Quality & Quantity **38:** 17–33, 2004.

Response Effects in Surveys on Children and Adolescents: The Effect of Number of Response Options, Negative Wording, and Neutral Mid-Point

NATACHA BORGERS* and JOOP HOX
Utrecht University

DIRK SIKKEL
Catholic University Brabant

Abstract. Social researchers increasingly survey children and young adolescents. They are convinced that information about perspectives, attitudes, and behaviors of children should be collected from the children themselves. Methodological expertise on surveying children is still scarce, and researchers rely on ad-hoc knowledge from fields such as child psychiatry and educational testing, or on methodological knowledge on surveying adults. Regarding adults, empirical evidence shows that respondent characteristics (cognitive abilities) as well as question characteristics (question difficulty) affect response quality.

This study reports on a methodological survey experiment on the effect of negatively formulated questions, the number of response options and offering a neutral midpoint as response option question characteristics on the reliability of the responses, using children and young adolescents as respondents.

The study shows no effects of negatively formulated questions on the reliability measures, although children respond consistently differently on negatively formulated questions than on positively formulated questions. Taking all results on the effects of number of response options and offering a neutral midpoints on the different reliability measures into consideration; it would appear that offering about four response options is optimal with children as respondents.

Key words: question characteristics, stability over time, internal consistency, response quality, reliability

1. Introduction

Surveying the general population involves a set of complex activities, although procedures to increase general response quality are well documented (Biemer et al., 1991; Groves, 1989; Lyberg et al., 1997). There is an increasing body of empirical evidence that both respondent and question characteristics affect the reliability of responses in surveys (Alwin & Krosnick, 1991; Krosnick, 1991; Krosnick & Alwin, 1987; Krosnick & Fabrigar, 1997; Narayan & Krosnick, 1996; Schwarz &

* Author for correspondence: University of Utrecht, Methodology & Statistics Department, P.O. Box 80.140, NL-3580 TC Utrecht, the Netherlands. E-mail: n.borgers@fss.uu.nl

Hippler, 1995; Schwarz & Knäuper, 1999; Schwarz et al., 1998; Schwarz et al., 1998). These studies show that cognitive abilities (often indicated by respondent education) and variations in question wording may affect responses. Although social researchers increasingly survey children and young adolescents, surveying such special populations is still at the frontier of survey methodology (Scott, 1997). Much methodological advice on how to best survey children is derived from methodological studies on adults and theories based on adults as respondents.

This study reports on a methodological survey experiment on the effect of several question characteristics on the reliability of the responses, using children and young adolescents as respondents. The theoretical background of our study is Krosnick's (1991) satisficing theory, which explains why the reliability of responses differs between respondents, and why it can be affected by question wording. The satisficing theory elaborates a standard question answering process-model developed by Tourangeau (1988; cf. Cannel et al., 1990; Krosnick et al., 1996; Schwarz & Knäuper, 1999; Schwarz et al., 1998; Sudman et al., 1996). In this standard question-answering model a sequence of four steps characterizes 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.

Krosnick's satisficing theory identifies two processes that explain differences in reliability of responses, namely optimizing and satisficing. Optimizing means that the respondent goes through all four cognitive steps needed to answer a survey question. In contrast to optimizing, satisficing means that a respondent gives more or less superficial responses that appear reasonable or acceptable, without going through all the steps involved in the question-answering process. Satisficing is related to three dimensions of the question-answering process: the motivation of the respondent, the difficulties of the task, and the cognitive abilities of the respondent. Low motivation, difficult questions, and low cognitive abilities may lead respondents to provide a satisfactory response instead of an optimal one. Using a satisficing strategy generally produces 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 the respondents are, the more sensitive they are to difficult or cognitively demanding questions, and the less reliable their responses tend to be.

Applying the satisficing theory to special populations, such as elderly or children, is of special interest, because both growing up and aging involve changes in cognitive functioning, and cognitive ability is a central respondent characteristic that affects the reliability of responses. Recently, several studies have shown that reduction in cognitive functioning due to the aging process is associated with a decline in the reliability of survey responses (Alwin & Krosnick, 1991; Knäuper et al., 1997; Krosnick, 1991; Schwarz et al., 1998). Like aging, growing up in-

volves changes in cognitive functioning. In children and young adolescents, both cognitive ability and communicative and social skills are still developing. As a consequence, cognitive ability and social skills vary considerably across children. These differences can lead to the use of different strategies in answering questions and therefore to differences in the reliability of responses obtained in surveys of children and young adolescents.

Piaget's (1929) theory of cognitive development provides a useful instrument to distinguish successive stages in children's cognitive development. According to Piaget, children's intellectual development evolves in a fixed sequence of stages. Piaget's theory has been under criticism from different points of view, but this critique focuses mostly on the timing of the successive stages; the transitions from one stage to another are not as clear as assumed. For our purpose, the directions of cognitive development are more important than the actual stages. Piaget's cognitive developmental theory enables us to combine developmental abilities and the cognitive demands of survey research. In combination with the question-answering model and the satisficing theory, it explains why younger children have more difficulties with cognitive demanding survey questions than older children.

Piaget's developmental theory distinguishes five stages (Flavell, 1985). The first two stages involve developments in early infancy, where verbal surveys are out of the question. In the third stage (*intuitive thought*), children aged from 4 until 7 are developing the basic skills necessary for successful verbal exchange. The age-group as a whole is still limited in their language development, which implies limitations in comprehension and in verbal memory. In the fourth stage (*concrete operations*), children aged 7 until 11 develop language and reading skills. Children of this age start to understand the concept of different points of views (such as self vs. others), they begin to learn classification and temporal relations, but they still have problems with logic forms, for instance negations. In addition, they tend to be very literal in the interpretation of words. In the fifth stage (*formal thought*), with children aged 11 until 15, cognitive functioning (e.g., formal thinking, negations, logic) is well developed. However, children in this age group are still very context sensitive and they may have their own norms. Children or rather adolescents aged 16 and up are treated as adults in survey research (Borgers et al., 1999), there is no reason to treat them as a special group.

In general, Piaget (1955) stated that language and other cognitive development have comparable developmental and transitional periods. Language and cognitive development are both involved in perception, storage and retrieval of information (Holaday & Turner-Henson, 1989). All characteristics of the three stages, *intuitive thought*, *concrete operations*, and *formal thought* appear to be important for the question answering process and may therefore affect this process.

The small number of studies that researched the effects of cognitive abilities of children in survey research support the hypothesis that growing up is related to an increase of the reliability of responses (e.g., Amato & Ochiltree, 1987; Borgers, 1997, 1998; Borgers & Hox, 1999, 2000; Borgers et al., 1999; De Leeuw & Otter,

1995; Otter et al., 1995; Otter, 1993; Vaillancourt, 1973). In general, even slight errors (such as negatively formulated questions) in the questionnaire are more difficult to compensate for children and have larger effects on the responses (Borgers & Hox, 2000; Marsh, 1986).

From Piaget's theory of cognitive development, one can derive directly that negatively formulated questions should pose serious problems in survey research with children. Especially the younger children (11 and younger) have not yet developed the formal thinking that is necessary to understand logical negations. In an earlier study, Borgers and Hox (2000) confirmed the expected negative effect of negatively formulated questions on response reliability. However, they did not find the expected interaction effect between age of the children and negatively formulated questions. A weak point in their study is that it is based on secondary analysis. The data were not collected for this research question, and researchers probably adapt the text of the questions to the cognitive abilities of the researched population.

In addition to the problem of negations, limitations of comprehension and verbal memory are expected to be one of the most important causes of children's difficulties in adequately responding to survey questions. Verbal comprehension and verbal memory are very important in the first two steps of the question-answering process: understanding the question and retrieval of relevant information from memory. For instance, longer questions can help respondents by providing memory cues or as 'anchor points' and by acting as a form of aided recall (Holaday & Turner-Henson, 1989). However, research on the effect of long questions indicates that response reliability in general declines as the length of the question increases (Borgers & Hox, 2000; Holaday & Turner-Henson, 1989; Knäuper et al., 1997), which can be explained by the increased demand that remembering longer questions makes on verbal memory. This leads to the general advice that survey questions should be short and clearly formulated (Borgers et al., 2000). The same contradiction might apply for the number of response options offered. In research with adults there is increasing evidence, and even consensus, that data quality improves as the number of response categories increases (Andrews, 1984; Krosnick & Fabrigar, 1997; Rodgers et al., 1989; Rodgers et al., 1988). However, the more options offered, the more burden is placed on verbal memory. In secondary research with children, the results indicate that offering more response options decreases the reliability of responses of children (Borgers & Hox, 2000).

Offering a neutral midpoint in the response scale can serve as an anchor point, but it seems more often found that offering a neutral midpoints tempts respondents to choose this category (Ayidiya & McClendon, 1990; Narayan & Krosnick, 1996; Raaijmakers et al., 2000). This neutral mid-point is apparently often used in the sense of undecided (Raaijmakers et al., 2000). Given the limitations of children's cognitive ability and communicative and social skills, we expect that they are sensitive to the temptation to satisfice by choosing a neutral mid-point when this is not the optimal answer.

An experimental design is needed to systematically research the effects of these question characteristics in survey research with children. In this study, we describe such an experiment with children aged between 8 and 16 years. The research questions can be summarized as follows: (1) What are the effects of negatively formulated questions, the number of response options and offering a neutral midpoint as response option on the reliability of children's responses, after controlling for age? (2) Is there evidence for an interaction between age and the effect of the question characteristics manipulated in the experiment?

2. Method

2.1. DATA COLLECTION AND PARTICIPANTS

The data have been collected by CentERdata, a University-based research center, which operates a telepanel. The telepanel consists of a random sample of approximately 2000 households who have an Internet connection at home. The panel members receive questionnaires via Internet on a weekly basis. New and replacement households are selected by telephone interviews (CATI). The sample for these interviews is based on random selection of telephone number of Dutch private households (Felix & Sikkel, 1999). Panel members who do not have a computer or an Internet connection borrow a settop box from CentERdata, which enables them to use e-mail and the Internet and to fill out electronic questionnaires.

For this experiment, children in the telepanel households between 8 and 16 years old were asked to answer the questionnaire. The questionnaire was administered twice (repeated measures). The first administration took place between the end of June and the end of September 2001. In the first administration 222 children participated, of whom 117 are boys and 105 are girls. In the second administration 91 children responded. The period between the first and second measurement varies between a minimum of 3 weeks and a maximum of 8 weeks.

2.2. DESIGN

A six by two factorial design was used. Six different number of response options by negatively and positively formulated items were used. Besides the question characteristic offering a midpoint was included. This was not an extra factor in the design because all odd number of response options include a midpoint while it is impossible to include a midpoint when offering an even number of response options. Table I shows in summary the design of the experiment and the construction of the 12 different versions of the questionnaires.

Two instruments were used for this study. The first instrument is the Dutch ten-item version of the Rosenberg Self-Esteem scale (Rosenberg, 1965, 1979), translated by Linden et al. (1983). For use with children the items were reformulated into shorter and more concrete items. One item was judged too complicated

Table I. Design experiment and construction of the different versions of the questionnaires

Version	1	2	3	4	5	6	7	8	9	10	11	12
Number of response options	2	2	3	3	4	4	5	5	6	6	7	7
Positively vs. negatively formulated items	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Offering a Midpoint	No midpoint	No midpoint	Midpoint	Midpoint	No midpoint	No midpoint	Midpoint	Midpoint	No midpoint	No midpoint	Midpoint	Midpoint

for the youngest children and is left out of the questionnaire. The original instrument offered four response options (strongly agree; agree; disagree; strongly disagree). The second instrument is derived from Bradburns well-being scale (Bradburn, 1969; Bradburn & Caplovitz, 1965) and translated into Dutch by Hox (1986). Bradburn's well-being scale originally contains 18 items, nine items ask for positive affect and nine for negative affect. Because we need to include positively versus negatively formulated items in the experiment, we just used the ten items that differ concerning the content of the items (the other Bradburn items are negatively formulated questions of available items). The Dutch version originally offered five response options (never; rarely; sometimes; often; very often). Again the items were reformulated into items that were more concrete and clear for children. The number of response options was adapted to the number of response options that the design allowed, and always included as many positive as negative options. The question order (in blocks of all questions in a scale) was randomly assigned to each of the children.

2.3. DEPENDENT VARIABLES AND ANALYSIS

Two measures of reliability are used as dependent variables (Borgers, 1997; Borgers & Hox, 2000; De Leeuw & Otter, 1995). The first is the item-rest correlation, a measure for internal consistency that is defined at the item level. The item-rest correlation is calculated on the responses of the first measurement occasion only. Item-rest correlations cannot be calculated for individual children, they can only be calculated on groups of children. Therefore, different groups of children are defined, and for each group of children within a questionnaire version, the item-rest correlations are computed for every item in the scale. Since correlations do not have a normal distribution and have a limited range ($-1 \leq r_{ir} \leq 1$) a normalizing transformation, the Fisher-Z transformation (Hayes, 1974)[1] is applied to the item-rest correlations.

The second reliability indicator is defined per child per question, being the absolute difference between the first and second measurement, which is a measure of consistency over time. The distribution of this absolute difference between the first and second measurement is a strongly skewed distribution, which means that most children are very consistent in their responses over time. In addition, the number of response options differs between the different versions, varying between two and seven response options. This means that the absolute differences between the first and second measure are not wholly comparable between the different versions. For this reason we use in addition to the absolute difference between both measures also the relative difference between the first and second measure, which is the absolute difference divided by the number of response options, which results in a value between 0 and 1 for the relative difference.

The analysis model used for all three dependent variables is a multilevel regression model. For an introduction to multilevel modeling, see Bryk and Raudenbush

(1992) and Hox (2002). The data in this study are a hierarchically structured with two levels: items that are nested within children. For the different types of dependent variables, different types of multilevel analysis have been performed.

First, a random-effects meta-analysis has been used for the transformed Fisher's Z as dependent variable. The use of a random-effect meta-analysis has the advantage that the coefficients are analyzed as outcomes, using the information available on their standard error. Bryk and Raudenbush (1992) and Raudenbush (1994) describe a two-level model for meta-analysis, which was used by De Leeuw (1995) and Otter (1995) and by Borgers and Hox (2000).

Secondly, a multilevel logistic regression model was used for the relative difference between the first and second measure. Because this variable is defined as a value between 0–1, it should treated as a proportion. Such data violate several assumptions of the normal regression method, and the appropriate analysis model is logistic regression. The multilevel logistic model is described in Goldstein (1995). The disadvantage of using logistic regression on the relative difference is the shrinkage of the differences between both measures.

Therefore we did a third type of multilevel analysis on the absolute difference between the first and second measure. Since we have to deal with a non-normal dependent variable, and a relatively small highest-level sample size, a nonparametric bootstrap method is used to estimate the parameters and standard errors in the final model. In the nonparametric bootstrap the multilevel regression estimate is carried out once on the total data set. The regression coefficient from this estimation is used to produce predicted values, and the residuals from this analysis are resampled in each bootstrap iteration (Hox, 2002).

The analysis strategy is the same for all three dependent variables. Three models are analyzed. The first model is the null-model. This model decomposes the variance in the dependent variables into the different levels. Secondly, we include the explanatory variables measured at the child level (age, gender, number of weeks between the first and second administration). In the final model the question characteristics (negatively formulated questions, offering a neutral midpoint and the number of response options) are included. We do not assume that the number of response options has a linear effect on the dependent variable. Including all possible dummy variables for the number of response options in the analysis together with the question characteristic 'offering a neutral midpoint' causes a collinearity problem. For that reason the number of response options is included in the analysis as the number of response options, centered around the overall (theoretical) mean, the squared values and the cubed values of the centered number of responses.

The last analysis step is to model these regression coefficients as varying across groups of children. This models different effects of the question characteristics across children, with varying slopes for the question characteristic effects indicating possible interaction effects for these variables with some child characteristics.

Table II. The effects of child and question characteristics on the Fisher-Z transformed item-rest correlations, based on the first measurement ($N = 222$)*

	Null-model	Final model	
Fixed part			
Intercept	**0.329 (0.033)**	**3.579 (0.775)**	**2.968 (0.596)**
Scale characteristics			
Rosenberg vs. Bradburn scale		**−0.141 (0.051)**	**−0.135 (0.034)**
Mean age per version		**−0.163 (0.052)**	**−0.108 (0.043)**
Question characteristics			
Negatively formulated			−0.035 (0.035)
Number response options			**0.054 (0.032)**[2]
Response options (squared)			**−0.019 (0.007)**
Response options (cubed)			−0.000 (0.007)
Neutral midpoint			−0.030 (0.042)
Random part			
Scale level	0.019 (0.008)	0.009 (0.005)	0.000 (0.000)
Deviance	77.56	63.75	43.39
Difference			
$\Delta\chi^2$; Δdf		$\chi^2 = 13.81$; $df = 2$	$\chi^2 = 20.36$; $df = 5$

* The significant ($\alpha = 0.05$) estimates are printed bold.

3. Results

3.1. INTERNAL CONSISTENCY OF THE SCALE

Table II presents the results of the random-effects meta-analysis for the Fisher-Z transformed item-rest correlations. The three columns correspond to the three models. In each column the estimates are given with their standard error within brackets. The lowest level variance (item level) is not included in the table, because in a meta-analysis model the lowest level variance is known and constrained to be equal to the lowest level sampling variance.

The results show negative and significant effects of the characteristics measured at the child level. The negative coefficient for the scale dummy should be interpreted as an instrument effect; the Bradburn well-being scale produces on average significantly lower Fisher-Z transformed item-rest correlations as the Rosenberg Self esteem scale, despite the fact that the Bradburn scale has nine items and the Rosenberg scale has ten.

The effect of the mean age per version should not be interpreted substantively in this analysis. The mean age of the children per version is included in the model as control variable, because despite using random assignment it turned out that the different ages were not completely equally divided over the different versions, older children were more often part of the more difficult conditions.

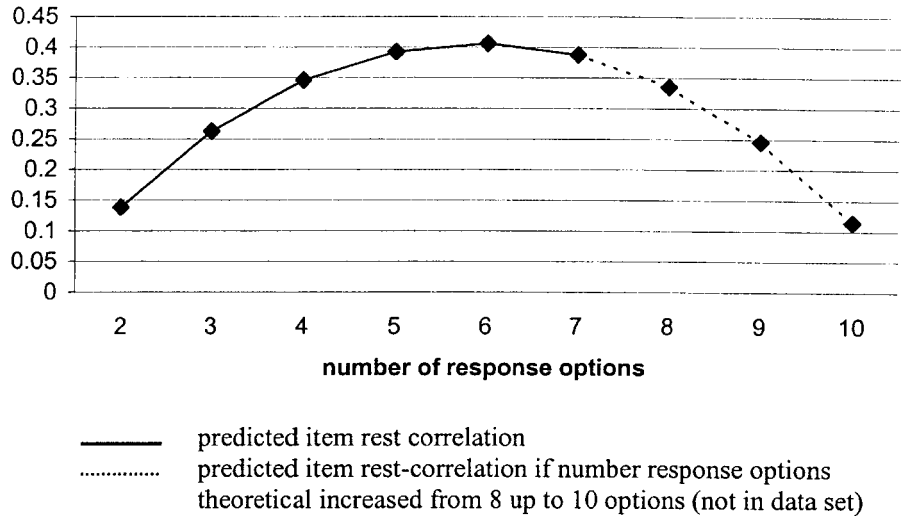

Figure 1. Predicted item rest-correlation for different numbers of response options.

The only significant effect of the question characteristics is the number of response options offered: a positive effect of the number of response options and negative effect of the squared number of response options. Offering negatively formulated questions or a neutral midpoint as a response option, does not show a significant negative effect on the item rest-correlation. However, in general the more elaborate model provides a significantly better fit to the data than the first two models.

To visualize the effect of the number of response options on the items rest-correlation Figure 1 is presented. The figure shows the effect of number of response options varying between two and ten options. The predicted item rest-correlation, with all other explanatory variables set equal to their overall mean, increases from 0.14 to 0.40 when the number of response options goes from two to six. In our experiment, the maximum number of response options was seven. In other words, there is no empirical evidence for the effect of offering eight response options and more. We included these number of response options in the figure to show the (theoretical) predicted decrease after offering six response options, and to show that six response options seems to be the optimum and turning point concerning the effect on item rest-correlation.

3.2. CONSISTENCY OF RESPONSES OVER TIME

Table III shows the results of the multilevel logistic regression analysis on the effects of child and question characteristics on the relative stability of responses over time. The three columns correspond to the three estimated models.

Table III shows only two significant effects on the relative difference between the first and the second measure, an effect of gender and an effect of offering a

Table III. Results of multilevel logistic regression analyses on the effects of child and question characteristics on the relative differences between the first and second measure divided by the range of response options offered ($N = 91$)*

	Null-model parameter (s.e.)	Final model (s.e.)	
Fixed part			
Intercept	**1.59 (0.06)**	**−1.53 (0.34)**	**−1.744 (0.030)**
Child characteristics			
Age		−0.03 (0.03)	−0.03 (0.02)
Gender		0.14 (0.11)	**0.21 (0.11)**
Weeks		0.02 (0.03)	0.01 (0.03)
Question characteristics			
Negatively formulated			0.12 (0.11)
Number response options			−0.02 (0.09)
Response options (squared)			−0.02 (0.03)
Response options (cubed)			−0.01 (0.02)
Neutral midpoint			**0.224 (0.121)**
Random part			
Child level	0.161 (0.040)	0.152 (0.039)	0.124 (0.034)
Question level	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)

* The significant ($\alpha = 0.05$) estimates are printed bold.

neutral midpoint. Girls produce a larger relative difference between the first and the second measure and offering a midpoint in the response scale also produces a bigger difference. Consistent with the results in Table II, negatively formulated questions do not affect the stability over time. Increasing the number of response options, however, also does not show an effect on the relative difference between the first and second measure. This result is likely the effect of the small size of the relative instability over time, which almost disappears after the logistic transformation. For that reason we also performed analyses on the absolute difference, which include bootstrapped estimates. Because of the small sample size and non-normality of the dependent variable, we have more confidence in the bootstrapped results than the asymptotic estimates. The results of these analyses are shown in Table IV.

Table IV shows that most of the variance in the absolute difference between the first and second measure is at the question level. Besides, the variance at the question level is not at all explained by the explanatory variables. 45% percent of the variance at child level is explained by our explanatory variables, just 8.3% of the total variance is explained by our final model.

The asymptotic results show a significant result of age and gender, partly consistent with the results in Table III, and a significant effect of the number of

Table IV. Results of the multilevel regression analysis of the effect of child and question characteristics on the absolute difference between both measures

	Null-model parameter (s.e.)		Final model	Bootstrap estimates	95%-confidence interval
Fixed part					
Intercept	**0.621 (0.041)**	**0.85 (0.26)**	**0.51 (0.18)**	**0.905**	**0.468/1.339**
Child characteristics					
Age		−0.03 (0.02)	**−0.02 (0.01)**	−0.017	−0.043/0.012
Gender		0.01 (0.08)	**0.12 (0.06)**	−0.079	−0.206/0.060
Weeks		0.03 (0.02)	0.00 (0.02)	−0.011	−0.041/0.021
Question characteristics					
Negatively formulated			0.08 (0.06)	0.056	−0.069/0.186
Number response options			**0.17 (0.05)**	**0.192**	**0.075/0.308**
Response options (squared)			−0.00 (0.01)	−0.003	−0.039/0.033
Response options (cubed)			−0.00 (0.01)	−0.006	−0.026/0.015
Neutral midpoint			**0.11 (0.06)**	0.091	−0.070/0.255
Random part					
Child level	**0.133 (0.022)**	**0.105 (0.021)**	**0.026 (0.010)**	**0.060**	**0.029/0.089**
Question level	**0.720 (0.025)**	**0.720 (0.025)**	**0.720 (0.025)**	**0.722**	**0.671/0.770**

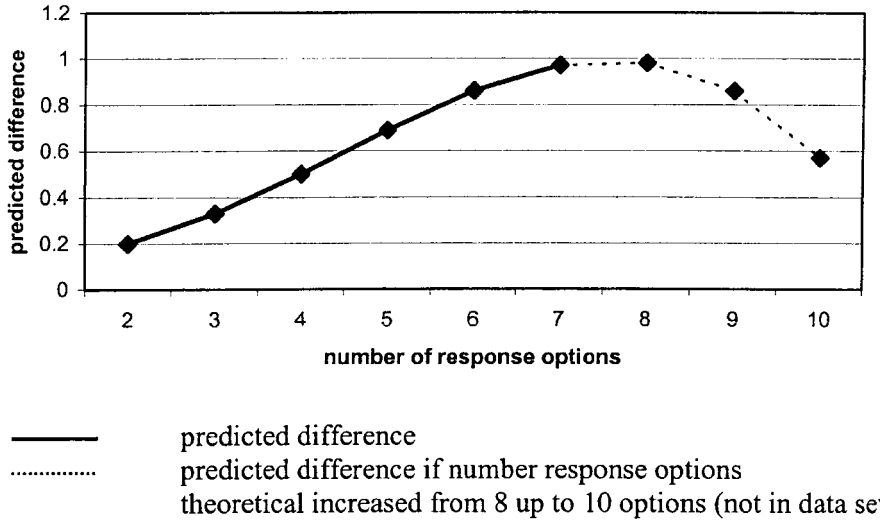

Figure 2. Predicted absolute difference between both measures for the different number of response options.

response options and offering a neutral midpoint. Conversely, the bootstrapped results show only a significant positive effect of the number of response options.

To visualize the effect of the number of response options on the items rest-correlation Figure 2 is presented. The figure shows the effect of number of response options varying between two and ten options, with all other explanatory variables equal to their overall mean. In our experiment the maximum number of response options were seven. In other words, there is no empirical evidence for the effect of offering eight response options and more. We included these number of response options in the figure to show the (theoretical) predicted decrease in the difference between both measures after offering seven/eight response options, and to show that offering seven or eight response options seems to be the theoretical turning point concerning the effect on the difference between both measures.

Figure 2 shows a rising line, the more options are offered the bigger the difference between the first and the second measurement. Theoretically, the difference decreases after eight options.

4. Conclusion

In this study the effects of three question characteristics, negatively formulated questions, offering a neutral midpoint and the number of response options, on the reliability was tested.

Surprisingly, the results in this study do not show an effect of negatively formulated questions on the reliability (stability) of responses in general. This result is surprising because it is not in concordance with the satisficing theory and not with empirical results found with adults (Andrews, 1984; Knäuper et al., 1997)

and children (Benson & Hocevar, 1985; Borgers & Hox, 2000; Marsh, 1986). That this result does not appear in our study can be the result of the definition of our dependent variable as stability over time. Apparently, children can respond stably over time and within a scale, while responding consistently different on negatively formulated questions. So, the interpretation of negative questions is stable over all questions, but different from the positively formulated questions. On more than half of the questions significant differences in responses between the negatively formulated and positively formulated questions were found. Marsh (1986) also found that the interpretation of negatively formulated question is very different from positively formulated questions.

The most stable result in this study is the effect of the number of response options on the reliability of responses. However, it is not a linear effect. The stability of responses within the scale increased with the number of responses options offered, up to six options. Offering seven or more options appears to cause a decrease in scale reliability. The effect of increasing the number of response options on the stability over time shows the opposite effect. Increasing the number of response options cause an increase of the absolute difference between the first and the second measurement up to offering seven response options. This is logical, since offering more response options provides room for larger absolute differences. After about eight response options, the predicted difference declined. For the relative difference no effects were found.

Earlier studies showed that more response options offered, the less reliable (stability within the scale) the responses were and more item non-response was produced. For developing questionnaires for children this means that the researcher should weigh one effect against the other. The results are not as clear with children as they are for adults, where increasing the number of response options definitely results in an increase of reliability of responses with an optimum around seven response options. Besides, the results on the relative difference between both measures show a positive effect of offering a midpoint. Questions with a response scale that offers a neutral midpoint produce a bigger relative difference between both measures. Take all above-mentioned results into consideration; it would appear that offering about four response options is optimal with children as respondents.

The variance of the outcome over children is significant, indicating that different children do not react the same to the questions. This indicates possible interactions between child and question characteristics. However, given the limitations of our sample, it was not possible to model these interactions with the data at our disposal. A second limitation of this study is that the data are collected using computer-assisted self-interviewing (CASI) with a standing panel of respondents. In general, using computer assisted questionnaires produces a better quality of responses (de Leeuw et al., 1998). In addition, panel respondents are trained respondents compared to respondents in general. Both aspects can affect the responses and the results in this study, resulting in smaller effects that would be found otherwise.

Notes

1. $Z = (\ln(1+r_{ir}) - \ln(1-r_{ir}))$; and the inverse transformation is $r_{ir} = (\exp(2*Z)-1)/(\exp(2*Z)+1)$.
2. The Wald test is used to test the hypothesis that both coefficients (the number of response options and the squared number of response options) are zero (Goldstein, 1995; Hox, 2002). This resulted in $\chi^2 = 8.45$ $df = 2$; $p = 0.015$, which shows that both coefficients do not equal zero.

References

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.

Amato, P. R. & Ochiltree, G. (1987). Interviewing children about their families: A note on data quality. *Journal of Marriage and the Family* 49: 669–675.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly* 48: 409–442.

Ayidiya, S. A. & McClendon, M. J. (1990). Response effects in mail surveys. *Public Opinion Quarterly* 54: 229–247.

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.

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 Preliminary 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. & Hox, J. J. (2000, 3–6 October). *Reliability of Responses in Questionnaire Research with Children*. Paper presented at the Fifth International Conference on Logic and Methodology, Cologne, Germany.

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 (eds.), *Official Statistics in a Changing World*. Stockholm: SCB, pp. 133–140.

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.

Bradburn, N. M. (1969). *The Structure of Well-Being*. Chicago: Aldine.

Bradburn, N. M. & Caplovitz, D. (1965). *Reports of Happiness*. Chicago: Aldine.

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*.

De 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.

Felix, J. & Sikkel, D. (1999). Attrition bias in telepanel research. *Kwantitatieve Methoden* 20(61): 101–110.

Flavell, J. H. (1985). *Cognitive Development*, 2nd edn. Englewood Cliffs, New Jersey: Prentice-Hall International, Inc.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd edn. London: Edward Arnold.

Groves, R. (1989). *Survey Error and Survey Costs.* New York: Wiley.

Hayes, W. L. (1974). *Statistics for Social Sciences*, 2nd edn. London: Holt, Rinehart and Winston.

Holaday, B. & Turner-Henson, A. (1989). Response effects in surveys with school-age children. *Nursing Research* 38(4): 248–250.

Hox, J. J. (1986). *Het gebruik van hulptheorieen bij operationalisering. Een studie rond het begrip subjectief welbevinden*. Unpublished PhD, University of Amsterdam, Amsterdam.

Hox, J. J. (2002, forthcoming). *Multilevel Analysis of Regression and Structural Equation Models*. Mahwah, NJ: Erlbaum.

Knäper, 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. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5: 213–236.

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

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

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

Leeuw, E. D. de, Hox, J. J. & Snijkers, G. (1998). The effect of computer-assisted interviewing on data quality. A review. In: B. Blyth (ed.), *Market Research and Information Technology*. ESOMAR Monograph. Amsterdam: Esomar, pp. 173–198.

Linden, F. J. V. D., Dijkman, T. & Roeders, P. J. B. (1983). *Metingen van kenmerken van het persoonssysteem en sociale systeem*. Nijmegen: Hoogveld Instituut.

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: Wiley.

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

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

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.

Piaget, J. (1929). *Introduction to the Child's Conception of the World.* New York: Harcourt.

Piaget, J. (1955). *The Language and Thought of the Child*. New York: Meridian Books.

Raaijmakers, Q. A. W., Van Hoof, A., 't Hart, H., Verbogt, T. F. M. A. & Vollebergh, W. A. M. (2000). Adolescents' midpoint responses on Liker-type scale items: Neutral or missing values? *International Journal of Public Opinion Research* 12(2).

Raudenbush, S. W. (1994). Random effects models. In: C. H. & L. V. Hedges (eds.), *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

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

Rodgers, W. L., Herzog, A. R. & Andrews, F. M. (1988). Interviewing older adults: Validity of self-reports of satisfaction. *Psychology and Aging* 3(1): 264–272.

Rosenberg, M. (1965). *Society and the Adolescents Self Image*. Princeton, NJ: Princeton University Press.

Rosenberg, M. (1979). *Conceiving the Self*. New York: Basic.

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. (1999). Cognitive aging and accuracy of self-report. In: D. Park & N. Schwarz (eds.), *Cognitive Aging: A Primer*. Philadelphia, PA: Psychology Press, pp. 233–252.

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, Stockholm.

Schwarz, N., Park, D., Knäuper, B. & Sudman, S. (1998). *Cognition, Aging, and Self-Reports*. Philadelphia, PA: Psychology Press.

Scott, J. (1997). Children as respondents: Methods for improving data quality. In: L. Lyberg (ed.), *Survey Measurements and Process Quality*. New York: Wiley.

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.

Vaillancourt, P. M. (1973). Stability of children's survey responses. *Public Opinion Quarterly* 373–387.