The Big Ask, Qualitative Analysis Methodology

Data collection

'The Big Ask' survey was a publicly available, online questionnaire that was launched by the Office of the Children's Commissioner on 19 April 2021 and closed on 28 May 2021. It was supported by a range of visits to schools and local areas, as well as media activity using national and local press. The Children's Commissioner's Office used regular communication with schools, local authorities, charities, and other organisations working with children to help disseminate the survey and maximise awareness and response rates.

The aim of the survey was to collect information from children aged 4 – 17. As this age range crosses several developmental levels, different survey forms were designed for the age groups 4-5, 6-8, 9-12 and 13-17.

The age 6-8 and 9+ surveys included one entirely qualitative question each. To ensure the questions were appropriate, the questions asked to the 6-8 age group and children aged 9+ were different, though designed to capture the same theme, to allow them to be analysed as a single set of qualitative responses.

The age 6-8 survey asked:

- 'If you could change anything to make your life better when you grow up, what would it be?'

The age 9-12 and age 13-17 surveys asked:

- 'What do you think stops children/young people in England achieving the things they want to achieve when they grow up?'

The aim of the qualitative analysis was to identify the topics and themes that children mentioned most frequently, and understand not just what these are but how they are experienced by children and how they impact children's lives.

The analytical approach for this report was determined heavily by the scale of data gathered, which rules out normal methods of qualitative analysis and synthesis. It was not possible to read every single response transcript or feed the 260,000 responses into standard qualitative analysis software such as NVivo. Data of the scale and nature gathered here required a new, novel and more experimental approach combining 'big data' analytical methods with 'traditional' qualitative methods. The former were used to generate a set of high-level topics under which the majority of responses could be categorised, while the latter enabled deeper analyses of the specific responses categorised under a topic.

Identifying and defining the topics

To identify the topics, we first generated lists of the most frequently used words across all responses and then by age-group (6-8, 9-12, 13-15 and 16-17). A simple word count would not be appropriate because it would also include 'stop words' ('I', 'think', 'to', 'and', etc) as well as multiple versions of the same word such as plurals and different

tenses. Therefore responses were processed using the 'ud-pipe'[1] model in R which performed lemmatization, the process of identifying the root word being used. For example, 'exams' and 'examination' share the root word 'exam' and 'schools' and 'schooling' share the root word 'school'. Lemmatization accounts for the different ways of referring to the same root whereas a simple word count would provide a separate word count for 'exam' and 'exams', following lemmatization uses of 'exam', 'exams' and 'examination' were counted together. A word count from the lemmatized list of words was run with stop-words removed using the 'tidytext'[2] package to identify stop-words.

The lemmatized data was also explored with the 'Rapid Automatic Keyword Extraction' (RAKE) algorithm built into the ud-pipe package[3]. RAKE is a domain independent keyword extraction algorithm which works to determine key phrases in a body of text by analysing the frequency of word appearance and its co-occurrence with other words in the text. The RAKE model produced a list of phrases (n-grams set to 3) containing nouns, verbs and adjectives. An n-gram in this context is a continuous sequence of *n* items from a sample of text after stop words have been removed.

We then had three lists of frequently used words and phrases which were the possible index words. In this context, an index word or phrase is one that denotes a topic or category. These words were then grouped according to the concepts, i.e., topics that they related to. Simple words to assign to a topic were those which were generally used to refer to only one topic. For example, 'school', 'education', 'teacher', 'gcse', 'a-levels', 'grades' are commonly used words when discussing education and are straightforward to group under that topic. Similarly, 'parent', 'family', 'mum', 'dad', 'sister', 'brother' were all used to describe families and were grouped under a 'family' topic. The initial grouping was agreed collaboratively between several researchers and topic experts who considered the use of terms in context and potential alternative uses. Random samples of responses mentioning a particular word were reviewed manually, to help the researchers understand all of the contexts in which that word was used.

In some cases, a single word could be used in multiple contexts. Where this is the case, pairs of words have been used to ensure correct categorisation. For example, children often used the word 'class' both as a reference to school (e.g. 'pay attention in class') and to society in general (e.g. 'higher social class'). These responses therefore needed to also contain another education-relevant word (e.g. 'teacher', 'lesson', 'disrupt', 'learning' etc) in order to be categorised under education, and they needed to or a society-relevant word (e.g., 'social', 'middle', 'lower', 'upper', 'poor' etc) to be categorised under the topic 'politics and society'. This is not a perfect solution as there could be overlap between the chosen 'education-relevant' or 'society-relevant' words, for example, 'children from a poor family might fall behind in class' and so this approach has likely introduced an amount of noise into the final categorisation. However, from the review of the correlations and random samples of the data we concluded that the risk of introducing noise with this approach was outweighed by the

---

[1] Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit, *R Package Documentation*, link

[2] Introduction to tidytext, *CRAN R-Project*, 10 April 2021, link

[3] Keywords_rake: Keyword identification using Rapid Automatic Keyword Extraction, *R Package Documentation*, link

benefit from capturing the large number of additional responses that this approach facilitated.

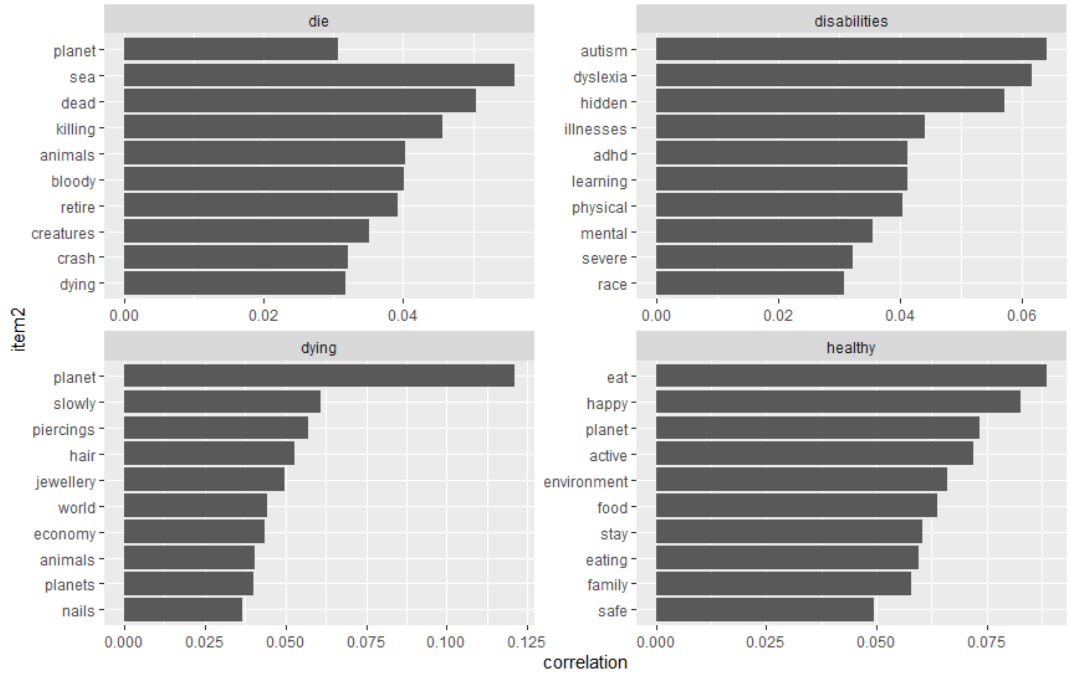
Once the words identified from the frequency/RAKE analysis and had been assigned to topics, they were then further reviewed and tested iteratively, to ensure, as far as possible that each word was a useful indicator of a given topic and was not likely to introduce significant amounts of noise into the categories.

There were two stages of testing, the first of which was pairwise correlation. This involved calculating the correlation among the index word and the other words used by the respondent in their response. The correlation test indicated how often words appeared together relative to how often they appeared separately. This was implemented using the 'tidytext' mining methodology[4] which utilises the phi coefficient, how much more likely it is that either both word X and Y appear, or neither do, than that one appears without the other. By looking at the words correlated with an index word across responses we can assess the extent to which the index word accurately represents the topic.

For example, the words 'die' and 'dying' were originally allocated to the topic of health. However, as Figure 1 shows, these words were most correlated with words like 'planet', 'sea', 'dead' or 'piercings', 'hair' and 'jewellery', respectively. In other words, the pairwise correlations revealed that the words 'die' and 'dying' were most likely to be used by children talking about the environment or about physical appearances. Therefore, these words were removed from the health topic. The word 'healthy' is introducing some noise to the health topic where responses talking about the environment are erroneously categorised to the health topic however this is relatively infrequent and is very unlikely to change the overall balance of responses within the topic.

[4] Text Mining with R: A tidy approach, *Silge & Robinson*, 2017, link

Figure 1 - Phi-correlation scores for the top 10 words most strongly correlated with 'die', 'dying', 'disabilities' and 'healthy" across all free-text responses



The pairwise correlation provided a good indication of the performance of a given index word for identifying a topic. However, as Figure 1 demonstrates, there is still an element of ambiguity around some words which required further exploration. The second stage of testing was to manually review a random sample of 200 responses which contained the index word of interest and assess the consistency of the use of the word to denote a single topic. For example, words such as ‘abuse’, ‘struggling’, ‘sad’, ‘men’ and ‘earn’ were manually reviewed and removed from the list of index words as they did not consistently refer to a single topic but were used in multiple contexts transcending several topics. ‘Earn’ for example, was not only used to refer to jobs and careers, but also to parental income as well as political issues. Therefore, these words could not be searched for in order to extract a set of responses all referring to the same topic.

In total, after the testing was complete, 9 topics were identified. These topics, and the index words phrases categorised underneath them, are given in Table 1. A response was categorised to one of these topics by searching for the presence of an index word, if the response contained one of the index words then it was tagged as referring to that topic.

Table 1 - The overarching themes, topics and index words used to categorise responses to the topic

Topic	**Index words and phrases under this topic**	**Number of responses categorised under this topic (as % of all responses)**
Education and	School Learn	79,995 (31%)

school	Education	Maths	
	Homework	A-levels	
	Teach	Academic	
	GCSE	Educated	
	Grade	Uniform	
	Study	Pupil	
	Skills	Curriculum	
	College	lessons	
Politics and society	Government	Prejudice	31,849 (12%)
	Racism	Background	
	Society	Race	
	Discrimination	Conservative	
	Capitalism	Social pressure	
	Stereotype	Politics	
	Equal	Economy	
	Gender	Classism	
	Skin	Social standard	
	Colour	Homophobia	
	Black	Economic	
	Homeless	LBGTQ	
	Racist	Racial	
	White	Culture	
	Minister	Sexist	
	Treated fairly	Societal	
	Social mobility	Religion	
	Community	Sexuality	
	Communities	Class*	
Financial considerations	Money	Debt	38,998 (15%)
	Poverty	Income	
	House	Nice home	
	Billionaire	Nice food	

	Rich	Prices	
	Wealth	Living conditions	
	Expensive	Wage	
	Cost	Better house	
	Financial	Buy things	
	Afford	Minimum pay	
	Bills	Housing	
Local area	Crime	Safety	20,488 (8%)
	Gang	Surroundings	
	Knife	Clubs	
	Drug	Sport	
	Alcohol	Facilities	
	Chav	Smoking	
	Area	Park	
	Violence	Local	
	Play	Town	
	Environment	Place	
	live	street	
	Road man		
	Facility		
Jobs and careers	Job	Career	29,714 (11%)
	Footballer	Famous	
	Vet	University	
	Doctor	Universities	
	Artist	Football player	
	Player	Work experience	
	Police officer	Qualification	
	Gamer	Horse rider	
	Singer	Dancer	
	Police man	Youtuber	
	Train driver	Nurse	
Feeling and	Mental	Physical health	48,790 (19%)

being well	Depression Anxiety Worry Stress Suicide Pressure Depressed wellbeing	Disability Illness Healthy Disabilities Junk food	
Family	Parent Family Social class Dad Mum Home life Sister Brother families	Familys Upbringing Mom Baby Divorce Husband Sibling Child abuse	39,569 (15%)
Social influences	Bully Social media Peer Influence Friend People opinion Judgement Mean people Other people Social pressure	Rude people Negative people Expectation Bullied Bullies Judging People think People discourage Wrong crowd hater	46,002 (18%)
Self-belief	Motivation Confidence Laziness Fear	Determination Motivate High expectation Work ethic	42,474 (16%)

	Believe	Negative thought	
	Failure	Lack encouragement	
	Doubt	Shyness	
	Procrastination	Belief	
	Dream	Effort	
	Attitude	Ambition	
	Esteem		

* Class was used with the rule that the response also included system, social, lower, middle, upper, divide, gap, poor or working, to differentiate from uses of the word class to refer to education.

Responses could be categorised across several topics, which is an advantage of this approach as opposed to, for example, cluster analysis or topic modelling which assigns responses to a single topic. 50% of responses referred to one topic, 28% referred to two topics, 13% referred to three topics and 1% referred to 4 or more topics. Overall, 85% of responses were categorised to at least one topic, with some response categorised under multiple topics.

Responses from 6-8 year olds less likely to be categorised compared to responses from other age groups. A review of these uncategorised responses shows that this was largely due to:

- Spelling mistakes. Common spelling errors, e.g., 'enviroment' or 'freind' were included in the index words, but this approach could not capture every single spelling mistake in every single response were made.
- Using a wider range of language (and rarer words) to talk about certain topics. For example, when talking about jobs and careers, children aged 6-8 often gave a specific career that they aspired to, such as 'footballer', 'vet', 'teacher', 'doctor' etc. But these individual career types were not mentioned frequently enough to appear in the top 1,000 most frequently used words.

Girls and 16-17 year olds were more likely to have their response categorised. This is due to:

- Girls overall were more likely to respond to The Big Ask, were more likely to give any free text response (55% of free text responses were from girls, compared to 40% of boys) and were more likely to give a longer or more detailed response (22% of girls responses were categorised to 3 or more topics compared to 14% of boys).
- Young people aged 16-17 were likely to give longer responses (30% mentioned 3 or more topics) and were less likely to make spelling mistakes.

Table 2 - Percentage of children who responded to The Big Ask, provided a free-text response and whose response was categorised, by demographic characteristic

Characteristic	Percentage of children who responded to The	Percentage of children who provided a free-text	Percentage of children whose free-text response was

	Big Ask	response	categorised
Age			
6-8	19%	12%	10%
9-11	34%	34%	34%
12-15	38%	43%	45%
16-17	7%	9%	9%
Gender			
Male	44%	40%	39%
Female	51%	55%	56%
Self-identified gender	2%	3%	3%
Ethnicity			
Asian	10%	11%	11%
Black	4%	4%	4%
Mixed	5%	4%	4%
Other	1%	1%	1%
White	70%	70%	70%
Ethnicity not given	6%	4%	5%
IDACI quintile (1 = most deprived, 5 = least deprived)			
1	15%	15%	15%
2	16%	16%	17%
3	17%	18%	19%
4	19%	20%	20%
5	21%	23%	23%

Table 3 - Percentage of sub-group responses categorised and uncategorised

Characteristic	Percentage of responses categorised	Percentage of responses uncategorised
Age		
6-8	68%	32%
9-11	85%	15%

12-15	89%	11%
16-17	91%	9%
Age not given	81%	19%
Gender		
Male	82%	18%
Female	87%	13%
Self-identified gender	87%	13%
Gender not given	84%	16%
Ethnicity		
Asian	86%	14%
Black	85%	15%
Mixed	86%	14%
Other	83%	17%
White	85%	15%
Ethnicity not given	80%	20%
Total	85%	15%

Table 4 - Percentage of sub-group by number of topics response was categorised to

Characteristic	Percentage categorised to number of topics					
	1	2	3	4	5	6+
Age						
6-8	52%	13%	3%	1%	0	0
9-11	42%	25%	11%	4%	1%	1%
12-15	40%	26%	13%	6%	2%	1%
16-17	36%	25%	15%	8%	4%	3%
Gender						
Male	46%	23%	9%	3%	1%	1%
Female	39%	25%	13%	6%	2%	1%
Self-identified gender	39%	22%	13%	7%	3%	2%
Ethnicity						

Asian	40%	25%	12%	5%	2%	1%
Black	42%	25%	11%	5%	2%	1%
Mixed	41%	25%	12%	5%	2%	1%
Other	41%	24%	12%	4%	1%	1%
White	42%	24%	11%	5%	2%	1%
Total	42%	24%	11%	5%	2%	1%

Consideration of the uncategorised responses

The most common reason for responses not being categorised to a topic are:

1. The list of index words is incomplete. There is a diminishing return on the utility of index words where the first 1,000 most frequently mentioned words cover 91% of all words used, the next 1,000 most frequently mentioned only cover 4% of the words used, and so on. For example, for the career aspirations of the 6-8 year olds we included the most commonly mentioned careers, 'Footballer', 'Vet', 'Doctor', 'Artist', 'Player', 'Police officer', 'Gamer', 'Singer', 'Police man', 'Youtuber', 'Horse rider', 'Dancer', and 'Nurse'. However, we did not include other careers mentioned far less frequently such as 'Fireman' which was mentioned 30 times or 'Builder' which was mentioned 13 times.
2. Children entering responses which cannot be categorised, such as only entering punctuation, entering a web address or a joke response
3. Themes which were identified but did not map to a topic conceptually and were too small to be a topic on their own. For example, children 6-8 frequently mentioned getting pets such as dogs, cats or fish, e.g., '*2 dogs and 5. pet fish please*' Boy, 6.

Further work is required to examine the uncategorised responses and identify whether any themes have been excluded which could inform the analysis and understanding following the completion of this initial report.

Depth-analysis

While the first stage gave a good indication of the kinds of subjects and topics that children were referring to, it could not explain what children felt about those topics. The aim of the depth-analysis was to get to closer to identifying the views and sentiments that children expressed within each topic. For this stage we employed a more traditional 'qualitative' with the researchers reading large samples of the responses and grouping the responses into sub-themes.

As the data were too large to read every response assigned to a topic, we used a number of techniques to assess the entire topic and sub-themes. Firstly, word clouds and bigram maps were produced which give an indication of the potential sub-themes present within the topic. For example, when looking at education, the largest topic, the bigram map revealed the use of the term 'education system' and the words used alongside school such as 'private' and 'grammar'. The bigram map highlighted language which

researchers could then look for in the responses, to understand how children used these words and the wider context.

For each topic, researchers took one random sample of 1,000 responses and one random sample of 500 responses from girls and boys separately as the starting point for responses to read. Then additional samples of around 200 responses were drawn on specific words to allow further exploration of a sub-theme. For example, within the education topic, additional samples were drawn for the sub-themes:

- 'System'
- 'Grammar'
- 'Memory'
- 'Lucky'
- 'Curriculum'
- 'Teacher'
- 'Distract', 'naughty' or 'messing'
- 'Autism', 'adhd', 'dyslexia' or 'aspergers'
- 'Trade', 'vocation' or 'apprenticeship'

These words were identified from the initial sample of 1,000 as indicative of a particular sub-theme. This approach allowed review of a large number of responses in a very targeted way in a short time frame. However, because this was based on an initial sample of 1,000 responses (out of 73,000 responses categorised to education), it is not impossible that other sub-themes may have been missed (if they did not appear in the initial word cloud, bigram or 1,000 response sample). This is inevitable with data of this scale.

To maximise the amount of insight within the time available, the approach was designed to deliver the maximum amount of detail utilising the majority of the qualitative data. Out of scope for this report was detailed mixed-methods analysis looking at how children responded to both the qualitative and quantitative questions, in-depth analysis dedicated to specific vulnerable subgroups or other cohorts of children, and exploring alternative methods for processing the text, such as 'topic modelling'.

Next steps

- Conduct further depth-analysis of the issues raised by children from vulnerable sub-groups such as young carers, looked after children, and children who self-identified their gender.
- Conduct mixed methods analysis utilising the qualitative and quantitative data to understand the relationships between, for example, children's current worries and the barriers they perceive for achieving in the future.
- Depth-analysis of the 'uncategorised' responses to draw out any additional themes or topics.