality in survey research with children, whereas survey researchers increasingly collect perspectives, attitudes and behaviors from the children themselves (Scott, 1997).

Krosnick (1991) utilizes a satisficing theory to explain why the reliability of responses differs between respondents, and why it can be affected by question wording. The satisficing theory elaborates on a standard *question answering process-model* developed by Tourangeau (1988) (cf.Cannel, Miller, & Oksenberg, 1990; Krosnick, Narayan, & Smith, 1996; Schwarz, Knäuper, & Park, 1998; Sudman, Bradburn, & Schwarz, 1996). Four steps characterize an optimal question answering process:

1. Understanding and interpreting the question being asked
2. Retrieving the relevant information from memory
3. Integrating this information into a summarized judgment
4 Reporting this judgment by translating it to the format of the presented response scale.

According to this satisficing theory there are three factors that affect the question answering process. The first is the motivation of the respondent to perform the task, the second is the difficulty of the task, and the third is the respondent's cognitive ability to perform the task. Besides, satisficing theory identifies two processes that explain differences in reliability of responses, namely *optimizing* and *satisficing*. Optimizing means that the respondent goes through all the four cognitive steps needed to answer a survey question. Contrary to optimizing, satisficing means that a respondent gives more or less superficial responses that however, appear reasonable or acceptable to the researcher, without going through all the steps involved in the question-answering process. Satisficing is related to the motivation of the respondent, the difficulties of the task and the cognitive abilities of the respondent. Difficult questions, low cognitive abilities, and low motivation may lead respondents to

provide a satisfactory response instead of an optimal one. Using a satisficing strategy may lead to less reliable responses than using an optimizing strategy.
Implicitly, the satisficing theory assumes an interaction effect between respondent characteristics and question characteristics, which can be described as follows: the less cognitively sophisticated respondents are, the more sensitive to difficult or cognitively demanding questions they will be, and the less reliable their responses will be. The few projects that reported on the interaction between respondent and question characteristics presented empirical evidence for this effect (Borgers, Leeuw, & Hox, 1999; Borgers, Leeuw, & Hox, 2000; Knäuper, Belli, Hill, & Herzog, 1997; Marsh, 1986; Schwarz et al., 1998). These studies showed that the less cognitively sophisticated respondents, elderly or young children gave less reliable responses on more difficult questions than the more sophisticated respondents.

In the last decade methodological knowledge on the effects of reduced cognitive ability on response reliability, increased. This knowledge is mainly based on research results reporting special populations in survey research, such as the effects of the cognitive decline of elderly (Herzog & Rodgers, 1992; Knäuper et al., 1997). Several studies show that reduction in cognitive functioning is associated with a decline in the reliability of responses (Alwin & Krosnick, 1991; Knäuper et al., 1997; Krosnick, 1991; Schwarz et al., 1998).
Like ageing, growing up involves changes in cognitive functioning. Within children cognitive, communicative and social skills are developing. As a consequence, cognitive ability varies considerably across children. These differences across children can result in the use of different strategies in answering questions and by that, differences in the reliability of responses. It is reasonable that answering questions in surveys brings along additional problems concerning response quality. However, methodological knowledge on how to survey children is still scarce, while this kind of data collection displays a rapid growth (Scott, 1997). The few studies that are known (Borgers, 1998; Leeuw & Otter, 1995; Otter, Mellenberg, & Glopper, 1995; Otter, 1993) support the hypothesis that growing up is related to an increase of the reliability of responses. Therefore it is necessary to increase the methodological knowledge on response quality in survey research with children.
This study investigates the effects of child and question characteristics on the reliability of children's responses in self-administered questionnaires. Besides, the interaction effect between both characteristics on the reliability of responses, which follows from satisficing theory, will be researched. Based on satisficing theory and empirical results, hypotheses can

be formulated for the effect of child characteristics and question characteristics on the reliability of responses. Less cognitively sophisticated children (youngest, less years of education) produce less reliable responses than the more cognitively sophisticated children. An overview of the hypothesized direction of the expectations for each question characteristic is given in Table 1. Besides their operationalization of the question characteristics are given in this table. The question characteristics are categorized according to the step of the question answering process model, were the concerned characteristic cognitively appeal to.

Table 1: Summary of the question characteristics, direction of the hypotheses and operationalization

Comprehension and interpretation of the question	**direction of the hypothesis[2]**	**Operationalization**
Question length	negative	number of words
	negative	number of sentences
Length of the introductory text	positive	number of words/100
Readability	positive	comprehensive readability, high score = easy to read
	positive	technical readability, high score = easy to read
Ambiguity	negative	of the question
	negative	of the response scale
Double barreled	negative	double versus single
Complex constructions	negative	complex versus simple
Negatively formulated question	negative	negative versus positive
Kind of information being asked	no direction	attitudes experiences opinions behavior attributions capacities
Retrieving relevant information from memory		
Complexity of the question	negative	complex versus simple
Reference period	positive	reference period versus no reference period
Numerical quantity	negative	numeric versus not numeric response
Judging the retrieved information		
Subjective question threat, sensitivity	positive	sum of 4 indicators[3]
Balance of the question	positive	balance versus unbalanced
Position in the questionnaire	negative	1st versus 2nd versus 3rd position in the questionnaire[4]
Communicate the final response		
Number of response categories	negative	2, 3, 4 ,5, 7 and 10 categories
Offering midpoints	positive	midpoint versus no midpoint offered
Offering Don't know filter	positive	don't know filter versus no don't know filter
Scale labels	positive	labeled versus partly labeled response scale

[2] negative: results in low reliability; positive: results in higher reliability

[3] 1: To personal for the respondent; 2: To threatening for the respondent; 3: Rather not answer the question; 4: Hard to give an honest answer to the question

[4] Every questionnaire was divided in three parts, with an equal amount of questions

The research question is '*What are the effects of child characteristics and question characteristics (and their interaction effects) on the reliability of responses produced by children in self-administered questionnaire research?*'

The respondent characteristics as mentioned in the satisficing theory, motivation and cognitive functioning, are difficult to measure directly within a survey. Thus, most attention is devoted to measurable characteristics of the respondent, such as age and education (c.f.Alwin & Krosnick, 1991; Andrews & Herzog, 1986; Krosnick & Alwin, 1987; Rodgers, Andrews, & Herzog, 1989). Age and education are then used as proxy variables for cognitive functioning. Following the known studies on response quality in survey research we used measurable child characteristics, such as age and years of education, as proxy variables for cognitive abilities. In addition we used sex of the children, because boys and girls develop in different phases at the school age. Next to child characteristics, 24 question characteristics such as, number of words in the introductory text, ambiguity of the question and the response scale and the number of response categories were coded. All question characteristics measure an aspect of cognitive demandingness of the question. Most question characteristics follow from research with adults (for and overview, seeKrosnick & Fabrigar, 1997). Concerning children, question characteristics, such as ambiguity of the question, complexity of the question and negatively formulated questions appear to have a negative effect on the reliability of responses (Benson & Hocevar, 1985; Leeuw & Otter, 1995; Marsh, 1986; Otter et al., 1995; Otter, 1993).

2. METHOD

2.1 Data sets

We use secondary analysis on five different data sets, collected in the field of educational research. These data sets consist of the responses of school children to written self-administered questionnaires, administered in class in the course of different educational research projects in the Netherlands and Belgium. The data sets are briefly described in Appendix 1[5]. Data are based on 13 questionnaires, which contain 51 different scales and 513 questions that were answered by 4644 children in total. Our sample is between 8 and 18 years of age ($\overline{X}$ = 12.1; s.d. = 2.2). For all the children the number of years of education was known,

[5] A full description is available from the first author

which is varying from 3 to 14 years of education ($\overline{X}$ = 8, s.d. = 2.2). The total data set contained 1911 boys and 1912 girls, and 821 children for which sex is not recorded in the data file.

2.2 Variables and Coding scheme

Three types of explanatory variables are distinguished:

a) Child characteristics
b) Scale characteristics
c) Question characteristics

ad. a) The child characteristics in the data sets that can be used as proxy variables for cognitive functioning, are age, number of years of education and sex of the children.
ad. b) Only scale characteristic is used, the number of items per scale.
ad. c) To code each of the 24 the question characteristics (see Table 1) a computerized coding scheme[6] is developed (Borgers, 1997). The question characteristics, number of words in the introductory text, question length and both readability indices were coded by one coder. Two coders coded all other question characteristics because these characteristics are to some degree subjective, e.g. question ambiguity. By combining the results of multiple coders, the reliability of the composite rating can still be satisfactory (Stock, 1994). In case of disagreement, the final code is the mean rating of both coders. This was possible in all cases except for the, nominal characteristic, information that is asked for. In case of disagreement a third coder was assigned, who was decisive for the final code. For each characteristic the intercoder reliability (Cohen's kappa) was determined. If the intercoder reliability is lower than 0.70, again a third coder is assigned, and the final code is the mean rating of all coders. For all question characteristics the intercoder reliability (Orwin, 1994) was in above 0.86, except for the four indicators for subjective question threat. However, these four indicators are combined into one rating. Consequently the reliability of the combined ratings[7] is used

[6] The full-computerized coding scheme (in Dutch) is available from the first author.

[7] $$\alpha_{sum} = 1 - \frac{\sum \sigma_j^2 - \sum \sigma_j^2 \alpha_j}{\sum \sigma_j^2 + 2 \sum_{k>j} \sum \sigma_j^2 \sigma_k^2 r_{jk}}$$

(Guilford, 1954). This results in an intercoder reliability of 0.97, which is sufficient for our purpose. A summary of all intercoder reliabilities is given in Appendix 2.

The three categorical variables (position in the questionnaire, number of response categories and the kind of information that is being asked for) in the inventory of the question characteristics are represented by dummy variables in the analysis. Position is represented by two dummy variables: one for the second and one for the third position in the questionnaire. The first position is the base against which the other two positions are compared. With respect to the number of response options, two response options is the base against which the remaining five options are compared. The last categorical variable concerns the type of information that is being asked for. For this variable we computed six dummy variables with attitude questions as the baseline: opinion, behavior, attribution, capacity, experiences and empathy.

Two measures of reliability were used as dependent variables (Borgers, 1997). The first indicator is the reliability measured at the scale level, for which Cronbach's alpha was chosen. The second indicator is measured at item level, being the item-rest correlation. Cronbach's alpha as well as item-rest correlation cannot be defined for individual children; they are only defined for groups of children. In order to compare the reliability between several age categories, years of education categories and both sexes, groups of children had to be constructed. However, the groups could not be constructed by combining all three child characteristics. The result would have been too many groups with too few children in the different groups. Therefore three separate groupings were constructed, based on age of the children, years of education and gender. For each group of children Cronbach's alpha was computed for every scale in the questionnaire they answered. Likewise, for each group of children item-rest correlations were computed for every question. Therefore, in the analysis with Cronbach's alpha, the mean codes for the question characteristics of the scale were computed, while the individual codes were used for the analysis with item-rest correlation

The indicators Cronbach's alpha and item-rest correlation do not have a normal distribution and both have a limited range ($\alpha \leq 1$ and $-1 \leq r_{ir} \leq 1$). In regression analysis estimates can be found that are out of this limited range. To achieve normality and overcome the problems caused by the limitations of the dependent variables transformations were used.
For Cronbach's alpha a transformation proposed by Hakstian (1976) is used.

We use $\alpha^* = 1-(1-\alpha)^{\frac{1}{3}}$ as dependent variable,

and the standard error of alpha star: $se(\alpha^*) = \sqrt{\sigma^2} = \sigma$

The quantity $(1-\alpha)^{\frac{1}{3}}$ is normal distributed $(N(\mu, \sigma^2))$

with $\mu = c^{-1}(1-\alpha)^{\frac{1}{3}}\mu^*$ and $\sigma^2 = 2Jc^{-1}(1-\alpha)^{\frac{2}{3}}[9(n-1)(J-1)]^{-1}$

were $c = \frac{9n-11}{9n-9}$ and $\mu^* = 1-\frac{2}{9(n-1)(J-1)}$

and were n = sample size and J = number of items

For item-rest correlation (r_{ir}): we used the Fisher-Z transformation (Hayes, 1974)

$Z = \frac{1}{2}(\ln(1+r_{ir}) - \ln(1-r_{ir}))$ with the standard error = $\sqrt{\frac{1}{n-3}}$

were n = sample size.

2.3 Analysis

The analysis model is a meta-analysis with the transformed alpha stars and Fisher's Z as dependent variables. The use of a meta-analysis has the advantage that both coefficients are analyzed, using the information available on their standard errors. We used a hierarchical linear model or multilevel model for meta-analysis. Bryk (1992) describes a two level model for meta-analysis, which has been used by Leeuw (1995). However, the hierarchy of our data is more elaborated. First, there are three levels in our analysis; *questions* which are nested in *questionnaires*, and *questionnaires* which are nested within groups of *respondents*.[8] To analyze data with such a hierarchical structure, multilevel analysis is appropriate. In this model, questions are the lowest level. Secondly, the hierarchical structure of the two highest levels is in fact an oversimplification. Questionnaires are not really nested within groups of respondents and respondents are not nested within questionnaires. This kind of structure should be treated as a cross-classified structure. In our model questions were nested within

[8] In fact there are four levels in the analysis with Fisher- Z transformed item-rest correlations: *questions* which are nested in *scales*, *scales* which are nested in *questionnaires*, and *questionnaires* which are nested within groups of *respondents*. However, there is no variance at the scale level so we treated the data as if there were only three levels.

within the cross-classification of questionnaires and groups of children (c.f. Goldstein, 1995; Rasbash et al., 1999).

Our strategy of analysis can be divided into 2 steps. The first step was running the *intercept only model* for each group separately for both measures of reliability. These models distinguish the variances in the dependent variables that can be attributed to the different levels. The results are presented in Table 4 and Table 5. Secondly, we included all explanatory variables, the child characteristics, scale characteristics and all 24 question characteristics. The following step was to model these coefficients as varying across groups of children. We expect different effects of the question characteristics across children, and varying slopes for question effects indicate interaction effects for these variables with some child characteristic. However, there were no significant random coefficients. That is why we finally report the results of the second step. This resulted in two times three tables with regression coefficients and their standard error[9]. For alpha star as dependent variable we had a table for the results of the different age groups, the different sex groups and for the different years of educations. The same yields for the Fisher-Z transformed item-rest correlation. The results for both measures of reliability gave comparable but not identical estimates. Finally, we combined the results for the alpha star into one table, and for the Fisher-Z transformed item-rest correlation into one table. We averaged the estimates for the three groups of children. The mean p-values were calculated by the mean z-score of the coefficients. These results are presented in Table 6 and Table 7.

[9] The tables of the separate analysis are available from the first author.

3 RESULTS

To give insight in the distribution of the groups Table 2 and 3 give an overview of the different groups that are constructed for our analysis. The different ages and years of education that are included in our analysis are given and the number of scales that are answered by these different groups.

Table 2 The values of the independent (child characteristics) variables and the number of scales answered by these groups for alpha star as dependent variable

age												
age	6	7	8	9	10	11	12	13	14	15	16	Total
number of scales	1	1	1	4	5	6	8	3	4	4	4	41
years of education												
education	3	4	5	6	7	8	9	10	11	12		Total
number of scales	1	3	1	3	7	7	3	3	4	4		36
gender												
sex	boys				girls							Total
number of scales	12				12							24

Table 3 The values of the independent (child characteristics) variables and the number of scales answered by these groups for Fisher-Z transformed item-rest correlation as dependent variable

age												
age	6	7	8	9	10	11	12	13	14	15	16	Total
number of scales	1	1	1	4	4	6	9	3	4	4	4	41
years of education												
education	3	4	5	6	7	8	9	10	11	12		Total
number of scales	1	3	1	3	7	7	3	3	4	4		36
gender												
sex	boys				girls							Total
number of scales	12				12							24

Both tables show that the youngest groups are somewhat underrepresented while all the other groups are well and almost equally represented in our analysis. However, every constructed group consists of at least 45 children.

Table 4 and 5 present the variances and their standard errors that can be assigned to the different levels as a result of the intercept only model (c.f.Hox, 1995). The lowest level (question) variance is not included in both tables. In a hierarchical linear model for meta-analysis the lowest level variance is known but can change as a result of differences in sample sizes. Besides the percentage variance at both levels (questionnaire and child) is given.

Table 4 Variances in alpha star divided for questionnaire and child level for age, years of education and gender

	age (s.e.) percentage variance		years of education (s.e.) percentage variance		gender (s.e.) percentage variance	
Variance at questionnaire level	.03 (.01)	60%	.03 (.01)	50%	.03 (.01)	60%
Variance child level	.02 (.00)	40%	.03 (.00)	50%	.02 (.00)	40%

Table 5 Variances (standard error) in Fisher-z transformed item-rest correlation divided for questionnaire and child level for age, years of education and gender.

	age (s.e.) percentage variance		years of education (s.e.) percentage variance		gender (s.e.) percentage variance	
Variance at questionnaire level	.03 (.01)	60%	.03 (.01)	50%	.03 (.01)	60%
Variance child level	.02 (.00)	40%	.03 (.00)	50%	.02 (.00)	40%

For alpha star as well as for Fisher-Z transformed item-rest correlation the variance at both levels is almost equally divided in all the three child characteristics.

Table 6 presents the results of the combined multi level analyses for meta-analysis for alpha star. The regression coefficients are given in the first column and the one-tailed p-values in the second. The effect of information asked has been tested two-tailed. In the third column the results of the back-transformed regression coefficients into Cronbach's alphas[10] are shown. There are two important things that should be realized whilst interpreting these results. First, these coefficients are unstandardized. That is why the standardized beta coefficients are

presented in the last column. Second, the question characteristics are the mean ratings over scales.

Table 6 Results of the multi level meta-analysis for the effects of child and mean question characteristics on alpha star

	alpha star	p-value (1-tailed)	alpha	beta
Intercept	**-.328**	**.05**	**-1.342**	
Child characteristics				
Years of education	**.021**	**.00**	**.062**	**.045**
Age	**.016**	**.00**	**.047**	**.036**
Gender	**.013**	**.01**	**.038**	**.007**
Mean question characteristics per scale				
Number of words in the introductory text per 100 words	**.047**	**.00**	**.134**	**.084**
Number of words in question	-.004	.06		
Number of sentences in the question	.022	.63		
Douma (readability index)	**.003**	**.00**	**.009**	**.047**
Cilt (readability index)	.001	.27		
Ambiguity of the question	-.004	.29		
Ambiguity of the response scale	**-.171**	**.05**	**-.606**	**-.078**
Double barreled	.000	.72		
Complex construction	.208	.99	**.503**	**.073**
Negative formulated	.011	.80		
Complexity of the question	-.013	.47		
Reference period	**.064**	**.01**	**.180**	**.020**
Numeric	**-2.947**	**.00**	**-60.490**	**-.026**
Balance	.012	.70		
Midpoint	**-.095**	**.00**	**-.313**	**-.152**
Don't know filter	**.122**	**.00**	**.323**	**.058**
Label	**.270**	**.00**	**.611**	**.091**
Sensitivity	**.206**	**.00**	**.499**	**.032**
Position in the questionnaire				
Second	**.076**	**.00**	**.211**	**.033**
Third	**.119**	**.00**	**.316**	**.034**
Number of response options				
3	.072	.13		
4	**-.337**	**.00**	**-1.390**	**-.126**
5	**-.032**	**.05**	**-.099**	**-.016**
7	-.003	.65		
Information that is being asked for				
Opinion	**.044**	**.00**[a]	**.126**	**.009**
Behavior	.129	.56[a]		
Attribution	-.025	.56[a]		
Experiences	**-.016**	**.05**[a]	**-.049**	**-.005**

[a] two tailed p-value

Table 7 shows the combined results for the Fisher-Z transformed item rest-correlation. Again in the first column the regression coefficients are given and in the second the one-tailed p-

[10] Inverse of alpha star: $\alpha = 1-(1-\alpha^*)^3$

values, for the information asked the two-tailed p-values are given. In the third column the results of the back-transformed regression coefficients into item-rest correlations[11] are given. The last column presents the standardized beta coefficients.

[11] Inverse fisher-Z transformation: $r_{ir} = (\exp(2 * Z) - 1) / (\exp(2 * Z) + 1)$

Table 7 Results of the multi level meta-analysis for the effects of child and question characteristics on Item-rest correlations

	Fisher-Z transformed item-rest correlation	p-values (1-tailed)	Item-rest correlation	Beta
Intercept	**.253**	**.02**	**.248**	
Child characteristics				
Years of education	**.003**	**.01**	**.003**	**.006**
Age	**.037**	**.00**	**.037**	**.080**
Gender	**.023**	**.01**	**.023**	**.012**
Scale characteristics				
Number of items in the scale	**.019**	**.00**	**.019**	**.171**
Question characteristics				
Number of words in the introductory text per 100 words	**.050**	**.00**	**.050**	**.080**
Number of words in question	**-.003**	**.00**	**-.003**	**-.041**
Number of sentences in the question	-.000	.50		
Douma (readability index)	**.001**	**.01**	**.001**	**.016**
Cilt (readability index)	-.001	.88		
Ambiguity of the question	.015	.89		
Ambiguity of the response scale	-.030	.46		
Double barreled	.012	.20		
Complex construction	.039	.99	**.039**	**.016**
Negative formulated	**-.113**	**.00**	**-.113**	**-.033**
Complexity of the question	-.004	.35		
Reference period	**.055**	**.00**	**.055**	**.022**
Numeric	**-.626**	**.00**	**-.555**	**-.017**
Balance	.027	.32		
Midpoint	-.130	.06		
Don't know filter	-.089	.08		
Label	.077	.15		
Sensitivity	**.104**	**.00**	**.104**	**.025**
Position in the questionnaire				
Second	**.032**	**.00**	**.032**	**.016**
Third	**.067**	**.00**	**.067**	**.030**
Number of response options				
3	**-.425**	**.00**	**-.401**	**-.093**
4	**-.306**	**.00**	**-.297**	**-.093**
5	-.126	.11		
7	**-.261**	**.01**	**-.255**	**-.097**
10	.061	.32		
Information that is being asked for				
Opinion	**-.027**	**.10**[a]	**-.027**	**-.009**
Behavior	-.017	.28[a]		
Attribution	**.095**	**.10**[a]	**.095**	**.021**
Capacity	**.094**	**.02**[a]	**.094**	**.009**
Experiences	**.059**	**.00**[a]	**.059**	**.027**
Empathy	.017	.68[a]		

[a] two tailed p-value

For all the three child characteristics the effects on alpha star as well as on the Fisher-Z transformed item-rest correlation are significant. Older children produce more reliable responses than younger children. The same result holds for years of education: the longer

children took education the more reliable their responses are. Besides, girls produce more reliable responses than boys. Concerning the effects of question characteristics, the two measures of reliability show results in the same direction.

The number of words in the introductory text has a positive effect: the more words used in the introductory text the more reliable the responses will be. The comprehensive readability index also has a positive effect on both measures, as do questions with a given reference period in the questionnaire and sensitive questions. Negative effects in conformity with our hypothesis are found for questions that are asking for a numeric quantity as response and the effect of number of response options.
Opposite to our expectations are the positive effects of the position of questions in the questionnaire, and the existence of complex constructions in the question.

Next to these common effects there are some effects, which only yield for alpha or only for the item-rest correlation. Concerning alpha there are positive effects of fully labeled response categories as there is for the use of don't know filters. Negative effects are found for the use of an ambiguous response scale and for the use of midpoints. The kind of information that is being asked for also shows opposite effects for both measures, some positive and some negative effects. Questions that invoke for opinions have a positive effect on Cronbach's alpha while they have a negative effect on item-rest correlation. For experiences the effects are exactly the opposite, a negative effect on Cronbach's alpha and a positive effect on item rest-correlations. Questions that ask for attributes and questions that ask for capacities do have a positive effect on item-rest correlations. Negatively formulated questions do have negative effects on item-rest correlations while they do not have an effect on alpha.

4. DISCUSSION

In this article it has been shown that both child characteristics as well as question characteristics affect the reliability of responses of children in self-administered questionnaire research. However, the hypothesized interaction effects between child and question characteristics failed to appear in this study.
The effects of child characteristics are unequivocal. Younger children, the less cognitive sophisticated respondents, produce less reliable responses than older children. Besides, girls

give more consistent responses than boys. This supports the hypothesis that reliability increases with cognitive level. We can extend this conclusion to a more general one by including earlier results on item non-response (Borgers & Hox, 1999); data quality increases with cognitive level.

Although, the results for the question characteristics are not quite unequivocal they provide indications to improve questions and should be considered when designing questionnaires for children. Using a clear and extensive introductory text in a questionnaire improves reliability of responses, as do questions with a high readability index and questions that ask for a numeric quantity.

As expected, sensitive questions do have a positive effect on the reliability of responses. Besides, they produce less item non-response for the youngest children (Borgers & Hox, 1999), possibly as a result of their involvement in the topic of the question. This can result in the use of a optimizing strategy for responding to these kinds of questions. Another explanation could be that younger children are more sensitive to social desirability. Young children are very suggestible. They are often reluctant to express their own thoughts and feelings because they assume that adults know everything already and in addition they are afraid to say something wrong or foolish (Maccoby & Maccoby, 1954). This can cause consistent or reliable responses, but not their own thoughts or feelings.

The use of a reference period in a question produces more reliable responses. Such a reference can work as 'anchor points' for children. Young children (7-11) are very literal in the interpretation of words. These kinds of questions are very literal in what they ask and for which period. The same train of thought can applied to the use of labeled response options for children.

In conformity with the results of research with adults the use of *don't know filters* with children increases the reliability of responses. However *don't know filters* increase unusable responses. Furthermore Krosnick (1999, p.20) stated that the vast majority of *No responses* are not due to completely lacking an attitude. *Don't know filters* discourage respondents to report their opinion. For that reason it is not recommended to use explicit *don't know filters* in questionnaires despite the result that it increases the reliability of responses.

Some question characteristics should be avoided in questionnaires for children. The first is negatively formulated questions. This result is in concordance with satisficing theory

(Krosnick, 1991) and empirical results with adults (Andrews, 1984; Knäuper et al., 1997) and children (Benson & Hocevar, 1985; Marsh, 1986).
The second characteristic can be described as the use of ambiguous words. The effects of ambiguous questions failed to appear in this study. Nevertheless, ambiguous response scales do decrease the reliability of response. Furthermore, other studies showed a decrease in reliability of responses and an increase of response on ambiguous questions (Borgers & Hox, 1999; Leeuw & Otter, 1995). All in all we may conclude that ambiguous words in questions should be avoided.

A univocal result is the negative effect of the number of response options. This effect is discordant with research on the effects of response options on the reliability of responses in questionnaire research with adults (Alwin, 1997; Alwin & Krosnick, 1991; Miller, 1956). The more response options offered with adults, the higher the reliability of responses with and optimum around seven response options. Reasoned from the satisficing theory (Krosnick, 1991) this could mean that the more response options can serve as 'anchor points' for respondents to reminisce from memory and give more options to fit their own answer to the offered response options. The results in this study however, showed the opposite effect of the number of response options in questionnaire research with children. Besides, it confirms results in an earlier study, which research the effects on item non-response (Borgers & Hox, 1999). It seems to be clear that we should avoid too many response options in questionnaire research with children. More response options can place a burden on children because of the cognitive demands. Children have to read all the options and interpret the differences between them. The more options offered the more refined these differences are. Children in these age groups are limited in their language development, which implies limitations in comprehension and verbal memory. It is likely that these children do not recognize these refined differences between the response options and get confused by the unrecognized differences between the options. This can cause a satisficing strategy with these kinds of questions.

Contrary to the results from an earlier study on item non-response (Borgers & Hox, 1999) the position in the questionnaire does have a positive effect on the reliability of responses. The questions in the third part of the questionnaire produce the most reliable responses. However, this position in the questionnaire produces the most item non-response (Borgers & Hox, 1999). Randomizing the position in the questionnaire cannot avoid item non-response in the third part of the questionnaire and cannot increase the reliability of responses in the first part

of the questionnaire but it can randomize item non-response and reliability of responses over questions.

There are some expected question characteristic effects that failed to appear. For example double barreled questions and complexity of the question. Other results that failed to appear seem to have the opposite effect to our hypothesis, like questions with complex constructions. Also the expected interaction effect between child and question characteristics on the reliability of responses failed to appear in this study.

The reason that some of the results failed to appear and others or seem to be the opposite of our expectations can be twofold. The first reason could be that the formulated hypotheses were incorrect. For instance, questions with complex constructions. In terms of the satisficing theory this characteristic can be considered as cognitively demanding. Children in our population still develop language and reading skills. Questions with complex constructions require more reading skills than questions without these constructions. In this line of reasoning questions with complex constructions are difficult to respond and cause less reliable responses. However, the results in this article do not confirm this hypothesis. Possibly, these kinds of questions force children to use an optimizing strategy because they are more difficult than questions without these constructions. Children have to read the questions very well and have to concentrate to understand the meaning of these questions. If they are forced to use an optimizing strategy they will produce more reliable responses.

The second reason could be that some of the results are an artifact of the method or data used in this study. We used secondary analysis on data that are collected for other purposes than our study. For example, the hypothesized interaction effects that failed to appear in this study were found in earlier studies (Borgers, 1997; Borgers, 1998; Borgers & Hox, 1999; Leeuw & Otter, 1995). These studies give indications to support this hypothesis for children. That no interaction effect followed in this study can be caused by the fact that researchers, who developed the used questionnaires, adapt the design of the questions to the cognitive abilities of the researched population. In other words difficult questions are only asked to the oldest children. The data of the different studies are collected from different populations. In our data we found significant correlations between years of education and 18 question characteristics out of the 21. These correlations vary between r_{xy}=-.046 (readability index, Douma) and r_{xy}=.749 (balance of the question). From this we can conclude that the questionnaire designers

have been succesful in the adaptation of the questions to the concerned population. Secondly, this correlation probably causes that the interaction effect failed to appear in this study. In an experimental setting these adaptations can be precluded, because all respondents, with all kind of cognitive abilities will get the same questions. Then we are able to test if the result failed to appear only in this study or we cannot confirm this hypothesis that follows implicitly from the satisficing theory.

Beside the effects that failed to appear or seem to be opposite to our expectations, the results that were found could also be a result of an artifact. Correlated effects between independent variables cannot be prevented completely in secondary analysis. A moderator variable can be part of it. In following research, experiments are necessary to prevent the analysis for these moderator effects and find out if the results of this study can be supported.

References

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales. Which are better? *Sociological Methods & Research, 25*(3), 318-340.

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement. The influence of question and respondent attributes. *Sociological Methods and Research, 20*(1), 139-181.

Andrews, F. M. (1984). Construct validity and Error Components of Survey Measures: a structural Modeling approach. *Public Opinion Quarterly, 48*, 409-442.

Andrews, F. M., & Herzog, A. R. (1986). The quality of survey data as related to age of respondents. *Journal of the American Statistical Association, 81*(394), 403-410.

Benson, J., & Hocevar, D. (1985). The Impact of Item Phrasing on the Validity of Attitude Scales for Elementary School Children. *Journal of Educational Measurement, 22*(3), 231-240.

Bergh, B. v. d. (1995). *Onderzoek 'De leefsituatie van Kinderen op schoolleeftijd'. Beknopte beschrijving van het onderzoek + vragenlijsten + bijlagen* . Brussel: Centrum voor Bevolkings- en Gezinsstudie.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (Eds.). (1991). *Measurement Errors in Surveys*. New York: Wiley.

Borgers, N. (1997). *De invloed van taalvaardigheid op datakwaliteit bij vragenlijstonderzoek onder kinderen [in Dutch] (The influence of language and reading ability on data quality in questionnaire research with children)* (unpublished). Amsterdam: University of Amsterdam, Department of Education (POW).

Borgers, N. (1998). *The influence of child and question characteristics on item non-response and the reliability in self-administered questionnaires: Coding scheme and prelimanary results.* Paper presented at the SMABS Conference, Leuven.

Borgers, N., & Hox, J. J. (1999). *Item non-response in questionnaire research with children.* Paper presented at the International Conference on Survey Non-response, Portland.

Borgers, N., Leeuw, E. d., & Hox, J. J. (1999). Surveying children: Cognitive development and response quality in questionnaire research . In A. Christianson, J. R. Gustafson, A. Klevmarken, B. Rosén, K.-G. Hansson, L. Granquist, & K. K. (Eds.), *Official Statistics in a changing world.* (pp. 133-140). Stockholm: SCB.

Borgers, N., Leeuw, E. d., & Hox, J. J. (2000). Children as respondents in survey research: cognitive development and response quality. *Bulletin de méthodologie sociologique (BMS), 66*, 60-75.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear model: Applications and data analysis methods.* Newbury Park: Sage.

Cannel, Miller, & Oksenberg. (1990). *Research on interviewing techniques. Field experiment in health research 1971-1977.*

Fukkink, R. (1996). 'Peer tutoring' in het leesonderwijs: verslag van het eerste experimentele jaar van Stap Door! *Spiegel, 14*(3), 47-71.

Fukkink, R., & Vaessen. (1996). *Stap Door! Verslag van het experimentele jaar (1995-'96*. Utrecht: Sardes.

Goldstein, H. (1995). *Multilevel statistical models*. (2nd ed.). London: Edward Arnold.

Groves, R. (1989). *Survey error and survey costs.*: Wiley.

Hakstian, A. R., & Whalen, T. E. (1976). A K-sample significance test for independent Alpha coefficients. *Psychometrika, 41*(2), 219-231.

Hattum, M. J. C. v. (1997). *Pesten. Een onderzoek naar beleving, visie en handelen van leraren en leerlingen.* Unpublished Ph.D., University of Amsterdam, Amsterdam.

Hayes, W. L. (1974). *Statistics for social sciences*. (2nd ed.). London: Holt, Rinehart and Winston.

Herzog, A. R., & Rodgers, W. L. (1992). The use of survey methods in research on older Americans. In R. B. Wallace & R. F. Woolson (Eds.), *The Epidemiologic study of the elderly* (pp. 60-89). Oxford: Oxford University Press.

Hox, J. J. (1995). *Applied Multilevel Modeling*. Amsterdam: TT-Publikaties.

Knäuper, B., Belli, R. F., Hill, D. H., & Herzog, A. R. (1997). Question difficulty and respondents' cognitive ability : The effect of data quality. *Journal of Official Statistics: an international review, 13*(2).

Krosnick, J. (1999). *The causes of No-opinion responses to attitude measures in surveys: They are rarely what they appear to be.* Paper presented at the International Conference on Survey Non-response, Portland.

Krosnick, J., & Alwin, D. (1987). An evaluation of cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51(2)*, 201-219.

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

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing Rating Scaling for Effective Measurement in Surveys., *Survey Measurement and Process Quality* : John Wiley & Sons Inc.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: initial Evidence. *New Directions for Evaluation, 70*, 29-43.

Lagerweij, N. W. (1995). *Milieugedrag bij kinderen.* Unpublished Ph.D., University of Amsterdam, Amsterdam.

Leeuw, E. D., & Otter, M. E. (1995). The reliability of Children's Responses to Questionnaire Items; Question Effects in Children's Questionnaire Data. In J. J. Hox, B. F. v. d. Meulen, J. M. A. M. Janssens, J. J. F. t. Laak, & L. W. C. Tavecchio (Eds.), *Advances in Family Research* . Amsterdam: Thesis Publishers.

Lyberg, L., Biemer, P., Collins, M., Leeuw, E. d., Dippo, C., Schwarz, N., & Trewin, D. (Eds.). (1997). *Survey Measurement and Process Quality*. (Vol. 1). New york: John Wiley & Sons, Inc.

Maccoby, E. E., & Maccoby, N. (1954). The interview: A tool of social science. In G. Lindzey (Ed.), *Handbook of social psychology.* (Vol. 1 Theory and method, pp. 449-487). Cambridge: Addison-Wesley.

Marsh, H. W. (1986). Negative item bias in rating scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology, 22*(1), 37-49.

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review, 63*(2), 81-97.

Nunnally, J. C. (1978). *Psychometric Theory*. New-York: McGraw-Hill.

Orwin, R. G. (1994). Evaluation coding decisions. In H. H. Cooper, L.V. (Ed.), *The handbook of research synthesis* . New York: Russel Sage Foundation.

Otter, M., Mellenberg, D., & Glopper, K. d. (1995). The relation between information-processing variables and test-retest stability for questionnaire items. *Journal of Educational Measurement, 32*(2), 199-216.

Otter, M. E. (1993). *Leesvaardigheid, leesonderwijs en buitenschools lezen. Instrumentatie en effecten.* Unpublished Ph.D., University of Amsterdam, Amsterdam.

Peetsma, T. T. D. (1992). *Toekomst als motor? Toekomstperspectieven van lerlingen in het voortgezet onderwijs en hun inzet voor school.* Unpublished Ph.D, University of Amsterdam, Amsterdam.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., & Lewis, T. (1999). A user's guide to MLwiN (Version 2.0.001). Lodon: Institute of Education. university of London.

Rodgers, W. L., Andrews, F. M., & Herzog, A. R. (1989). *Quality of Survey Measures: A Structural Modeling Approach.*

Schoonen, R., Triescscheijn, B., Gelderen, A. v., & Klerk, A. d. (1993). *Evaluatie van de leeskisten Het Plein en De Klas van Meester Ed.* (Vol. Sco).

Schwarz, N., & Hippler, H. J. (1995). Subsequent Questions may influence answers to preceding questions in mail surveys. *Public Opinion Quarterly, 59*, 93-97.

Schwarz, N., Knäuper, B., & Park, D. (1998). *Aging, Cognition, and Context Effects: How Differential Context Effects Invite Misleading Conclusions About Cohort Differences.* Paper presented at the Conference on Methodological Issues in Official Statistics, Stokholm.

Scott, J. (1997). Children as respondents: methods for improving data quality. In L. Lyberg (Ed.), *Survey measurements and process quality* . New York: Wiley.

Stock, W. A. (1994). Systematic coding for research synthesis. In H. H. Cooper, L.V. (Ed.), *The handbook of research synthesis* . New York: Russel Sage Foundation.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers. The application of cognitive processes to survey methodology*. San Fransisco: Josey-Bass.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin, 103*, 299-314.

Appendix 1

Summary of the data sets

Data set	Topic/Contents	Reference
1	Perspectives and effort in school	(Peetsma, 1992)
	4 questionnaires; 18 scales, 142 questions, 606 pupils	
2	Environmental behavior	(Lagerweij, 1995)
	3 questionnaires; 15 scales, 175 different questions, 859 pupils	
3	Reading tutoring project	(Fukkink, 1996)
	2 questionnaires; 4 scales, 66 questions, 800 pupils	(Fukkink & Vaessen, 1996)
4	Perceived competence	(Bergh, 1995)
	3 questionnaires; 8 scales, 48 questions, 758 pupils	
5	Reading pleasure	(Schoonen, Triescscheijn, Gelderen, & Klerk, 1993)
	1 questionnaire; 1 scale, 20 questions, 443 pupils	
6	Bullying	(Hattum, 1997)
	1 questionnaire; 5 scales; 62 questions,	

Appendix 2

Summary of the intercoder reliability for all questionnaires after corrections
Number of items = 513

	Cohen's Kappa	Percentage corresponding codes
Number of words in the introduction text	1.00	
Instruction in the question	1.00	
Ambiguity of the question	.87	
Ambiguity of the response scale	.86	
Double barreled question	.92	
Complex constructions	.95	
Negative formulated question	.91	
Kind of information that is being asked for	1.00	
Complexity of the question	.88	
Reference period	1.00	
Numeric quantity		99%
To personal for the respondent To threatening for the respondent Rather not answer the question Hard to give an honest answer to the question	.97	
Position in the questionnaire	1.00	
Balanced question	1.00	
Number of answer categories	1.00	
Neutral midpoint	1.00	
Labeled response scale	1.00	
Don't know filter	1.00	
Something else category	1.00	

1 Coding scheme: a technical report

Natacha Borgers

University of Amsterdam
Faculty of Social and Behavioral sciences

Introduction

In large-scale surveys children are no longer neglected as respondents. Researchers are more and more convinced that information about perspectives, attitudes, and behavior of children should be collected from children themselves, and not from the responsible parents or other sources of information, as researchers were used to do until recently [Scott, 1997 #270. That is why it becomes more usual to let children participate in large-scale surveys.
However, methodological knowledge on how to survey young children is scarce. Researchers have to rely on ad-hoc knowledge from such diverse fields as child psychiatry and educational testing. Another way is to rely on methodological knowledge on how to survey adults.

According to Krosnicks' satisficing theory (1991) there are three factors, which affect the process of answering questions. The first is the motivation of the respondent to perform the task, the second is the difficulty of the task facing the respondent, and the last is the respondent's cognitive ability to perform the task. This theory tries to explain why some respondents perform the cognitive task of answering questions very well and others do not. This satisficing theory elaborates on the question answering process (Tourangeau & Rasinski, 1988). Four important steps can characterize the question answering process:

1. Understanding and carefully interpreting the question being asked
2. Retrieving the relevant information from memory
3. Integrating this information into a summarized judgement
4. Reporting this judgement by translating it to the offered response scale.

Implicitly this theory involves an interaction between respondent characteristics and question characteristics. In other words, the lower the respondents' cognitive abilities and motivation the more sensitive they will be for difficult questions and the more likely they will use a satisficing strategy to answer the questions. More specific, the less cognitive sophisticated respondents are, the more sensitive they are for difficult or cognitive demanding questions and the less reliable their responses will be. Till now there has been done little research on this interaction effect (with the exception of among others, Knäuper, Belli, Hill, & Herzog, 1997; Schwarz, Knäuper, & Park, 1998)

Resulting from the above-mentioned theory both aspects, respondent and question characteristics, are sources of error in questionnaire research.
Much research has been done on numerous question characteristics and data quality. It is now well known that even slight variations in the way an attitude question is asked can significantly change answers. Altering the order of questions, the order of response alternatives, the words used in the questions stems, the number of response options, and many other aspects of questions can change both the distributions of responses to a question and the reliability of those responses (Krosnick, 1991). Andrews (1984), for example, found that the characteristics of survey design account for a large part of the variation in the estimate of validity, method effects and residual error, and question characteristics are an important part of these characteristics. Nevertheless, in general these studies give fragmentary and often ambiguous results, so it tends not to be conclusive (Rodgers, Andrews, & Herzog, 1989). Besides, the errors produced by question characteristic vary, as a result of the cognitive activity question requires.

Questions are closely linked to the errors interrelated with respondents. The question identifies for the respondent the cognitive tasks to perform, which information to retrieve from memory, and what judgements are sought, how one should translate their own judgement into the given response scales. If questions are difficult and fail to identify one of these steps for the respondent there is a chance that respondents switch to a satisficing strategy.

To research effects of question characteristics on response quality in survey research with children we first had to develop a coding scheme to distinguish different characteristics within questionnaires. The purpose of this report is to give insight in the question characteristics we use in our coding scheme. The coding scheme is based on the steps distinguished in the already mentioned question answering process model (Tourangeau & Rasinski, 1988) and results of empirical research with adults on the effects of question characteristics on response quality. Because question characteristics are also related to the steps in the question answering process we will follow Knäuper et al. (1997) in their classification of question characteristics according these steps. Within these steps different characteristics will be exerted influence on the process.

In this report we will give an overview of the large number of question characteristics and their effect on response quality, without pretending to give a complete overview of researched question characteristics. Molenaar (1991) for example researched a lot of question characteristics using secondary analysis. Only a few of these characteristics will be discussed in this section because many of them have no effect at all on response quality, and have additionally never been used by others. Moreover, for our purpose, developing a coding scheme to code question characteristics within a questionnaire, it is important to get insight in the possibilities to measure these characteristics. Although most of the research concerning question characteristics has been pointed the first and the last step in the question answering process (question wording and response scales) but we will also pay attention to question characteristics concerning the second and the third step of the process.

Comprehension and Interpretation.
Comprehension and interpretation of the question are the central elements in the first step of the question answering process and can be influenced by various language aspects. The harder the question is to comprehend and interpret for the respondent, the faster respondents would be satisfice when they answer such a question. It concerns characteristics like wording aspects, among other thing length of questions, introductory phrase, instruction, ambiguity or vagueness of the question (e.g. cryptic questions, undefined abbreviations and jargon, double-barreled questions). These characteristics will subsequently be discussed.

In readability research there are many well-known readability indicators that can influence the comprehension of texts. Some of these indicators are also used as an indicator that can influence the comprehension and interpretation of questions.
The first characteristic that has been well researched and can be related to difficulty is the length of the question. Two opposite hypothesis concerning question length can be formulated. The first hypothesis assumes that longer questions communicate more information about the nature of the cognitive task that is asked from the respondent. Longer questions can help to explicate the meaning of the question and can serve as a communication resource. By that longer questions will result in better response quality. Groves (1989) suggests that most results confirm the first hypothesis, namely longer questions may improve response quality. Longer questions force respondents to engage in deeper cognitive processing and allow respondents to give their response in their own words. This can reduce effects of social desirability and improve recall. Longer question can serve as 'anchor points' for respondents to reminisce from memory.
The second hypotheses assume that longer questions place a burden on respondents because of the cognitive demands it bring about. For an adequate response it is necessary to keep all of the information in the working memory. The amount of information communicated by means of a long question exceeds the capacity of the respondent to retain it, and confuses the respondent (Groves, 1989). Besides linguistic studies have shown that the difficulty of a sentence is related to the sentence length in words, as well as the educational level of the reader. Lower skilled people have more difficulty in understanding longer sentences than higher skilled people do.

Different empirical studies support the first hypothesis. Andrews' results (1984), for example, show that medium or long questions (16-34, or 25+ words) yielded higher data quality than shorter questions. He combined in his study question length with the length of the introduction text and found besides two positive main effects also a first order interaction. The overall of the result patterns suggests that short introductions followed by short questions are not good neither are long introductions followed by long questions.
Knäuper et al. (1997) also shows that question length is a predictor of the propensity to say "don't know". They found an interaction effect between cognitive ability and question length. The effect, although not statistically significant, shows that lower cognitive ability respondents profit more from longer questions than higher cognitive ability respondents do. However, Rodgers, Andrews and Herzog (1989) show in their study that longer questions and longer introductions to sets of questions are associated with lower validities than questions with as few as 15 words, especially after controlling on other measure characteristics.

The differences between linguistic studies and studies who research question characteristics, maybe related to the differences between questions and sentences: sentences are a part of a whole text while questions stand more or less on their own. May be these contradictory results can be explained by the limit of the working memory. Miller (1956) found by means of experiments using magnitude estimates of judgements of a number of physical stimuli that the optimal number of elements was seven plus or minus two. Although it is not sure if these results can be generalized to survey questions, one can imagine that it could be the case with extremely long and complicated questions, in which more than seven elements are presented. This means that even though questions are long, generally they do not offer more than seven elements at the same moment.
In our coding scheme we both aspects, length in sentences and length in word will be included.

The train of thought that has been posed for question length can also be applied for introductory texts. Different studies have used the introductory text as a question characteristic, but all in a different way. We will next discuss a few of these used definitions. Knäuper et al. (1997) uses the introduction text in terms of 'whether or not the question contains a introductory phrase'. The results of this study show that the presence of an introductory phrase is a predictor for the propensity to say "don't know". They also found that the mean percentage of "don't know" responses is larger among lower cognitive sophisticated respondents compared with higher cognitive sophisticated respondents. They explain this effect from the idea that questions, which are introduced with a phrase, can provide difficulties because they introduce a number of details that are required to integrate with the question. In other words an introductory phrase makes the task more cognitively demanding for respondents. Molenaar (1991) used the introductory part in two ways as a question characteristic. First the length of the introductory text measured in classes of five words. Second whether knowledge that is judged to be relevant for the respondents to answer the question is transmitted in the introductory part of the question or not. This connects with the recommendation of Salant & Dillman (1994) that you should not assume that respondents know enough to answer your question. Likewise too much precision or too specific questions or introductory texts can confuse respondents.
May be question length and introductory text cannot be researched on their own. As pointed out before, Andrews (1984) found in his study that short introductory texts with short questions are not good, neither are long introductions followed by long questions. In our coding scheme the number of words in the introduction text in measured. In that way the presence and the length of the introduction text is included. Second we can also research the interaction effect between question length and introductory text length as Andrews (1984) did.

Besides word- and sentence length a lot of other characteristics can be used as indicators for readability of questions. Molenaar (1991) for example used in his study a Dutch version of the reading ease formula developed by Flesch. Flesch has combined in this formula word- and sentence length as variables in a particular mutual relationship. However, Molenaar (1991) did not find a significant effect of this characteristic on data quality indicators like the mean, the standard deviation and the percentage of non-substantive responses (no opinion, don't know, and no answer). In readability research wordfrequency and wordfamiliarity are frequently used indicators and appear to be powerful predictors for readability. As far as we know

these indicators are not used in survey methodology research. To determine the density of (in)frequent words, readability researchers use wordfrequency frames. This means that one supposes that infrequent words are unknown and frequent words are well known (Staphorsius, 1994). Another formula that combines sentence length as well as wordfrequency is a technical readability index. A second advantage of this index concerns the fact that it is not based on texts as criterion, like for instance the Flesch index. Therefore it seems more applicable for questions (Staphorsius, 1994)
That is why our coding scheme uses the technical readability index as an indicator for the readability of questions. However we also included the Dutch version of the Flesch index following Molenaar (1991) and because of the wide spread experience of this index.

The syntactical structure of a question can also cause difficulties for the respondent, especially for children who are still learning all kind of language rules. Thus the more complex questions are in their syntactical structure the more difficult they will be to read for respondents. Knäuper et al. (1997) included this question characteristic in their research. However, the level of agreement for this characteristic was unsatisfactory low and the chance adjusted Kappa was also very low for this measure. That is why they removed this code from analysis. We decided to include this characteristic in our coding scheme contrary to these results. We tried to give a complete as possible description of this characteristic for coders to overcome this problem.

Another characteristic that can confuse respondents in their understanding of the question is the use of questions that instruct respondents to include or exclude experiences or examples from consideration in their response. Knäuper et al. (1997) suggests that these kinds of questions are difficult to answer. They can mislead or distract respondents from the real question, and is accompanied with requiring a comprehension of all the details pointed out. In this way it places a great cognitive demand on the respondent. They define this characteristic as 'whether or not the question contains one or more instructions'. They found this characteristic to be one of the most important predictors of the propensity to say don't know. Following Knäuper et al. (1997) our coding scheme measures the presence of an instruction to the question.

There is a widespread agreement that negatively formulated questions are difficult to answer for respondents. Because respondents have to give a negative answer to a question when they want to give a positive response, it places a cognitive burden on respondents. Krosnick & Alwin (1987) state that people are cognitive misers, so respondents seek to provide minimally satisfactory answers and avoid additional cognitive effort that would be necessary for the task. Nevertheless, negative questions are still used and recommended, because the use of an equal number of positively and negatively worded questions could reduce the influence of response styles on the responses. Some research has been done on the effect of negative formulated questions on the response of children (Benson & Hocevar, 1985; Marsh, 1986). Although the amount of distortion varies with the age of children it is also substantial with older children. Moreover, it seems that negatively formulated questions measure other constructs than positively formulated question. Therefore, Benson & Hocevar (1985) and Marsh (1986) have recommended avoiding this kind

of formulations. They state that it may be useful to add some negatively formulated questions to the questionnaire only to reduce response bias, but remove them for further analysis. A negatively formulated question can be defined as follows; if a question in combination with its response options does have a possibility of double negatives. In other words, if respondents have to give a negative answer to a question to give a positive response questions will be coded as negative in our coding scheme.

Another question characteristic concerning the step of comprehension and interpretation refer to the clarity and ambiguity of questions. Again this characteristic concerns the wording aspect of the question*. The reader should not infer that ambiguity in words of survey questions is limited to the vague quantifiers that Bradburn and Miles discuss for attitudinal question. Questions that would often be labeled as factual that is, for which there are observable behaviors, are also subject to such problems* (Groves, 1989, p.454).
Different studies have shown that ambiguous questions have some negative effects on response quality. Knäuper et al. (1997) stated that some concepts may be difficult to translate into easy to understand questions. They characterize a question as ambiguous when it includes ambiguous or unfamiliar terms. The characterization of ambiguous questions is more clearly described by De Leeuw & Otter (1995) and Otter, Mellenbergh & de Glopper (1995). They characterize a question as ambiguous when either the question itself or its answer categories contain words, which are not likely to have the same clarity for all the respondents. A question is also characterized as ambiguous when it is subject to different interpretations. Besides vaguely worded questions and cryptic questions can be misunderstood by the respondents and produce useless information and interpreted as ambiguous (Salant & Dillman, 1994).
Knäuper et al. (1997) as well as De Leeuw & Otter (1995) and Otter, Mellenbergh & de Glopper (1995) found results that support the idea that ambiguous questions have negative effects on response quality. Knäuper et al. show that questions containing ambiguous terms are more likely to be answered with "don't know" than questions without ambiguous terms, especially when they are introduced by a phrase. Respondents lower in cognitive ability are more affected by ambiguous terms than respondent higher in cognitive ability. The studies of De Leeuw & Otter (1995) and Otter, Mellenbergh & de Glopper (1995) show the same results, although their research population concerns children. The test-retest correlation is higher for unambiguous questions than for ambiguous question. They have also found an interaction effect between age of the respondent and the clarity of the question. Younger children (9 years old) have far more difficulty with ambiguous questions than older children (14 years old). For our coding scheme we distinguished ambiguous questions from ambiguous response scales.

Salant & Dillman (1994) discuss a number of question characteristics that can be classified under ambiguous questions. They stated that 'double-barreled' questions would produce ambiguous or not mutually exclusive answers. Because of that we distinguished in our coding scheme double-barreled questions from ambiguous questions.

The last characteristic that can affect the comprehension and interpretation of a question concern the kind of information that is being asked. Actually most of the

questions can be divided into the following categories of information adapted from Dillman (1978):

- Attitudes
- Opinions
- Behavior
- Attributes
- Capabilities
- Experiences
- Empathy
- Knowledge

An attitude can be defined as the attitude of individuals compared towards something, while opinions refer to the beliefs about reality. In some cases responses are mixtures of attitudes and opinions. The most questions about behavior are actual questions about opinions of individuals about their own behavior. If you ask respondents according to the capabilities, it refers to what they are able to. Usually this means that respondents are asked to give their opinion about their own capabilities. Attributes are characteristics of the respondent. It concerns characteristics that respondents consider to be as a fact. Experiences refer to what respondents have lived through. This definition about the information that is asked for is exactly the way we used it in our coding scheme.

Retrieval
The second step in the question answering process appeal especially to the memory of the respondent. Just like in the step above there are a number of question characteristics that can influence the cognitive demands this step ask for. All the more the respondents have to search in their memory over a longer period or the search in their memory is complicated, the more respondents would switch over to a satisficing strategy. The question characteristics that involve this step are among other things: complexity of the question, retrospective reports, frequency reports, and appropriateness of time references.

Two different studies have used the complexity of the question as a question characteristic, which can influence response quality (Leeuw & Otter, 1995; Otter et al., 1995). These two studies used both the same descriptions for complexity and can be distinguished from complex syntactical structure of the question. A question was characterized as complex when the required information could not be extracted right away from memory. Otter, Mellenbergh & de Glopper (1995) showed that the consensus of the judges, between the three independent judges, in these studies was good . Complex questions are difficult tasks for respondents because they are expected to compile various information units, combine and then compare them (Leeuw & Otter, 1995). Krosnick & Alwin (1987) have showed that respondents do not put much cognitive effort into the task to respond on complex questions. This means that complex questions will contribute to response distortions, because respondents switch over to a satisficing strategy. Both mentioned studies have demonstrated that complexity affect data quality. The test-retest correlation is lower for complex questions than for simple questions. Furthermore the effect of complexity is more clearly with younger (9 years old) than with older (14 years old) children (Leeuw & Otter, 1995; Otter et al., 1995). In our coding scheme we used the same description for complex questions.

Different studies pointed out the effect of reference period on response quality (Knäuper et al., 1997; Rodgers et al., 1989). They stated that 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. Salant & Dillman (1994) recommend using a cognitive design if you want to ask retrospective question. Ask a number of questions related to the specific event to activate the memory and time period, so respondents can ultimately answer the major question.
The same applies more or less for questions asking for numerical quantity. These kinds of questions ask for information that cannot be retrieved from memory right away. According to Knäuper et al.(1997) questions asking for numerical quantity yield may also require an extensive memory search a counting or estimation procedures. Respondents lower in cognitive ability are significant more strongly affected by these effects than respondents higher in cognitive ability.
In our coding scheme we included both aspects, asking for a numeric quantity and we distinguish questions that ask for defined or undefined reference period.

Judgement

This step in the process is a not a well-researched step, with regard to question characteristics and response quality. The tone and implications of a question can influence judgements. If the question head for one particular opinion, for example, it can influence the response quality. Characteristics that can influence the judgement of the respondent are among other things: suggestive worded questions, questions with a subjective tone, sensitive questions and balance of questions, but also the position in the questionnaire. In general it is assumed to avoid emotional and biased words, otherwise the judgement of the respondent would be influenced. The characteristics that are included in our coding scheme, social sensitivity, balance of the question and position in the questionnaire will be discussed in succession.

Questions differ in the extent to which certain responses may be judged to more social desirable than other responses, and in this way it can affect response quality. Contrary to this train of thought, Andrews (1984) and Rodgers, Andrews & Herzog (1989) did not found evidence that measures judged to be more subject to social desirability effects necessarily any less valid than other measures. Even attempting to counteract such effects does not always increase the validity of responses. Generally, social desirability has been conceptualized as a respondent characteristic instead of a question characteristic. Hence, built on a publication of Bradburn, Sudman, Blair & Stocking (1978), de Leeuw & Hox (Leeuw & Hox, 1987) have formulated five questions which concern the question characteristic, question threat. This scale measures the subjective threat a question brings about. In our coding scheme the coders should give an impression if the question is threatening for children concerning, by answering these five questions.

Whether or not the item is included in a battery with other items along with the length of the battery appear to effect responses. In Andrews' (1984) study it is the third most important predictor for data quality in. The results show that how longer the battery, the lower the data quality. This is in contrast with the results of Rodgers, Andrews & Herzog (1989), they did not found any effects of placing a question in a battery on the quality of responses. The position in battery does not have a detrimental effect on the response quality (Andrews, 1984; Rodgers et al., 1989). The kinds of questionnaires we tend to use there are not really batteries of questions available. That is why we not included this characteristic in our coding scheme. What we did include instead is the position in the questionnaire as a question characteristic.

The last question characteristic that will be discussed concerns the balance of questions. Response bias will appear when questions are weighted in one direction. Such a question seems to pressure the respondent to answer in accordance with the suggested direction. This characteristic plays not only a role in the last step of the process but notably when the respondent is judging the question. As we have seen not only the response categories can be unbalanced but also the question itself, so we decided to discuss this characteristic in this step of the process and not into the last step. Mainly because the question-answering process does not only go from the first through the second and the third step to the fourth step, but respondents can go back to a preceding step in the process. Molenaar (1991) have demonstrated that respondents show a greater preference for the opinion that over represented in the question or the response categories, because of the unbalanced formulation. The same results are found when the object of the question is formulated in more positive terms. The respondent is inclined to choose the answer categories that express positive judgements. However, the counter argument poses that unbalanced questions blatantly show the other implicit side of the question that it is not necessary to balance the question (Schuman & Presser, 1981).

Responses.
A lot of research has been done on question characteristics concerning the last step in the question answering process. This step concern the translation of the judgement into the given response categories. Knäuper et al (1997) stated that providing response 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 responses. Likewise, open-ended questions can be very demanding for respondents, close-ended questions are less demanding for respondents and moreover they are much easier to code and analyze. Close-ended questions with unordered choices are usually more difficult to answer than those with ordered choices. The first kind of questions have to process more information. Even though questions are close-ended, too much precision can makes questions nearly impossible to answer. Broad categories make the respondent's job easier (Salant & Dillman, 1994).
If you offer respondents' response categories there are many ways to do so and many response characteristics can influence the data quality in this step. Characteristics that will subsequently be discussed are among other things, number of response choices, offering a midpoint, labeling scales, offering a no opinion filter, and offering a something else category.

One important decision that must be made when constructing a rating scale is how many scale points should be included. In spite of widespread use of rating scales there is little agreement about the optimal number of response alternatives. Nevertheless, the number of rating scale categories apparently affects response quality. Andrews (1984), for example, found that the number of rating scale categories have the most pronounced effects on data quality. There is increasing evidence, and even consensus, that data quality improves as the number of response categories increases (Alwin & Krosnick, 1991; Andrews, 1984; Krosnick & Fabrigar, 1997; Rodgers et al., 1989). The results of these studies show that from 3-point scales upwards, there is a general monotonic increase in reliability and validity, while residual error tends to decrease. The 2-point scales are a major exception to this pattern, as they have relatively reliable responses. (Alwin, 1997; Alwin & Krosnick, 1991; Krosnick & Fabrigar, 1997; Rodgers et al., 1989). The optimal length of a rating scale seems to be 5 to 7 points, because scales of this length appear to be more reliable and valid than shorter or longer ones (Alwin, 1997; Krosnick & Fabrigar, 1997; Rodgers et al., 1989). Besides, according cognitive theorists there appear to be an upper limit on the number of response alternatives people can handle (Alwin, 1997). Maybe this could be explained by the same argument as we borrow from Miller (1956) regarding question length. The optimal number of elements people can remember with their working memory equals seven plus or minus two. As we mentioned before it is not sure if the conclusions formulated in this study apply to survey questions and in this case the number of response alternatives.
Molenaar (1991) used a different approach in studying the number of response categories. His results showed that the size of the standard deviation decreases as the number of answer categories is increased. The relationship reversed, when the 'normal' standard deviations were calculated. These relationships are explained as follows: the further an answer category is from the middle of the scale, the more likely respondents are to avoid that category. In other words they tend to avoid the more

extreme categories. On the other hand, he also showed that the differences of opinion among respondents are more clearly expressed when more answer categories are provided. For our coding scheme the number of response options are counted.

With respect to midpoints, it is less clear whether researchers should include midpoints. Evidence on validity is mixed and the theory of satisficing suggests that including midpoints may decrease measurement quality. Offering a midpoint may discourage people from taking sides and may encourage them to satisfice, whereas if no midpoint is offered, respondents might optimize (Krosnick & Fabrigar, 1997). Andrews (1984) results show that offering respondents an "easy out" by including an explicit midpoint had only slight effects on data quality. Likewise, Narayan & Krosnick (1996) found evidence that education moderated the effect size of offering a midpoint alternative. The high-education group's average effect size was significant smaller than those for the medium- or low-education group.

Fully labeled scales seem to be more reliable and valid that partially labeled scales. Some results indicate significant differences in reliability in favor of fully labeled response scales, among 7-point scales (Alwin & Krosnick, 1991). Because numeric values can alter the meaning of labels, researchers should probably avoid using them altogether and simply present verbal response options alone.
Andrews (1984) found contrary to this and his own expectation that the results of a multivariate analysis suggest that data quality is below average when all categories are labeled. He stated that the data are not sufficient to clarify this matter and it merits further research.

The results concerning offering non-substantive responses diverge. Some studies show evidence for the hypothesis that offering a “don’t know” option decrease response quality, while others find evidence for the opposite. Different studies show that some real attitudes are missed by the inclusion of no-opinion filters. Furnishing a "don't know" option appears to lower the reliability (Alwin & Krosnick, 1991; Krosnick & Fabrigar, 1997; Rodgers et al., 1989). This seems to be evidence for the recommendation that such filters be omitted when possible. Krosnick & Fabrigar suggest using attitude strength techniques to overcome problem of non-attitudes. The satisficing theory supports this idea, because you give respondents an easy way out when you offer a no opinion filter. Molenaar (1991) found that the percentage of non-substantive responses is lower when the 'no opinion' option is absent. He explains this result by the possibly willingness of the interviewers to accept the no opinion response when that category is provide.
The results above are not consistent with previous research. According to Andrews (1984) the second most important survey characteristic in his study is whether the answer categories include and explicit "don't know" option. The effect is of this design is clear and consistent: inclusion of an explicit "don't know" category was associated with better data- higher validity, lower method effects and lower residual error.
We included a something else category in our coding scheme, although we did not found any study, which used this code. A something else category is a catch all category and a concealed don’t know filter. That is the reason why we included this category in our coding scheme

Coding procedure

On the base of these empirical results, from research with adults as respondents, we developed a computerized code scheme to code all the questions in the questionnaires for children (Borgers, 1997). The items in each questionnaire were coded on 26 characteristics by two independent coders. Some of the characteristics are more or less subjective codes. To assign a code to such characteristics, the coders have to interpret the question and make a judgement call. In our study characteristics like this are:

Ambiguity of the question,
Ambiguity of the response scale,
Complexity of the question,
Kind of information that is being asked for,
A judgement if the question is threatening for children,
Is this question hard to answer honestly for most children,
Is the question for most children too personal,
Do most children not like to answer this question.

In such cases, the individual reliability of the coders tends to be small. By combining the results of multiple coders, the reliability of the composite rating can still be satisfactory (Stock, 1994). In our case, discrepancies in these characteristics were handled in the following way. Questionnaires were coded per study. After the coding, a check was made to correct simple coding errors. Next, for each characteristic the intercoder reliability (Cohen's kappa) was determined. If this was lower than 0.70 (the lower bound, Nunally, 1978) considers acceptable for research purposes, a third coder was assigned to this coding task. The final code assigned is the mean rating of all coders. For the remaining characteristics, such as number of response categories, number of words in the question etc., little inference is needed from the coders to assign a code to such characteristics. In these cases, discrepancies between the coders are mostly caused by coding errors. In our case, such discrepancies were resolved by checking if an error was made and repairing that error. If no obvious errors were found, the discrepancy is assumed to be the result of intercoder differences in interpretation, and the characteristic is treated as one of the subjective characteristics. Appendix 2 contains a list of the question characteristics and the mean intercoder reliability for each characteristic separately (Orwin, 1994) which is a summary for 5 data sets and 348 items (Borgers & Hox, 1999).

Summary

In table I we will show an overview of all mentioned question characteristics in this section. It concern the question characteristics that are researched before of could be interesting characteristics regarding response quality.

Table 1 Summary of the question characteristics.

Comprehension and interpretation of the question	
Question length	-> in words
	-> in sentences
Length of the introductory text (in words)	
Presence of instructions in the question	
Readability	-> comprehensive readability
	-> technical readability
Ambiguity	-> of the question
	-> of the response scale
Double barreled	
Complex constructions	
Negatively formulated question	
Kind of information being asked	
Retrieving relevant information from memory	
Complexity of the question	
Reference period	
Numerical quantity	
Judging the retrieved information	
Subjective question threat (sum of 4 indications)	
Balance of the question	
Position in the questionnaire	
Communicate the final response	
Number of response categories	
Offering midpoints	
Offering Don't know filter	
Offering a something else category	
Scale labels	

References

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales. Which are better? *Sociological Methods & Research, 25*(3), 318-340.

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement. The influence of question and respondent attributes. *Sociological Methods and Research, 20*(1), 139-181.

Andrews, F. M. (1984). Construct validity and Error Components of Survey Measures: a structural Modeling approach. *Public Opinion Quarterly, 48*, 409-442.

Benson, J., & Hocevar, D. (1985). The Impact of Item Phrasing on the Validity of Attitude Scales for Elementary School Children. *Journal of Educational Measurement, 22*(3), 231-240.

Borgers, N. (1997). Codingscheme (Version 1.25). Amsterdam.

Borgers, N., & Hox, J. J. (1999). *Item non-response in questionnaire research with children.* Paper presented at the International Conference on Survey Non-response, Portland.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.

Groves, R. (1989). *Survey error and survey costs.*: Wiley.

Knäuper, B., Belli, R. F., Hill, D. H., & Herzog, A. R. (1997). Question difficulty and respondents' cognitive ability : The effect of data quality. *Journal of Official Statistics: an international review, 13*(2).

Krosnick, J., & Alwin, D. (1987). An evaluation of cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51(2)*, 201-219.

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

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing Rating Scaling for Effective Measurement in Surveys., *Survey Measurement and Process Quality* : John Wiley & Sons Inc.

Leeuw, E. D., & Otter, M. E. (1995). The reliability of Children's Responses to Questionnaire Items; Question Effects in Children's Questionnaire Data. In J. J. Hox, B. F. v. d. Meulen, J. M. A. M. Janssens, J. J. F. t. Laak, & L. W. C. Tavecchio (Eds.), *Advances in Family Research* . Amsterdam: Thesis Publishers.

Leeuw, E. D. d., & Hox, J. J. (1987). Artifacts in mail surveys. The Influence of Dillman's Total Design Method on the quaity of the responses. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric Research* (Vol. II,). London: MacMillan.

Marsh, H. W. (1986). Negative item bias in rating scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology, 22*(1), 37-49.

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review, 63*(2), 81-97.

Molenaar, N. (1991). Nonexperimental Research on Question Wording Effects: A Contribution to Solving the Generalizability Problem., *Measurement Errors in Surveys* : John Wiley & Sons Inc.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, 60*, 58-88.

Orwin, R. G. (1994). Evaluation coding decisions. In H. H. Cooper, L.V. (Ed.), *The handbook of research synthesis* . New York: Russel Sage Foundation.

Otter, M., Mellenberg, D., & Glopper, K. d. (1995). The relation between information-processing variables and test-retest stability for questionnaire items. *Journal of Educational Measurement, 32*(2), 199-216.

Rodgers, W. L., Andrews, F. M., & Herzog, A. R. (1989). *Quality of Survey Measures: A Structural Modeling Approach.*

Salant, P., & Dillman, D. (1994). *How to conduct your own survey.* (1st ed.). New York: Wiley.

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

Schwarz, N., Knäuper, B., & Park, D. (1998). *Aging, Cognition, and Context Effects: How Differential Context Effects Invite Misleading Conclusions About Cohort Differences.* Paper presented at the Conference on Methodological Issues in Official Statistics, Stokholm.

Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument.* Unpublished Ph.D. Dissertation, Universiteit Twente.

Stock, W. A. (1994). Systematic coding for research synthesis. In H. H. Cooper, L.V. (Ed.), *The handbook of research synthesis* . New York: Russel Sage Foundation.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin, 103*, 299-314.

Appendix 1 Coding scheme question characteristics

Question characteristics	codes	Description
Introduction text length (Andrews, 1984; Knäuper, Belli & Hill, 1997; Rodgers, Andrews & Herzog, 1989; Salant & Dillman, 1994; Molenaar, 1991)	Number of words in the introduction text	What is the number of words in the introductory text? The introduction for a whole scale or a number of questions is also counted. A number of questions get the same code.
Question length (Groves, 1989; Miller, 1956; Andrews, 1984; Knäuper, Belli & Hill, 1997; Rodgers, Andrews & Herzog, 1989)	Number of words in the question	What is the number of words in the question? Only the stem of the question counts and not the response scale.
Number of sentences	Number of sentences in the question	What is the number of sentences in the question? Only the stem of the question counts and not the response scale.
Presence of instructions in the question (Knäuper, Belli & Hill, 1997)	no = 0 yes = 1	Is there an instruction in the question?
Comprehensive readability (Staphorsius, 1994; Molenaar, 1991)	1-100	* Douma: 206.835 - 0.770 x LGHW - 0.930 x GZW LGHW = number of syllables per hundred words. GZW = the mean sentence length in words. Only the stem of the question counts and not the response
Technical readability (Staphorsius, 1997)	1-100	* CILT: 114.49 + 0.28 x WFREQ - 12.33 x GWL WFREQ = percentage frequent words GWL = mean word length Only the stem of the question counts and not the response
Ambiguity of the question (Otter, 1992; Knäuper et al., 1997; De Leeuw & Otter, 1995; Otter, Mellenbergh & De Glopper, 1995)	not ambiguous = 0 ambiguous = 1	A question should be coded as ambiguous if the stem of the question can be interpreted in different ways. In that case the meaning of the question is not the same for all respondents. Vague quantifiers as 'sometimes' or 'often' but also sentences like 'doing thins at home very often' should be coded as ambiguous.
Ambiguity of the response scale (Otter, 1992; Knäuper et al., 1997; De Leeuw & Otter, 1995; Otter, Mellenbergh & De Glopper, 1995)	not ambiguous = 0 ambiguous = 1	A question should be coded as ambiguous if the response categories of the question can be interpreted in different ways. In that case the meaning of the question is not the same for all respondents. Example: vague quantifiers as 'sometimes' or 'often'
Double barreled question	not double barreled = 0 double barreled = 1	If the question include more than one question, it should be coded as double-barreled.
Complex construction of the question	Simple = 0 Complex = 1	A question should be coded as complex if the question includes complex constructions. Examples are constructions with semicolons, colons, brackets, parenthesis, or subordinate clauses.

Question characteristics	codes	Description
Negatively formulated question (Krosnick & Alwin, 1987; Benson & Hocevar, 1985; Marsh, 1986)	Positive formulated =0 negative formulated = 1	If there is an option of double negative when answering the question it should be coded as negative. Example: I do not like to swim? Yes/No. If I like to swim I should answer no, and that is a double negative option, no I don't.
Kind of information that is being asked for (Dillman, 1978; Hox, 19..)	1 = attitudes 2 = opinions 3 = behavior 4 = attributes 5 = capacities 6 = experiences 7 = knowledge 8 = empathy	What kind of information is being asked for in the question? 1: the attitude of respondents toward something; 2: what respondents think about reality; 3: what the respondent thinks about his own behavior; 4: personal characteristics of the respondent; 5: shown what people are capable of ; 6: what respondents experienced; 7: what the respondent know about certain things; 8: empathy
Complexity of the question (Otter, 1992; De Leeuw & Otter, 1995; Otter, Mellenbergh & De Glopper, 1995)	Simple = 0 Complex = 1	A question should be coded complex if the information that is being asked for in the question can not directly be retrieved from memory. In complex questions the expectation of the researcher is that the respondent all retrieve kinds of information units from memory, combine these information units and compare them. Examples: how often do you go to the library?
Reference period (Knäuper et al, 1997; Rodgers, Andrews & Herzog, 1989)	Not applicable. = 0 Undefined time = 1 Defined time = 2	Does the question ask for an undefined or defined reference period?
Numerical quantity	No = 0 Yes = 1	Does the question ask for a numeric quantity?
Too personal question (Hox & De Leeuw, 198?; Andrews, 1984; Rodgers, Andrews & Herzog, 1989)	Not = 0 Somewhat = 1 Yes = 2	Is this question too personal for most of children?
Too threatening question (Hox & De Leeuw, 198?; Andrews, 1984; Rodgers, Andrews & Herzog, 1989)	Not = 0 Somewhat = 1 Yes = 2	Is this question too threatening for most of children?
Rather not answering this question (Hox & De Leeuw, 198?; Andrews, 1984; Rodgers, Andrews & Herzog, 1989)	Not = 0 Somewhat = 1 Yes = 2	Do most children rather not answer this question?
Hard to give an honest answer (Hox & De Leeuw, 198?; Andrews, 1984; Rodgers, Andrews & Herzog, 1989)	Not = 0 Somewhat = 1 Yes = 2	Do most children find it hard to give an honest answer to this question?
Balance of the question (Molenaar, 1991; Schumann & Presser, 1981)	Balanced = 0 Unbalanced = 1	A question should be coded as balanced if an equal number of positive response options as negative options is offered. In other words a symmetric response scale.

Question characteristics	codes	Description
Position in the questionnaire	1st part = 1, 2nd part = 2 or 3rd part = 3	Classify the questionnaire in three parts with each the same number of questions. The codes of the questions are analogue with the part of the questionnaire. If the questionnaire does not have the same number of questions for every part, the questions that are to much should be put in the second part (at most this could be two questions)
Number of response categories (Alwin, 1984; Andrews, 1984; Rodgers, Andrews & Herzog, 1989; Krosnick & Fabrigar, 1997, Alwin, 1997; Alwin & Krosnick, 1991; Molenaar, 1991; Miller, 1956)	Number of response categories	What is the number of response categories?
Offering midpoints (Krosnick & Fabrigar, 1997; Andrews, 1984, Narayan & Krosnick, 1996)	No = 0 Yes = 1 Not applicable = 9	Does the response scale include a neutral midpoint?
Scale labels (Alwin & Krosnick, 1991; Andrews, 1984)	not = 0 partly = 1 completely = 2	Are the response categories completely, partly or not labeled?
Offering don't know filter (Alwin & Krosnick, 1991; Rodgers, Andrews & Herzog, 1989; Molenaar, 1991; Krosnick & Fabrigar, 1997)	No = 0 Yes = 1	Is an explicit Don't know filter offered in the question?
Offering a something else category (Alwin & Krosnick, 1991; Rodgers, Andrews & Herzog, 1989; Molenaar, 1991; Krosnick & Fabrigar, 1997)	No = 0 Yes = 1	Is a something else category offered in the response scale?

Appendix 2

Summary Intercoder reliability of all questionnaires after corrections
Number of items = 348

	Cohen's Kappa	Percentage corresponding codes
Number of words in the introduction text	1.00	
Instruction in the question	1.00	
Ambiguity of the question	.86	
Ambiguity of the response scale	.90	
Double barreled question	.91	
Complex constructions	.95	
Negative formulated question	.91	
Kind of information that is being asked for	1.00	
Complexity of the question	.87	
Reference period	1.00	
Numeric quantity		99%
To personal for the respondent		91%
To threatening for the respondent		97%
Rather not answer the question	.61	
Hard to give an honest answer to the question		85%
Position in the questionnaire	1.00	
Balanced question	1.00	
Number of answer categories	1.00	
Neutral midpoint	1.00	
Labeled response scale	1.00	
Don't know filter	1.00	
Something else category	1.00	