Leading education and social research
Institute of Education
University of London

Millennium Cohort Study

How do children answer questions about frequencies and quantities?

Evidence from a large-scale field test

Kate Smith and Lucinda Platt

December 2013

Centre for Longitudinal Studies
Following lives from birth and through the adult years
www.cls.ioe.ac.uk

CLS is an ESRC Resource Centre based at the Institute of Education, University of London

How do children answer questions about frequencies and quantities? Evidence from a large-scale field test

Kate Smith and Lucinda Platt

December 2013

Contact the authors:
Kate Smith
Institute of Education, University of London
Email: k.smith@ioe.ac.uk

Lucinda Platt
Institute of Education, University of London
Email: l.platt@ioe.ac.uk

First published in December 2013 by the
Centre for Longitudinal Studies,
Institute of Education, University of London
20 Bedford Way
London WC1H 0AL
www.cls.ioe.ac.uk

ISBN 978-1-906929-73-2

The Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre based at the Institution of Education (IOE), University of London. It manages three internationally-renowned birth cohort studies: the 1958 National Child Development Study, the 1970 British Cohort Study and the Millennium Cohort Study. For more information, visit www.cls.ioe.ac.uk.

The views expressed in this work are those of the author and do not necessarily reflect the views of CLS, the IOE or the ESRC. All errors and omissions remain those of the author.

This document is available in alternative formats.
Please contact the Centre for Longitudinal Studies.
tel: +44 (0)20 7612 6875
email: clsfeedback@ioe.ac.uk

Contents

Abstract

There is still much we have to learn about the best ways of obtaining accurate and comprehensive information from children without undue burden. This paper describes the findings from two experiments undertaken to develop response categories to maximise data quality for a series of questions intended to be useable with 11 year olds. The experiments were part of the instrument development for the child self-completion questionnaire for the fifth (Age 11) Survey of the Millennium Cohort Study (MCS), a large-scale national birth cohort study. The aim was to ascertain how to get the best quality estimates of frequencies of activity or quantities of consumption from children in response to potentially sensitive questions in a self-completion context. Specifically we set out to compare more and less specific response categories of two different types: descriptive versus numerical frequency responses, and banded versus open ended quantitative questions. To this end, we conducted two randomised experiments to evaluate responses to questions on bullying and alcohol consumption with a large-scale sample of children aged 11-16. We find that how response options to questions are presented matters for children's responses. Questions need to be clear to avoid ambiguity in interpretation. When asking questions about quantities, such as amount of alcoholic drinks consumed, we conclude that better quality data is provided by fixed category rather than open ended response options.

Background

The aim of survey questions is to achieve specificity and accuracy of the data that they generate. There may, however, be a trade-off between these two aims when asking questions which require respondents to give a specific response to an amount or the periodicity of an event, e.g. how often they are bullied or how much they drink. Specificity may involve greater cognitive burden which can lead to greater error and/or more non response. On the other hand, a requirement for lower specificity provides less information and therefore results in less analytical leverage. This raises questions about what is optimal.

The survey literature suggests that issues of cognitive burden may be enhanced when seeking responses from children and also when questions are sensitive (Borgers et al 2000; Scott 1997). Moreover, children answer questions differently from adults (Borgers et al. 2000 & 2004). There is some evidence that challenging questions (including sensitive ones) may lead respondents to display satisficing – ticking any box or selecting the first answer in a list. Borgers et al (2003 & 2004) have found that children are particularly prone to satisficing as they have not yet developed the ability to fully think through their answers.

Research also suggests that children may be able to handle different types of questions at different ages (Borgers et al 2003) and change in the way they can answer or handle questions at around the age of 11 (Scott 2008). There appears to be a cusp in developmental capabilities at this age and Scott suggests that children of 11 and older can answer standardised surveys much more like adults. There is, however, clearly a range, with some children of this age responding more like younger children and some more like teenagers, depending on such issues as literacy levels, cognitive ability and the context of the questions. Moreover, this does not imply that there are no measurement issues associated with setting questions and establishing response for children aged 11-16.

Children's cognitive, social and communication skills are not yet fully developed in the early adolescent period. For instance de Leeuw (2011) suggests that early adolescents (children over the age of 12) process the information needed to respond to questions around 1.5 times slower than adults, so they need adequate time to answer questions. Fuchs (2005) found that when presented with a long list, children aged 10-13 were twice as likely as their older peers to select the first item in a list. Children aged between 11 and 16 are also particularly context sensitive (Borgers et al 2000), which means that the sensitivity of a topic and its placement become especially important in influencing response.

In addition, general pre-testing for questions addressed to children is vital, since children tend to understand questions differently from the way the researcher may have intended. For instance, Scott (1997) found that if an interviewer read out a question about 'people my age' children tried to guess the age of the interviewer rather than thinking about children their own age. It has also been found that small errors in question design such as ambiguity, can lead to greater errors from children and adolescents. (Borgers et al 2000).

When faced with developing a self-completion questionnaire for children aged 11 in a large, nationally representative UK child cohort study, the Millennium Cohort Study (MCS),[1] we therefore recognised the importance of paying careful attention to question design and wording. Specifically, we addressed issues not only of question wording but also of the response categories supplied. In addition, the subject of the questions was also relevant to responses. At age 7, the MCS children completed their own paper questionnaire for the first time. For the Age 11 Survey they were to be asked to complete a considerably longer questionnaire covering topics relevant to their age, including those of a sensitive nature such as questions about risky behaviours including alcohol consumption and experience of bullying, both as victim and perpetrator.

The standard practice on large scale surveys is to evaluate new data collection instruments using cognitive testing (Presser et al 2004) to investigate how well questions perform and are understood. This is typically implemented by conducting in-depth, semi structured interviews with a small number of respondents who share key characteristics with those to be surveyed in the main study. Most cognitive testing tends to be small scale and lacks the sample sizes to test different versions of questions or quantitatively evaluate questions. It is also important to recognise that children, even of a similar age, are highly varied and small scale cognitive testing may not pick out the extent of individual variation or heterogeneity within the age group. Fowler (2004) has argued that such detailed cognitive testing of working comprehension should be supplemented and supported by greater experimental testing of questionnaire content. Experimental field tests have the benefit that they can provide explicit criteria by which to evaluate different options.

With this in mind, we therefore conducted two randomised experiments to evaluate responses to questions on bullying and alcohol consumption with a large-scale sample of children aged 11-16. This enabled us to evaluate different response options to questions

[1] The MCS is a national longitudinal cohort study of nearly 19,000 children born in 400 areas of the UK between 2000/2001. It is a large scale, multi-purpose, multi-disciplinary quantitative study. To date there have been five waves of data collection, at 9 months and ages 3, 5, 7 and 11. The fifth survey at age 11 took place during 2012, when the children were in their last year of primary school. Interviews were carried out in the home with resident parents. At each survey since the age of 3, cognitive assessments and physical measurements have been conducted with the children.

that had already had prototype versions used in other studies and which we considered likely to be included in the age 11 sweep of the Millennium Cohort Study. There was thus a clear practical imperative to ensure the effective structuring of these questions and their responses.

The rest of the paper describes the two experimental question tests, how they were designed to evaluate different response options, and our findings. Specifically, section 2 outlines the data and approach; section 3 describes the first test of frequency responses in relation to questions about bullying; section 4 outlines the testing of alternative quantity response options designed to capture alcohol consumption; and section 5 briefly concludes.

Data and methods

In order to assess the acceptability and relevance of proposed questions for the Age 11 Survey of the MCS, a variety of question testing was carried out as part of the development work. Qualitative pre-pilot research was conducted with 11 year olds and their parents in order to test the acceptability and relevance of new and potentially sensitive topics such as the onset of puberty, experience of bullying, anti-social behaviour and risky behaviours including, drinking, smoking, and drug taking. Having ascertained potentially relevant question domains and exemplar questions, cognitive testing was used to assess the understanding and appropriateness of particular exemplar questions for our target population. At the same time, additional work was carried out to investigate the effect of modifying answer categories. It is this third element of the development work that we describe here.

The experiments were fielded in a large scale omnibus survey with a panel of children. The Omnibus question testing was carried out by Ipsos MORI between 23/01/2011 to 15/04/2011 with 2,739 school children aged between 10-15 years in 114 schools in England and Wales. The children were given a 20 page paper self-completion questionnaire to complete, which contained, embedded amongst other topics of relevance to this age group, our experimental questions. Specifically, the questionnaire included tests of a) response categories for frequency of bullying questions; and b) a comparison of responses for quantities of alcohol consumed over different periods offered as either banded options in grids or as direct entry in open-answer boxes. We discuss these two experiments and their findings in turn.

Experiment 1: Comparing descriptive response categories versus numeric response categories for questions on frequency of experiencing bullying by children

The issue of bullying amongst children is a major policy concern. There is evidence that it is frequently experienced by children of all ages, but was especially high among our target age range for the MCS. Just under half of all children say that they have been bullied at some point in school and around 20 per cent have experienced this outside school. These figures rise to over a half and over a quarter for those aged around 11 (DCSF 2010). Bullying is damaging for children's well-being and can affect their academic performance and school

participation (Wolke et al 2000; Woods and Wolke 2004). It potentially has long-term consequences (Wolke et al. 2013), and may therefore be an important route to understanding adult wellbeing.

This it was considered a key area of value for inclusion in MCS by the scientific community. A proposed prototype on bullying types and frequencies was that already fielded in a self-completion instrument targeted at 10-15-year olds in *Understanding Society*: the UK Household Longitudinal Study (UKHLS). The UKHLS is a large scale household panel study of around 40,000 British households. The UKHLS bullying questions had already been developed by experts in the field and covered experience of bullying both as victim and perpetrator, as well as distinguishing between physical, verbal or 'relational' bullying, and that taking place at school and by siblings (Woods and Wolke 2004 op cit; Wolke and Skew 2011).

It is clearly important to ask questions that can be understood by children and answered in a meaningful way to capture the true prevalence and the immediate and longer term impact of bullying. We were therefore concerned to ensure that the questions met stringent standards for question design and clarity, particularly given the fact that they were to be posed to relatively young children.

The original questions used response categories which combined the quantity of frequency with descriptive judgement – or a "vague quantifier". For example "Not much" was classed as "(1-3 times in the last 6 months)", "Quite a lot" was further defined as "more than 4 times in the last 6 months"; and "A lot" was considered to be "a few times every week". It was not clear that these 'mixed' response categories were unambiguous, since children might regard 'A lot' in ways other than those provided by the additional definition, and it would not necessarily be clear to them or to researchers which element they should regard as being the appropriate one for how they evaluated their experience. Bradburn and Miles (1979) found that offering 'vague quantifiers' (such as 'very often', 'fairly often', 'not at all often') as responses to questions about frequency of behaviours appeared to be interpreted very differently by adults when a value was assigned to the categories and Borgers (2011) suggests that these type of responses are particularly problematic for children as they need clear definitions.

We therefore set out to test the responses to the same question when the separate components of the response types – descriptive categories versus numeric categories – were offered to the same children. This enabled us to investigate whether respondents gave consistent answers over the two different ways of framing the question. That is, when asked independently do children regard the vague quantifier "not much" as equating to "1-3 times in the last 6 months"? If not, we wanted to ascertain which might be the more appropriate – or consistent – way to seek the information. Our hypothesis was that they would interpret the two elements differently, giving inconsistent answers across measures. But we were unclear as to which element might be prioritised when faced with a combined choice.

We therefore positioned the two forms of question – one with the descriptive response categories only and one with the frequency response categories only – proximately in the questionnaire, so that we could compare responses to the different forms of the question by the same children. Figure one illustrates the three versions of the question for physical

bullying.[2] That is, it shows the version with the descriptive response categories, the version with the frequency response categories, and the original 'combined' version from *Understanding Society*.

Figure1

Physical bullying question omnibus version 1: descriptive response categories only

How much have you been physically bullied at school, for example, getting pushed around, hit or threatened, or having belongings stolen? Please answer about the last 6 months

PLEASE TICK □ ONE BOX ONLY

Never	
Not much	
Quite a lot	
A lot	

Physical bullying question omnibus version 2: frequency response categories only

And can I just check, how often in the last 6 months have you been physically bullied at school, for example, getting pushed around, hit or threatened, or having belongings stolen?

PLEASE TICK □ ONE BOX ONLY

Never	
1-3 times	
More than 4 times	
A few times every week	

Physical bullying question *Understanding Society* combined version: descriptive and frequency responses

How much have you been physically bullied at school, for example, getting pushed around, hit or threatened, or having belongings stolen?

Never	
Not much (1-3 times in last 6 months)	
Quite a lot (more than 4 times in last 6 months)	
A lot (a few times every week)	

[2] There were also separate questions for verbal bullying, which took the form: 'How much have you been bullied in other ways at school, such as getting called names, getting left out of games or having nasty stories spread about you on purpose'. In this paper we focus on physical bullying, although we also completed the analysis for the verbal bullying and refer to the verbal bullying responses where appropriate.

Results of experiment 1

In order to investigate how children interpret questions that combine a description and a frequency, we compared the responses by the same children across the two separate forms of the questions and the extent to which there was congruence between them. We also compared the distributions for the two different individual questions to the combined version of the question in *Understanding Society* (where the children cover the same age range) in order to ascertain which distribution (descriptive or frequency) mapped more closely onto that combined version. Table 1 shows the distribution for physical bullying according to *Understanding Society* (combined) and the two experimental versions (descriptive and frequency) asked in the Omnibus survey.

Table 1: Distributions of bullying according to different question approaches

Frequency reported being bullied	**Combined Q (UKHLS) %**	**Descriptive Q (Omnibus) %**	**Frequency Q (Omnibus) %**
Never	79.7	59.6	74.4
Not much / 1-3 times	15.5	30.3	18.8
Quite a lot / 4 or more times	2.8	7.5	3.5
A lot / a few times a week	2.0	2.6	3.3
Total (n)	**100 (4,854)**	**100 (2,639)**	**100 (2,639)**
Missing	*0.8 (45)*	*4 (100)*	*4 (100)*

Source: Understanding Society, Wave 1, Youth Questionnaire, Weighted percentages; Schools Omnibus Survey, Weighted percentages

Table 1 shows a striking similarity overall between the distributions for the combined (UKHLS) and frequency (Omnibus) form of the questions, although the percentages for "a lot" are more similar across the combined and descriptive measures.

What is also worth noting is the fact that the descriptive form of asking seems to influence the tendency to report "never". That is, there are more children who report "never" when the other options have the number of times attached rather than just a description.

This can be seen further in Table 2, where the evaluative and numeric responses from the same children in the Omnibus survey are tabulated against each other. Of those reporting 'never' on the numeric measure nearly 20 per cent considered this to be 'not much' when no numbers were attached to the categories. (By contrast, over 98 per cent of children reporting 'never' on the evaluative category also put 'never' when offered the numeric category: see Table 3.) Overall, what Table 2 shows is that there is a strong diagonal distribution – that is, broadly speaking children 'agree' that 1-3 times in the last 6 months is 'not much', that 4 or more times is 'quite a lot' and that a few times as week is 'a lot'. However, there was substantial variation. Of the children who reported being bullied 4 or more times in the last 6 months, 35 per cent regarded this as 'not much'; and of those who reported being bullied a few times a week fewer than half thought this was 'a lot', with over a third thinking it only 'quite a lot' and 17 per cent deeming it to be 'not much'.

We cannot assume that the numeric answer is more accurate in some way than the evaluative answer – but it does suggest that there is some mismatch between evaluations of

what constitutes a lot of bullying, and the expert judgement that it is 'a few times a week' – by the victims themselves. For many of them, 'a few times a week' is not evaluated as 'a lot' – a potentially damning judgement on the regularity of bullying.

Table 2: Proportion of children in each numeric category reporting responses in the different evaluative categories (column percentages)

	Never **%**	**1-3 times** **%**	**4 or more times** **%**	**A few times a week** **%**
Never	**77.9**	5.9	0	0
Not much	19.1	**74.2**	34.9	16.7
Quite a lot	1.7	16.8	**54.3**	36.4
A lot	0.2	1.7	10.9	**45.8**
Missing	*1.3*	*1.4*	*0*	*1.1*
Total (n)	100 (1978)	100 (491)	100 (86)	100 (84)
Total off diagonal	22.1	25.8	45.7	54.2

If we take the opposite perspective, and look at the numeric interpretation of the evaluative labels we see, from Table 3, that of those children thinking they were bullied 'quite a lot', over 40 per cent deemed this to be 1-3 times in the last 6 months – rather than the 'correct' 4 or more times. On the other hand 16 per cent of these children thought that 'quite a lot' equated to a few times a week. For children reporting being bullied 'a lot' nearly two thirds categorised this as 'a few times a week'. Thus there is a closer match between the definition of 'a lot' as a few times a week, for children selecting 'a lot', than there is between those who are bullied a few times a week claiming it as 'a lot'.

Table 3: Proportion in each numeric category reporting responses in the different evaluative categories (row percentages)

	Never **%**	**1-3 times** **%**	**4 or more times** **%**	**A few times a week** **%**	***Missing*** ***%***	**Total (n)**	**Total off diagonal** **%**
Never	**97.2**	1.9	0	0	*0.9*	100 (1579)	2.8
Not much	46.8	**46.2**	4.1	1.8	*1.2*	100 (796)	54.8
Quite a lot	16.4	42.3	**25.4**	15.8	*0*	100 (193)	74.6
A lot	4.9	12.4	14.5	**56.7**	*11.5*	100 (71)	43.3

What is also interesting in Table 3 is the relatively high proportion (11.5 per cent) of those children stating they have been bullied 'a lot' who can't put a value on it and are coded as missing. They seem to feel they are being bullied 'a lot' but that feeling does not correspond to the rates offered. In addition a substantial proportion (47 per cent) of children who say 'not much' to the evaluative question classify this as 'never' in the numeric question. This

suggests that there is a sense for these children of being bullied, which can't be enumerated or of it being 'around' an amount or varying which also can't adequately be enumerated by the responses on offer. This reinforces the impression that the evaluative and 'objective' frequencies may be capturing different aspects of the children's experiences.

In relation to what this tells us about the design of bullying questions, it clearly matters quite substantially how the information is presented. It is worth noting that the results for the verbal bullying questions were very similar and consistent with those for the physical bullying ones.[3] The comparison of the distributions with those in *Understanding Society* suggests that in some ways providing a 'numeric' account provides more consistent information than the evaluative format. This is important if we want to be able to compare frequencies over time or across sources. It confirms other research which has found that with regard to children it is particularly important to fully label scales. It appears that children will use the numeric frequency information and respond to that, rather than take the evaluative information. Moreover the differences do not just apply to the categories of 'victimhood', since it would seem that the numeric frequency responses encourage children to select 'never' – or discourage them from selecting a substantive category, compared to when evaluative options are offered. We thus get different rates of /any/ bullying from the two approaches, not just difference of frequency / intensity. This echoes the findings that children (particularly between the ages of 7-10) struggle to answer questions about frequencies that offer vague quantifiers as responses as they need clear definitions (de Leeuw 2011).

It is likely the when children respond to the evaluative questions they are taking into account some measure of intensity as well as actual frequency, since this is implied by phrases such as 'not much'. This could help to explain why different frequencies map onto particular evaluations: particularly horrible or upsetting experiences could feel 'a lot' even if they occurred relatively infrequently. Similarly, 'never' may have come to seem a more appropriate category with the numeric options if the frequencies did not fit well onto children's perceptions of 'a bit'. Thus the experiment itself raises questions about what we are trying to measure in terms of impact when assessing bullying.

While the final question used in MCS differed from these prototypes tested, we were influenced by the results to frame the question response categories to the bullying question as frequencies rather than evaluations (see www.cls.ioe.ac.uk/mcs). This can be thought of as providing an objective measure of specific frequencies, without necessarily telling us about the child's own evaluation of their position. If anything it seems it may understate experienced bullying by requiring it to happen at particular frequencies rather than evaluations of intensity. But it offers the advantages that consistency is likely to be greater and thus change over time or across studies is likely to be better captured. While, it may be that frequency questions are more stable – in some ways and give more defined responses –we may nonetheless garner additional information about the child's, rather than the analyst's perception of their situation if we ask them evaluative questions.

[3] The tables showing the results for the verbal bullying questions are available from the authors on request.

Experiment 2: Comparing pre coded versus open ended response categories for questions on frequency of alcohol consumption in children.

There is increasing research and policy interest in the age of onset of 'risky' behaviours in children. Evidence has shown that the early onset of risky behaviours can lead to poor outcomes later in life (Agrawal et al. 2006). One of these risky behaviours is alcohol consumption (Hingson et al. 2009; Donovan 2004). There is concern both about the age that children start drinking and the quantity that they are drinking. This concern is not limited to the UK experience. Questions on alcohol consumption were planned for inclusion in the MCS Age 11 survey with the aim of investigating the variation in prevalence of early alcohol initiation and use and attitudes to alcohol. It also sought to collect prospective, nationally representative data on young adolescents' alcohol consumption behaviour, which would facilitate cross-national as well as longitudinal understanding of its consequences.

One of the proposed questions asked children to recall the number of occasions that they had had an alcoholic drink: in their lifetime, during the last 12 months and during the last 4 weeks. The question, a standard one that has been used in other studies in the US as well as the UK,[4] uses a grid ranging from '0 occasions' to '40 or more occasions'.

We considered that offering responses as explicit options might encourage children to report alcohol consumption with less accuracy – either more or less often than they had actually drunk, as the response categories may convey information about the expected answers (Krosnick and Presser 2010). Specifically, we considered that offering a "40 or more" category could increase the average alcohol reporting as this would be presented as a viable option, even though it equates to 10 drinks a week for the last four-week period. Offering open ended responses does not convey any 'expected' amounts, in the same way, and might lead to more accurate responses. We therefore hypothesised that using an open-ended question might result in lower average reported alcohol consumption across the sample, greater differentiation of responses, and more accurate information. On the other hand, we recognised that an open-ended question might increase cognitive burden by providing no prompts about viable amounts, and might hence lead to greater non-response or selecting rounded amounts (e.g. 10s, 20s etc.) (Bradburn and Miles 1979 op cit).

Two versions of the question were, therefore, tested to compare how children reported their alcohol consumption over the three periods. One version gave children the grid of pre coded responses. The second question asked children to write their answers in open ended response boxes. The two questions were randomly allocated to children based on their day of birth: those with odd birth days of the month received the closed grid question while those with even birth days received the open ended version. Figure 2 shows the two versions of the questions.

[4] The questions were funded by the US National Institute of Health (NIH) and were proposed by Dr Jennifer Maggs, Penn State University. (Grant no: 1 R01 AA019606-01A1)

Figure 2

Alcohol question version 1: pre-coded responses (odd birth days)

On how many occasions have you had an alcoholic drink…
PLEASE TICK ☐ ONE BOX ONLY FOR EACH ROW

	0 occasions	1-2 occasions	3-5 occasions	6-9 occasions	10-19 occasions	20-39 occasions	40 or more occasions
…in your lifetime?	☐	☐	☐	☐	☐	☐	☐
…during the last 12 months?	☐	☐	☐	☐	☐	☐	☐
…during the last 4 weeks?	☐	☐	☐	☐	☐	☐	☐

Alcohol question version 2: open-ended responses (even birth days)

On how many occasions have you had an alcoholic drink…
PLEASE WRITE YOUR ANSWERS BELOW

…in your lifetime? …… occasions

…during the last 12 months? …… occasions

…during the last 4 weeks? …… occasions

One disadvantage of the second version was that the children did not get an instruction to enter ‘0’ or write ‘none’ if they had not had an alcoholic drink in any of the periods listed. We considered that this might influence the responses as those who had not consumed any alcohol might provide no response rather than ‘correctly’ filling in a zero.

Results of experiment 2

We first compared the number of children who gave responses to the number of occasions they had a drink for each of the three time frames. We recoded the open responses to the same categories as the closed options.

Table 4 shows that there was significantly greater non response in the open ended version of the question than in the closed. In the case of lifetime consumption of alcohol, more than

four times as many children did not respond to the open question as compared to the grid (17.8 per cent vs 3.7 per cent).

Table 4 Number of times child has had an alcoholic drink in lifetime, last year and last 4 weeks

	In lifetime		**In last 12 months**		**In last 4 weeks**	
	Closed %	Open %	Closed %	Open %	Closed %	Open %
0	**22.1**	**18.6**	**29.7**	**26.8**	**58.4**	**54.2**
1-2	20.3	13.9	26.6	20.7	21.8	17.4
3-5	13.3	14.8	15.5	13.8	8.4	7.5
6-9	12.4	6.5	9.5	5.4	3.6	1.7
10-19	12.1	12.7	7.2	9.3	1.1	3.2
20-39	8.3	9.2	4.4	5.1	0.1	1.3
40+	7.9	6.6	2.0	2.7	0.5	0.7
Mean[5] number of drinks	1.1	1.7	5.6	7.9	12.3	14.7
Missing	***3.7***	***17.8***	***5.2***	***16.2***	***6.2***	***14.0***
Total (n)	**100** (N=1333)	**100** (N=1239)	**100** (N=1333)	**100** (N=1239)	**100** (N=1333)	**100** (N=1239)

It would appear from this that the grid performed better by providing more information from the children. Moreover, contrary to our hypothesis, average number of drinks was not higher in the banded compared to the open ended questions. The estimates were largely consistent, but were slightly lower across the periods for the banded responses.

However, as noted, it was possible that the lack of instruction to enter '0 occasions' in the open ended question, if the child had not consumed any alcohol in the period, may have led to an artificially higher rate of non-response for the open ended question. We therefore compared the distributions across the grid format and the open-ended version excluding all missing responses. If non-responders to the open-ended format were primarily non-drinkers

[5] The mean was calculated for the banded responses by taking the midpoint of the band as the response, and by using the midpoint between 40 and the actual upper value in the open ended response for the top band. For the last four weeks option, there were only a few responses specifying over 40 in either the banded or the open ended options, so in this case, for robustness, responses over 40 were simply excluded from the calculations of means.

we would expect to see between around 8 and 14 per cent lower reporting of 0 among responders to the open-ended question compared to the closed format. Table 5 shows the distributions for the different periods when missing responses were excluded from the analysis.

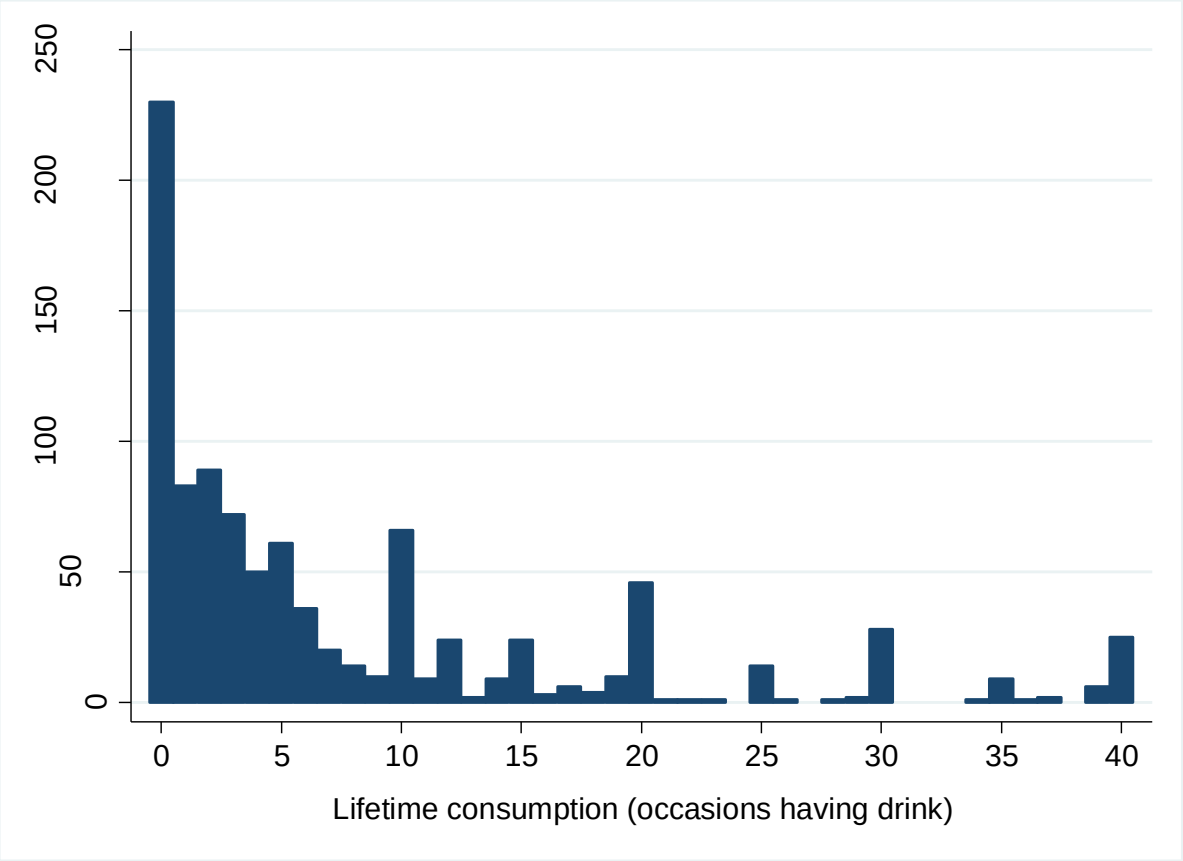
Table 5 Number of times child has had an alcoholic drink in lifetime, last year and last 4 weeks - excluding missing cases

	Lifetime		**In last year**		**In last 4 weeks**	
	Closed %	Open %	Closed %	Open %	Closed %	Open %
0	23.0	22.6	31.3	32.0	62.2	63.0
1-2	21.1	16.9	28.0	24.7	23.2	20.2
3-5	13.8	18.0	16.4	16.5	9.0	8.7
6-9	12.9	7.9	10.1	6.5	3.8	2.0
10-19	12.5	15.4	7.6	11.1	1.1	3.8
20-39	8.6	11.2	4.6	6.1	0.1	1.5
40+	8.2	8.1	2.1	3.3	0.6	0.8
Total (n)	100 (N=1,284)	100 (N=1,018)	100 (N=1,264)	100 (N=1,038)	100 (N=1,250)	100 (N=1,066)

In fact, we see that the distributions are remarkably similar across the two question formats once non-responders are excluded. This suggests that the non-responders to the open-ended question were distributed across consumers and non-consumers, rather than being concentrated among the latter. Thus, children did not appear to be using non-response to the open question as a proxy for not having consumed any alcohol. Instead, it suggested that the cognitive burden of the question was more challenging regardless of alcohol consumption rates.

We also noted that the distributions that included a 5 or a 10 in the category had higher rates in the open-ended than in the closed questions. This indicated that there might have been some clustering among responses to the open-ended question. We explored this point: Figure 3 illustrates how the open question elicited bunching of responses around the round numbers (5, 10, 20, 30 etc).

Figure 3: Heaping in open ended responses



While it is complex to try to evaluate which is the 'true' distribution of consumption when only bands are provided, the results suggest that estimates are better served by offering a range rather than the open option which might lead to rounding (either up or down).

Overall, it appeared that the grid seemed to work better than the open ended question. We hypothesised that the open ended question might result in lower average estimates and in more definition of responses. Neither of these results were found. Conversely, we found that the open ended question incurred a higher degree of non-response, higher average consumption estimates, and in a preponderance of 'focal' estimates (at #5 and #0). The model of survey response popularized by Tourangeau, Rips and Rasinski (2004) breaks down the response process into four sub-processes: comprehension of the item, retrieval of relevant information, use of that information to make required judgements, and selection and reporting of an answer. These processes are affected by characteristics of the interviewer, question wording and survey instrument and, of course, the characteristics of the respondent themselves. Our test was conducted as a self-completion so there was no interviewer effect; and because the allocation was random we could discount variations in respondent characteristics. This left, as the source of variation, the question wording.

The retrieval of relevant information for a frequency of occurrence across a series of specified recall periods can be relatively demanding. This is likely to be particularly the case for children, for whom such questions are likely to be more challenging. Therefore selecting an answer from a list, compared to retrieving it without any prompts, eases the burden. The existence of the list also helps to make it clear what information is sought. The fact that the

proportions of children reporting having had no alcoholic drinks were very similar when missing cases were excluded suggests that those who drank were as likely as non-drinkers to be non-respondents. Thus we cannot attribute the differences to lack of clarity in the question for those with no alcohol consumption. Rather, the open ended question seemed to be more generally demanding.

We also cognitively tested these questions and while an equal number of children expressed a preference for both the closed and open ended questions, the data collected in the grid was better completed and more accurate. Overall then, there was a compelling case that we gained better quality data from the grid version, and that the potential benefit of greater specificity from the open-ended question was not fulfilled in practice, even when we only considered those who gave a response. The practical consequence of this experiment was that the banded grid response options were adopted in the Age 11 Survey of the MCS. We thereby expected to elicit the greatest information from the children in a way that was clearest and provided the least cognitive burden.

Conclusions

The experiments we have conducted emphasised the power of quantitative question testing to establish preferred design, specifically in relation to the issue of response categories, rather than the perhaps more familiar territory of question wording. While cognitive testing continues to provide an important role in ascertaining comprehension and misapprehension of questions, field testing enables evaluation against specific quality criteria in order to help implement the best measure for a large scale survey.

This was particularly important in these examples, as we were concerned with providing optimised questions for children, for whom responding clearly and consistently is more challenging – ambiguities are enhanced, satisficing is more likely, and responses are more sensitive to cognitive burden.

Questions for adults could also benefit from more similar approaches, alongside the standard approach of utilising cognitive testing, since we may be making assumptions about their response capabilities that we are more cautious about making with children, but which are no better founded. Our findings endorse Presser's call for greater attention to large-scale field testing of questionnaire content, and have shown their potential for highlighting the strengths and weaknesses of alternative response categories and giving new insight into how these shape our understanding of the frequency and amount of different activities.

References

Agrawal, A., Grant, J. D., Waldron, M., Duncan, A. E., Scherrer, J. F., Lynskey, M. T., et al. (2006). Risk for initiation of substance use as a function of age of onset of cigarette, alcohol and cannabis use: Findings in a Midwestern female twin cohort. *Preventive Medicine*, 43, 125-128.

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

Borgers, N; Hox, J & Sikkel ,D (2003). Response Quality in Survey Research with Children and Adolescents: the effect of labelled response options and vague quantifiers. *International Journal of Public Opinion Research* Vol.15 No 1:83-94.

Bradburn N.M., Miles, C. (1979). Vague Quantifiers. *Public Opinion Quarterly Vol 43 (1) 92-101*

Bradburn, N. M., Sudman, S., and Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design.* San Francisco, CA: Jossey-Bass.

Chamberlain, T., George, N., Golden, S., Walker, F., Benton, T. (2010). *Tell Us4 National Report .* Research Report DCSF-RR218, NFER.

de Leeuw, E. (2011). *Improving Data Quality when Surveying Children and Adolescents: Cognitive and Social Development and its Role in Questionnaire Construction and Pretesting.* Report for the Annual Meeting of the Academy of Finland.

Donovan, J. E. (2004). Adolescent alcohol initiation: A review of psychosocial risk factors. *Journal of Adolescent Health*, 35, 529.e7–529.e18.

Fowler, F. J. (2004). 'The Case for More Split-Sample Experiments in Developing Survey Instruments' in *Methods for Testing and Evaluating Survey Questionnaires* (eds S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin and E. Singer), John Wiley & Sons, Inc., Hoboken, NJ, USA. Ch9

Fuchs, M. (2005). Children and Adolescents as Respondents, Experiments on Question Order, Response Order, Scale effects and the Effect of Numeric Values Associated with Response Options. *Journal of Official Statistics* Vol. 21(4): 701-725.

Hingson, R., Edwards, E. M., Heeren, T., & Rosenbloom, D. (2009). Age of drinking onset and injuries, motor vehicle crashes, and physical fights after drinking and when not drinking. *Alcoholism: Clinical and Experimental Research*, 33, 783-790.

Krosnick, J.A. & Presser, S. (2010). 'Question and Questionnaire Design' in *Handbook of Survey Research, Second Edition.* Eds. Marsden, P. V. and Wright, J. D., Emerald Group. Ch 9

McGee A. & d'Ardenne, J. (2009). '*Netting a winner': tackling ways to question children online.* London: NatCen

Presser et al. 2004. Methods for testing and evaluating survey questions. *Public Opinion Quarterly* 68(1): 109-130

Scott, J. (1997). 'Children as Respondents: Methods for Improving Data Quality' in *Survey Measurement and process quality*. Eds. Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz and Trewlin, John Wiley

Scott, J. (2008). 'Children as Respondents: the Challenge of Quantitative Methods' in *Research with Children: Perspectives and Practices*, Eds. Christensen, P. and James, A. Second Edition. Routledge 87-108

Tourangeau, R., L. Rips and K. Rasinski (2004). *The Psychology of Survey Response.* Cambridge University Press.

Wolke, D. Copeland, W.E., Angold, A. and Costello, E.J. (2013) Impact of bullying in childhood on adult health, wealth, crime and social outcomes. *Psychological Science* DOI: 10.1177/0956797613481608

Wolke, D., Woods, S., Bloomfield, L., & Karstadt, L. (2000). The association between direct and relational bullying and behaviour problems among primary school children. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 41*(8), 989-1002.

Woods, S., & Wolke, D. (2004). Direct and relational bullying among primary school children and academic achievement. *Journal of School Psychology, 42*(2), 135-155

Wolke, D., & Skew, A. (2011). Bullied at home and at school: Relationship to Behaviour Problems & Unhappiness. *Understanding Society: Early Findings from the First Wave of the UK's Household Longitudinal Study* (Vol. 1), ISER

Centre for Longitudinal Studies
Institute of Education
20 Bedford Way
London WC1H 0AL
Tel: 020 7612 6860
Fax: 020 7612 6880
Email cls@ioe.ac.uk
Web www.cls.ioe.ac.uk