Report on the Analysis of Qualitative Data from StudentSurvey.ie COVID-19 questions 2021

Authors: Dr. Stephen Erskine

David Harmon

Insight Statistical Consulting

60 Merrion Square

Dublin 2, Ireland

www.insightsc.ie

Date: April 2021

Due to the volume of responses to StudentSurvey.ie, steps have to be taken to handle the corpus of data in a suitable manner as hand-coding of the material is too unwieldly to be practical. As such, this report has combined researcher-led coding of the material with computer-assisted content analysis methods to draw out the key themes within student responses.

The computational analysis of the responses to the two open-text questions was carried out using the open-source statistical software R in conjunction with Insight Statistical Consulting to ensure that the coding frame developed was a good fit for the whole dataset. Supporting documentation relating to this analysis is available on www.studentsurvey.ie.

The preliminary steps taken were identical for both open-text questions in that first of all, a familiarity with the content of the responses was obtained thorough reading through a large proportion of the comments while conducting a spell-check. This is very time-intensive but invaluable in understanding how students approached answering each question at hand, and from a practical perspective ensured that no words were lost to the analysis through common spelling errors.

Due to time constraints, the minority of students who provided their answers in Irish were filtered out of the computational analysis. Other strings of text were filtered out at the preliminary stage if they provided no substantive content, for example, if they were random or meaningless character strings like ‘qwerty’/’eeeeee’, or if they were under four characters, because at this length it is impossible to ascertain what students were trying to say. A similar problem was evident with students who replied either ‘yes’ or ‘no’.

Once the preliminary stages of cleaning the corpus of data had been completed, the data was ready to be entered into the statistical software R. The first step upon entry into the software was to remove punctuation and symbols that cluttered the corpus. The second step was to segment the corpus of data into individual tokens, usually words separated by white space. However, through immersion in the material from the outset, we noted some words which tended to be associated with one another. To avoid these being lost to the analysis, underscores were inserted between these words, so that these multiword phrases would be compounded and seen by the software as being one unit. Some of these, for example 'continuous assessment' are visible in the results below. The final step was to remove common 'stop words' that are of little intrinsic value to the analysis (such as "the", "a", "an", and "in") whose presence would suppress words and phrases that are important to the analysis.

The first open-text question asked of students was:

> **Q1: What are the positive elements of the online/ blended learning experience you want to keep when on-campus studies resume?**

Of the 43,791 students how completed the survey, 35,496 students provided a response to this question, which when cleaned provided 34,655 comments suitable for analysis. Some summary statistics of the number of characters used in students' responses to Q1 are provided in Table 1.

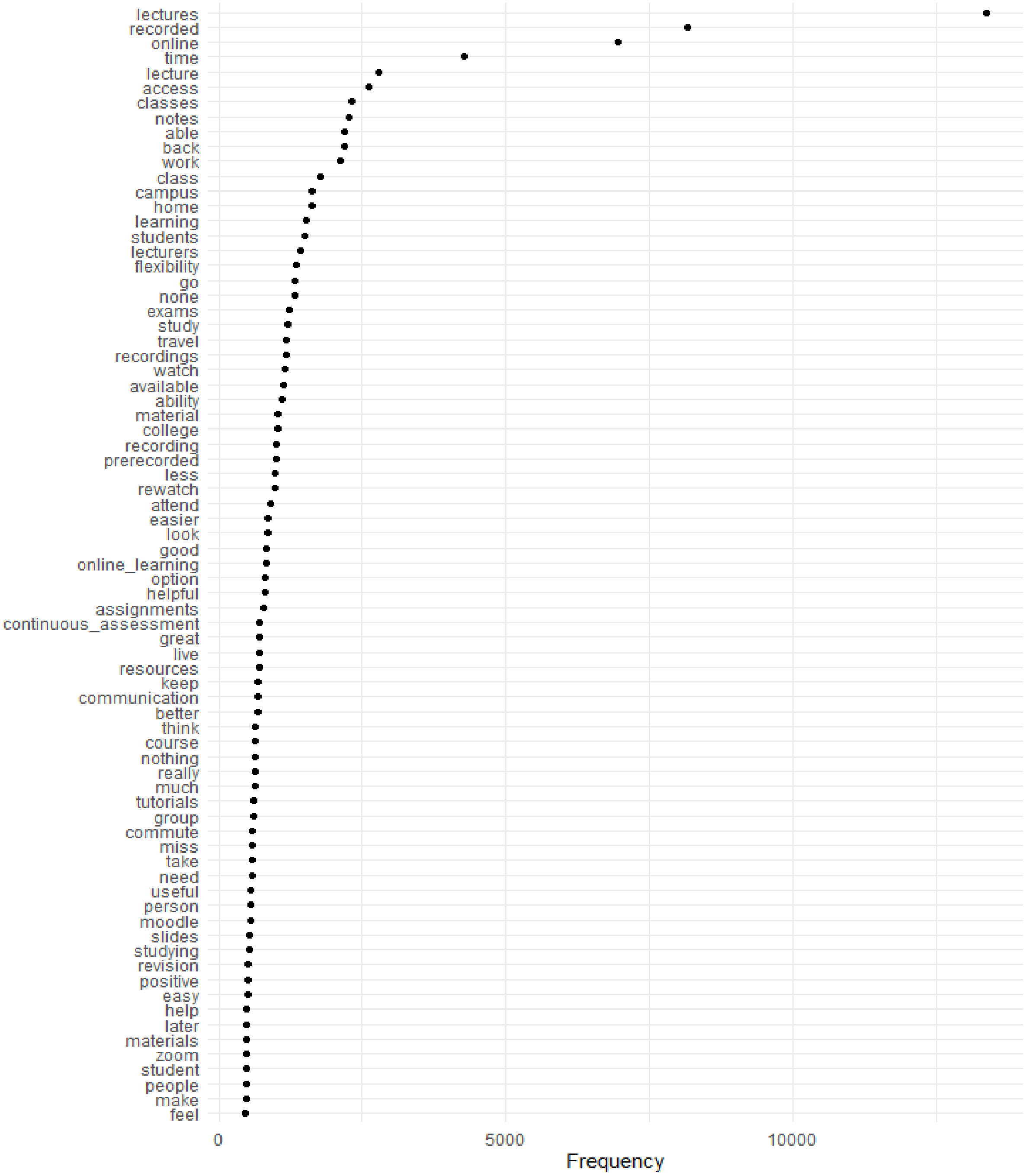
Table 1: Summary statistics of the number of characters in responses to Q1

Mean	Median	Standard deviation	Interquartile Range	Minimum	Maximum	Total N
65	46	67	59	4	1382	34,655

Figure 1 plots the relative frequency of the top 75 words. From this we can see that "lectures" is the most used word with over 12,500 instances in the corpus. After this there is a large gap between the most used and second most used with "recorded" being used over 7,500 times. In third place, close to "recorded" is "online" with just under 7,500 instances.

There are 6,200 unique words used in the clean Q1 corpus, and these words are used close to 200,000 times in total. However, the top 75 words account for over fifty percent of the words used in the Q1 corpus.

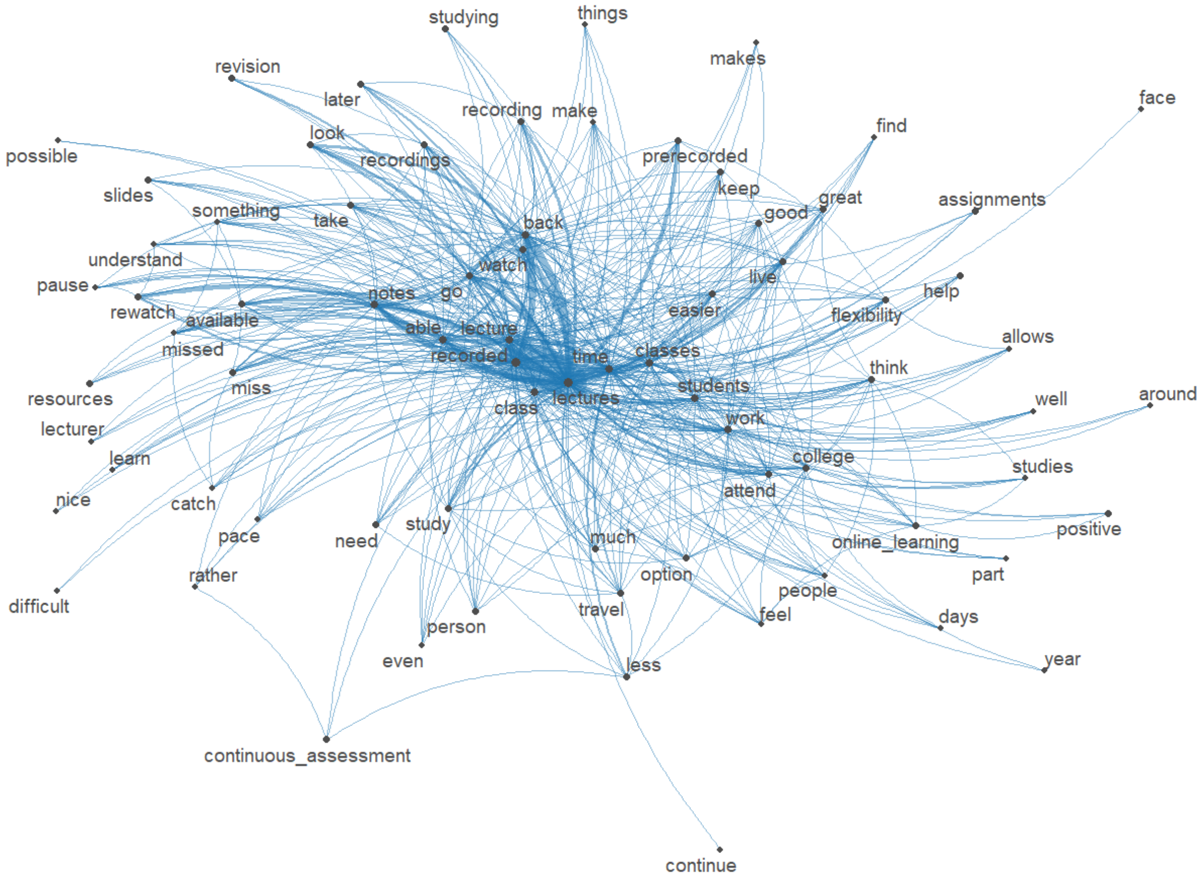
Figure 1: Relative frequency of the top 75 most frequently used words (Q1)



The analysis of individual words is illuminating but is also divorces words from the other words around them in sentences. To bring this back into the analysis, the next step is then to identify which words are most closely associated with one another. Within the statistical software this was done by creating a feature co-occurrence matrix which records the number of co-occurrences of tokens.

This feature co-occurrence matrix can then be visualised in a semantic network to illustrate which words are most associated with one another. The width of the bars linking words indicates the strength of the relationship between the words. Figure 2 presents a semantic network for Q1 and as we would expect the three most frequently used words in the corpus thus far, “lectures”, “recorded” and “online” form the central axis from which all the other words branch.

Figure 2: Semantic network of feature co-occurrence matrix



An alternative way of visualising the association between words is to use latent semantic scaling, which is a flexible and cost-efficient semi-supervised document scaling technique. This method has a couple of distinct advantages which the charts below demonstrate further. First of all, the scaling allows us to plot the frequency of words associated with keywords of interest. Secondly, we can combine this with sentiment analysis through applying a sentiment dictionary to the corpus. This sentiment dictionary identifies words with specific positive connotations which then allows us to identify the most used positive associated with key words of interest.

Figure 3 presents the words most commonly associated with 'lectures' and in the first chart highlights the most commonly used words in the Q1 corpus if associated with the keyword 'lectures', the second chart highlights words related to lectures in the corpus with positive connotations. As can be seen from the charts, the keyword 'lectures' tends to be used with 'recorded', 'access', 'notes', 'able' and 'back'. In addition, there are a lot of positively associated words used when describing lectures.

Figure 3: Latent semantic scaling of words associated with 'Lectures'

Most frequently used words	Words associated with positive sentiments

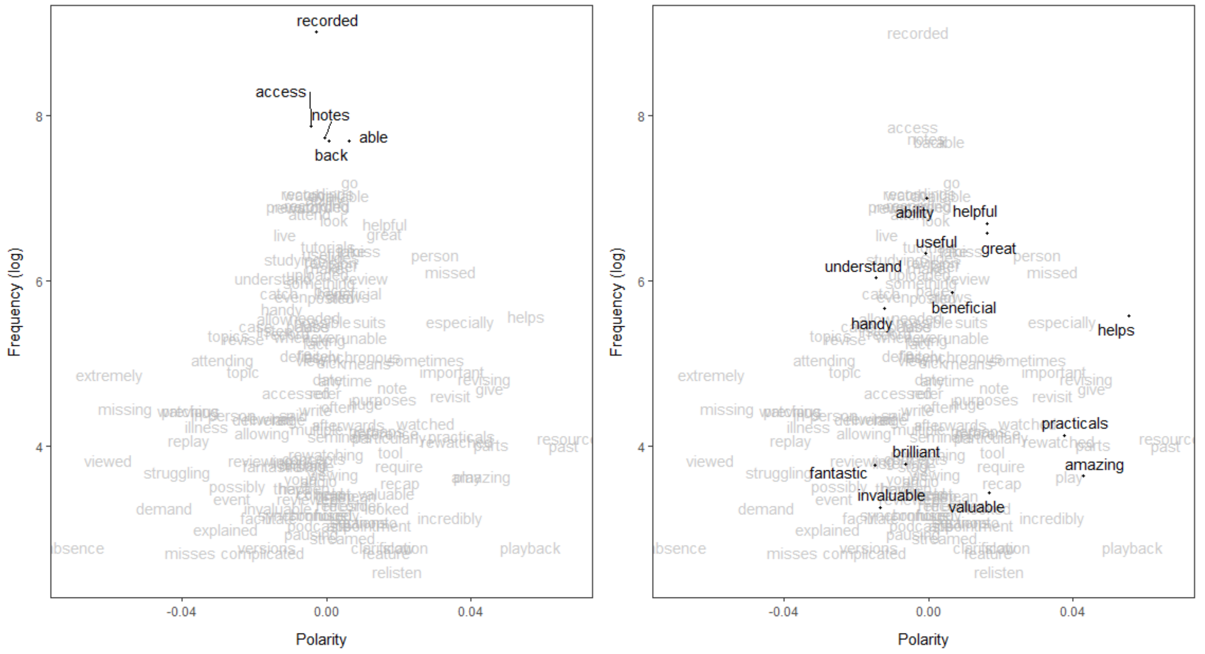


Figure 4 presents the words most commonly associated with 'recorded'. The first chart highlights that the word 'recorded' tends to be used along with the words 'lectures', 'classes' and 'back'. The second chart highlights words related to 'recorded' in the corpus with positive connotations.

Figure 4: Latent semantic scaling of words associated with ‘Recorded’

Most frequently used words	Words associated with positive sentiments

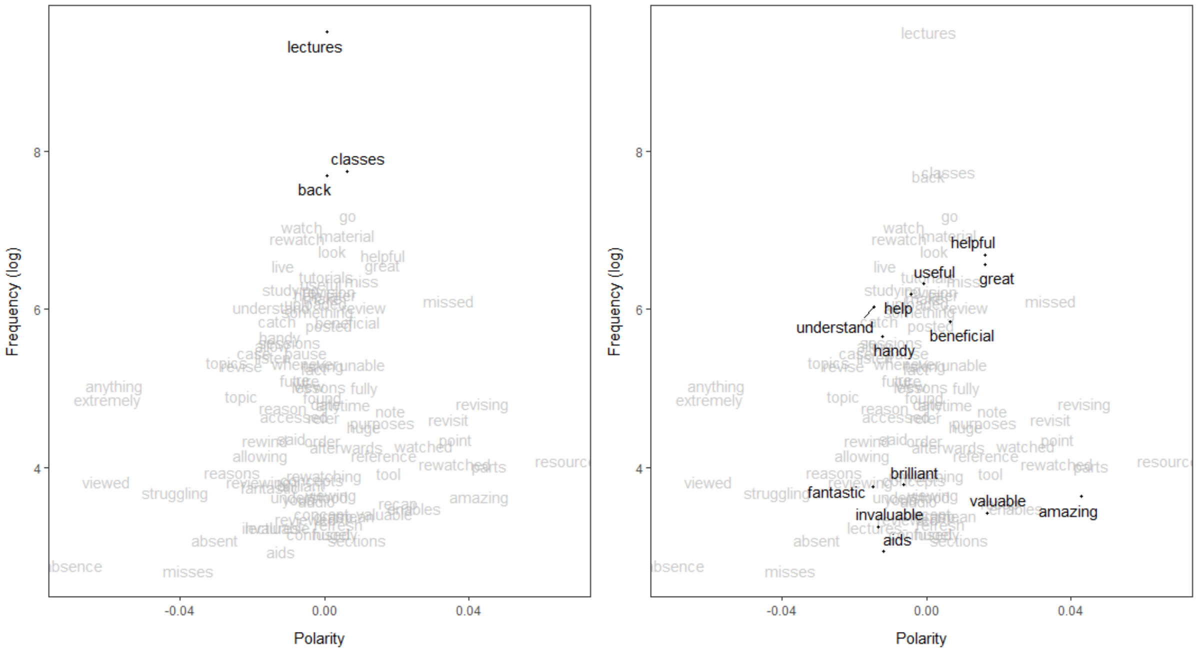
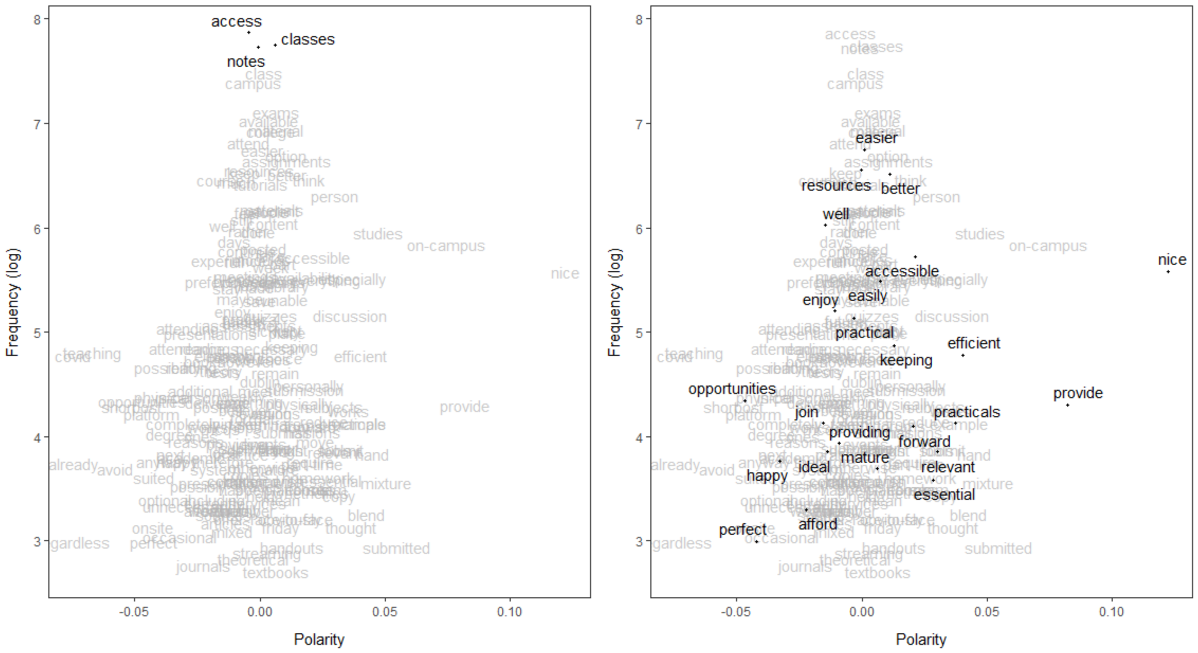


Figure 5 presents the words most commonly associated with ‘online’. The first chart highlights that the word ‘online’ tends to be used along with the words ‘access’, ‘classes’ and ‘notes’. The second chart highlights words related to ‘recorded’ in the corpus with positive connotations.

Figure 5: Latent semantic scaling of words associated with ‘Online’

Most frequently used words	Words associated with positive sentiments



These charts give us some indication of the content of students' comments when using certain popular words. Another way of examining this is to move beyond individual words in the corpus and tokenise consecutive sequences of words within each comment provided by students, and then examine these sequences. These sequences are called n-grams and the ones with greatest use here, are bigrams which are sequences of two words. Thus by seeing how often word X if followed by word Y, we can then begin to build a model of the relationships between the words, and obtain a clearer picture of the answers students are providing for Q1.

Figure 6: Relative frequency of most frequently occurring bigrams

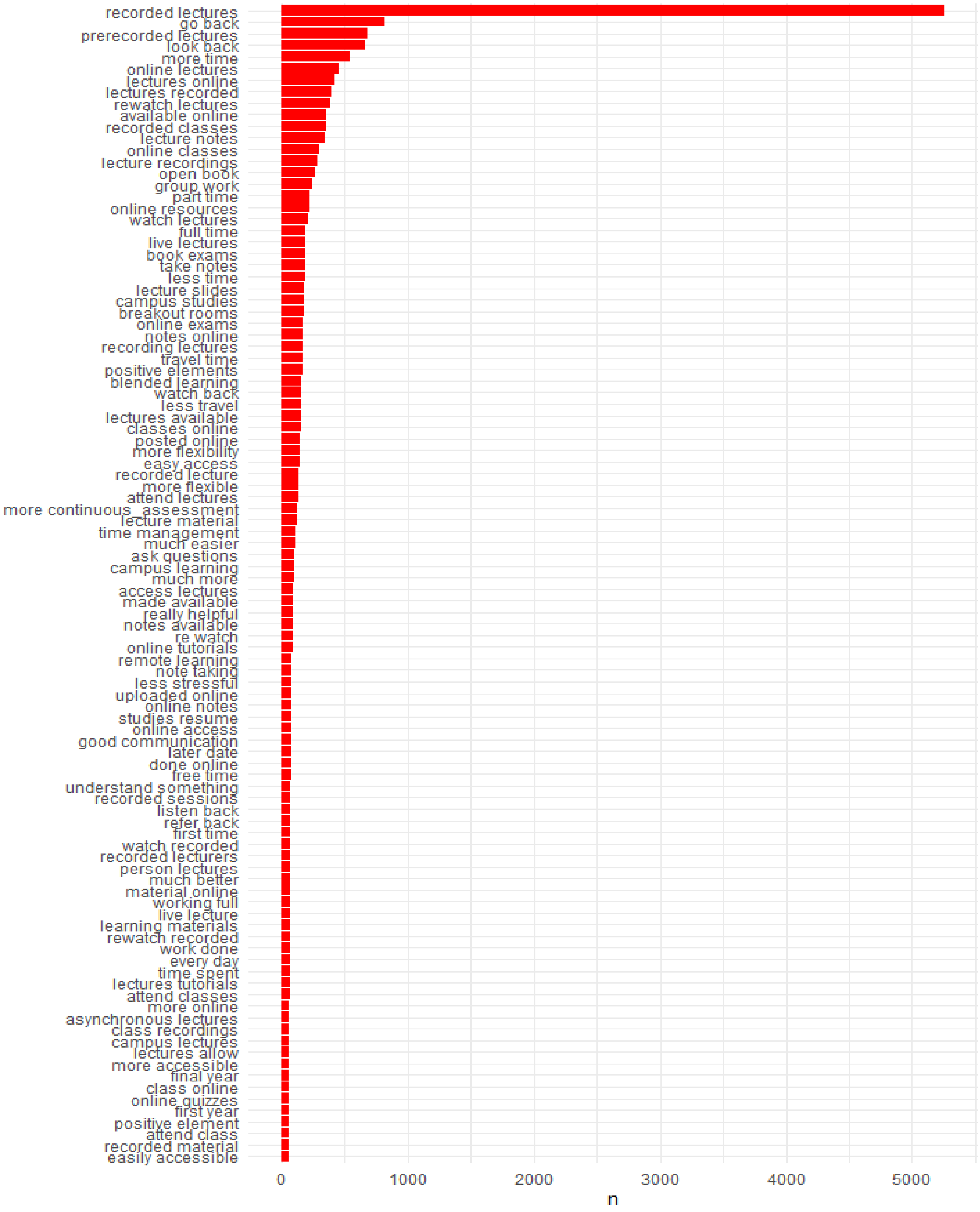
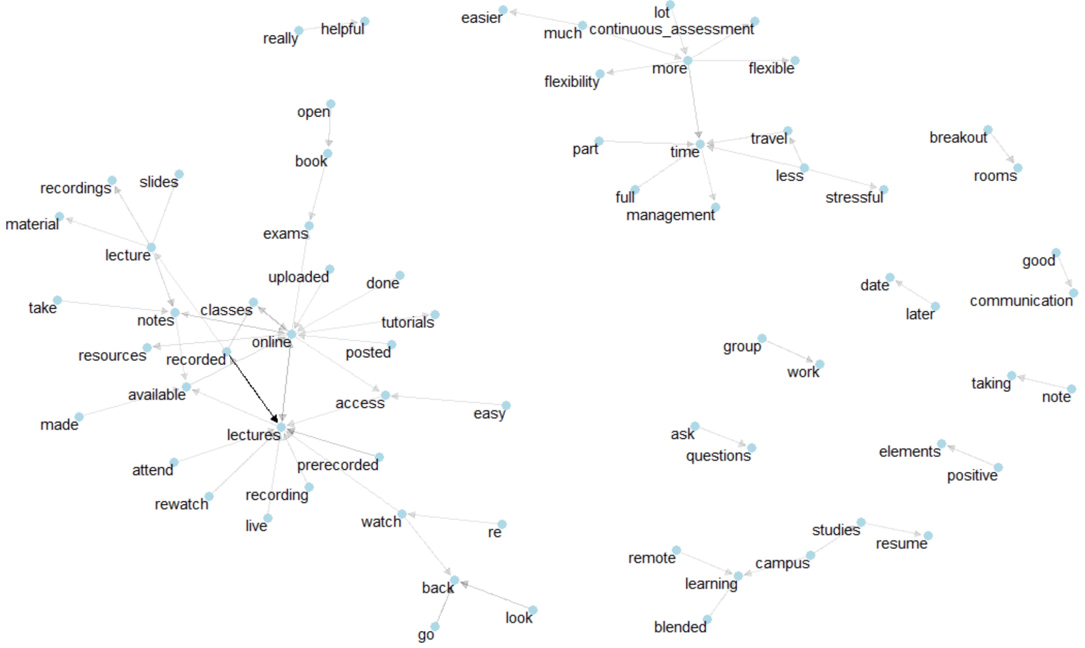


Figure 6 presents the most frequently occurring bigrams in the corpus, and quickly illustrates how often recorded lectures are mentioned within the corpus. Furthermore, almost all of the most frequently occurring bigrams appear to be either mentioning online recorded lectures or describing their benefits, they allow students to go back and rewatch them and so on.

The interrelationship between bigrams can be further illustrated through how they relate to one another rather than being viewed in isolation. Figure 7 presents bigrams where there are at least 250 cases and visualises the relationships among words simultaneously, rather than just a selected word. As such, the chart is a visualisation of a Markov chain, which is a common model in text processing, where the choice of a word only depends on its previous word. The arrows linking words show the direction of association. For example, the word 'really' typically precedes 'helpful'. This chart shows two large networks of bigrams, the smaller on the top right-hand side covering the flexibility of continuous assessment and how less travelling affects students time management and stress levels. The larger network covers online recorded classes and the positive aspects of this and other online resources.

Figure 7: Markov chain bigram network (Q1)



The analysis of Q1 so far has shown that at the aggregate level, students' responses to the question asked of them have predominantly focussed on one aspect of the online learning experience. This appears to be a commonality across all student groups. However, it is also worth examining where groups differ in perspective. One way of doing this is to examine how the words chosen by respondents demonstrate the different priorities students may have at each stage within third level education (i.e. first year undergraduate, final year undergraduate, or taught postgraduate).

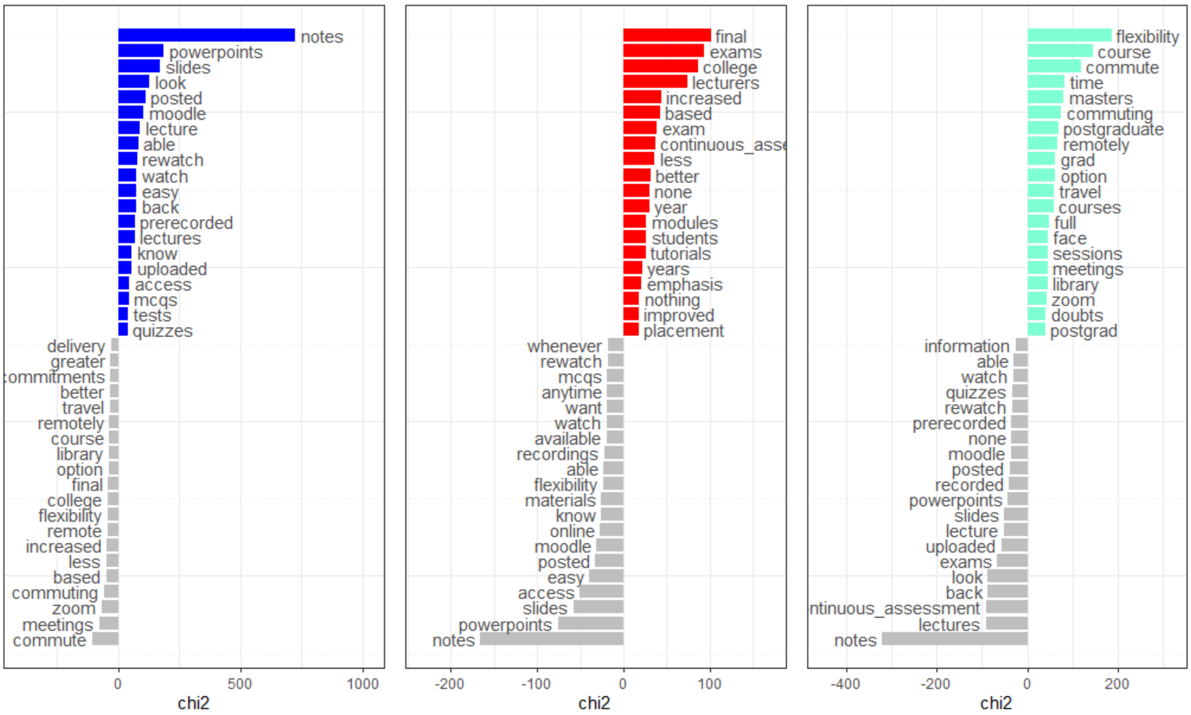
To identify significant differences across groups contained within a corpus a statistical measure called *keyness* was used. This measure uses the relative frequency of words across two parts of a corpus to see if there are differential associations of keywords between a target and a reference group.

Figure 8 examines this by comparing each individual group against a reference group (i.e. the other two groups) to see the relative frequencies of keywords. The bars in grey are terms frequently used by the reference groups and those in colour are the target group for that chart.

From the first chart on the left, we can see that first year undergraduates mention notes, PowerPoint, and slides much more than final year or taught postgraduates. For final year undergraduates we can see that final and exams or exam along with continuous assessment are mentioned a lot. Taught postgraduates mention the flexibility of working remotely along with numerous mentions of travel, commuting, or their commute.

Figure 8: Relative frequency analysis (keyness) by student status

First year undergraduates	**Final year undergraduates**	**Taught Postgraduates**



From the word choices across groups, we get some indications of how groups differ in their priorities when answering the question, in that first-year undergraduates talk about the accessibility of notes and lecture slides when summarising the positive elements of an online learning experience whereas final year undergraduates are more likely to make reference to assessment be it final exams or continuous assessment, and taught postgraduates are likely to mention the flexibility of online learning and how it reduces the time spent commuting, which frees up time for study.

In previous work conducted on earlier iterations of the open-text questions contained in the standard Irish Study of Student Engagement from 2016 to 2020, each year's corpus has largely split neatly into a number of themes. This does not appear to be the case for students answering the question at hand here. Within responses to the question "what are the positive elements of the online/blended learning experience you want to keep when on-campus studies resume?", one answer appears to dominate and that is students would like to have recorded lectures available to them. Over forty percent of respondents mentioned lectures at least once in their responses and typically noted that have recorded lectures available to them provided them with greater flexibility as they did not have to be there in person (so had to travel less) and allowed them to approach their studies in their own time and at their own pace. Other themes were mentioned but to a much lesser extent, for example, students also mentioned wanting to have more continuous assessment rather than final exams.

The second open-text question asked of students was:

> **Q2: In what way(s) could your higher education institution improve its support for you during the current circumstances?**

In contrast to Q1, which is worded in a manner to evoke a precise response, such as an item, or institutional provision that students would like to see when they return to in-person and on-campus teaching, Q2 poses a more hypothetical query, which asks about a change that could be made by students' HEIs to improve their circumstances. The analysis of this question follows a similar format to that used for Q1 but with a few significant deviations to capture this fact that standard responses from students provide **items** that students think should be altered to improve the support provided by an HEI and some **measure of direction and degree** (more/less/better/fewer and so on).

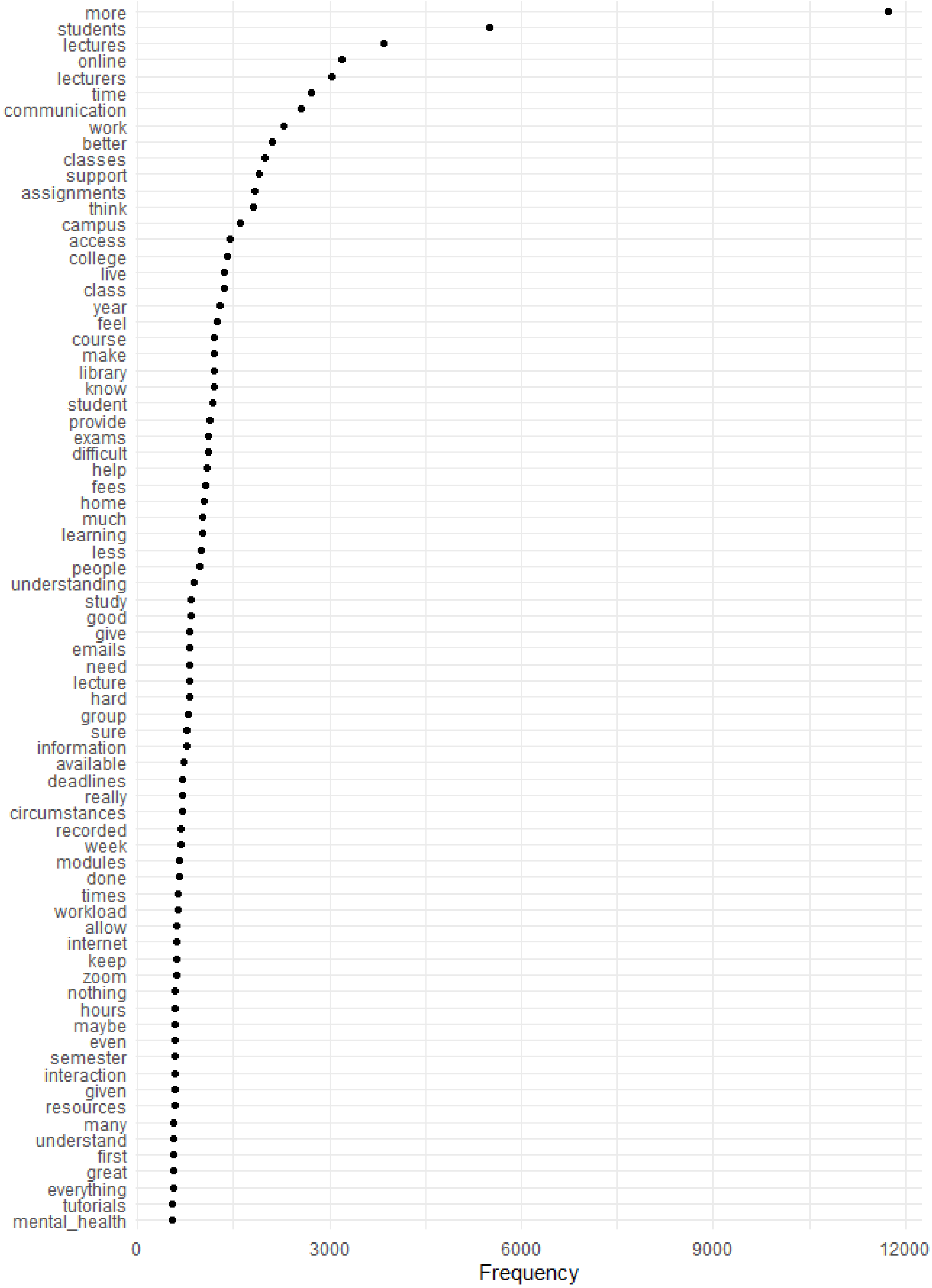
Of the 43,791 students how completed the survey, 32,307 students provided a response to this question, which when cleaned provided 30,955 comments suitable for analysis. Some summary statistics of the number of characters used in students' responses to Q1 are provided in Table 2.

Table **Error! No text of specified style in document.**: Summary statistics of the number of characters in responses to Q2

Mean	Median	Standard deviation	Interquartile Range	Minimum	Maximum	Total N
93	61	110	81	4	2000	30,955

Figure 9 plots the relative frequency of the top 75 words. From this we can see that "more" is the most used word with almost 12,000 instances in the Q2 corpus. After this there is a large gap between the most used and second most used with "students" being used almost 6,000 times. In third place, close to "students" is "lectures" around 4,000 instances. There are over 9,100 unique words used in the clean Q2 corpus, and these words are used close to 253,000 times in total. In contrast to Q1, the broader selection of words used in students' responses means that the top 75 words accounts for only thirty-eight percent of the words used in the Q2 corpus.

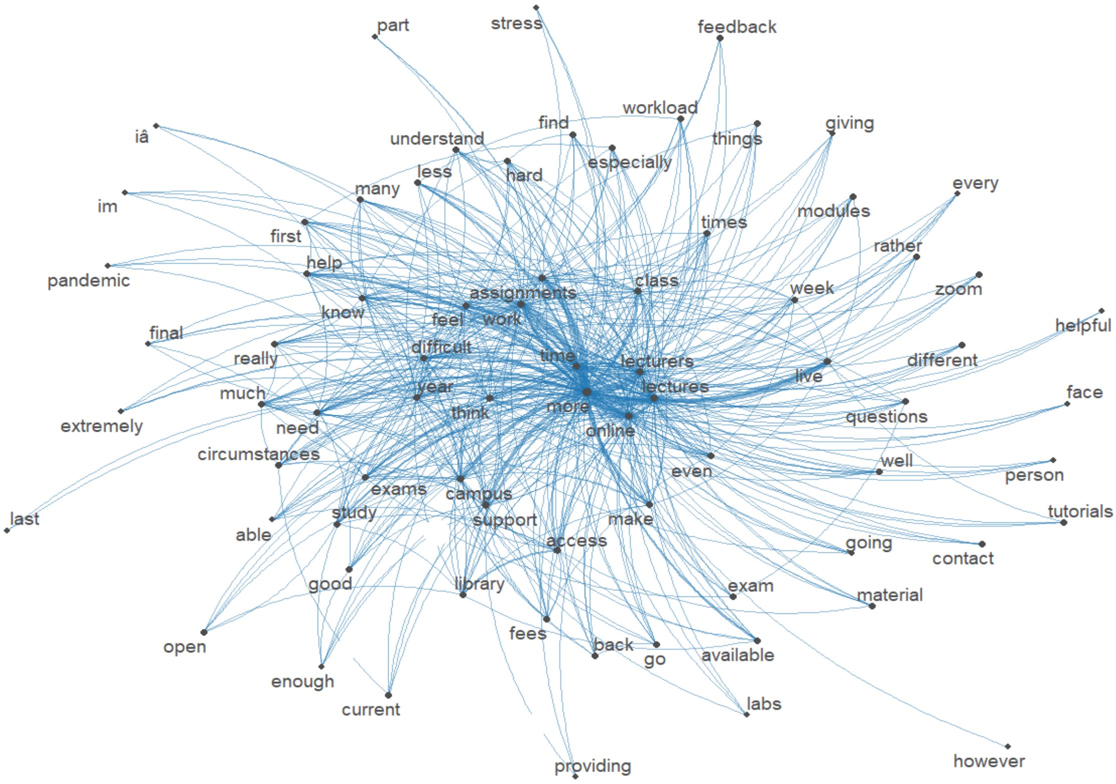
Figure 9: Relative frequency of the top 75 most frequently used words (Q2)



As noted above, the analysis of individual words is illuminating but is also divorces words from the other words around them in sentences. To bring this back into the analysis, the next step is then to identify which words are most associated with one another. Within the statistical software this was done by creating a feature co-occurrence matrix which records the number of co-occurrences of tokens. This feature co-occurrence matrix can then be visualised in a semantic network to illustrate which words are most associated with one another. The width of the bars linking words indicates the strength of the relationship between the words. Figure 10 presents a semantic network of the Q2 corpus and the frequency that 'more' is found in the corpus means that it is linked with a large number of other keywords and forms the central hub from which all other words branch out from.

This chart broadly illustrates that a large number of students answered Q2 by writing "more" and following this up with a mention of some item which if provided by their HEI would improve student support. All of which, supported our hypothesis that to get a handle on how students answered Q2 it was necessary to examine both the items that were being mentioned but also the degree and direction associated with them.

Figure 10: Semantic network of feature co-occurrence matrix



Because students tended to follow a very similar format in answering Q2 through mentioning some form of direction and degree along with an item, it is possible to move beyond individual words in the corpus and tokenise consecutive sequences of words within each comment provided by students, and then examine these sequences as they tend to capture the format with which students provide answers to Q2 very well.

Figure 11 presents the most frequently occurring bigrams in the Q2 corpus. Some bigrams are completely descriptive such as ‘first year, ‘work load and ‘group work. However, what is striking is how often modifiers are used within the corpus. Throughout this list ‘more’ is often the first or second word in the bigram. Other modifiers used include ‘smaller’, ‘better’, and ‘less’.

Figure 11: Relative frequency of most frequently occurring bigrams (Q2)

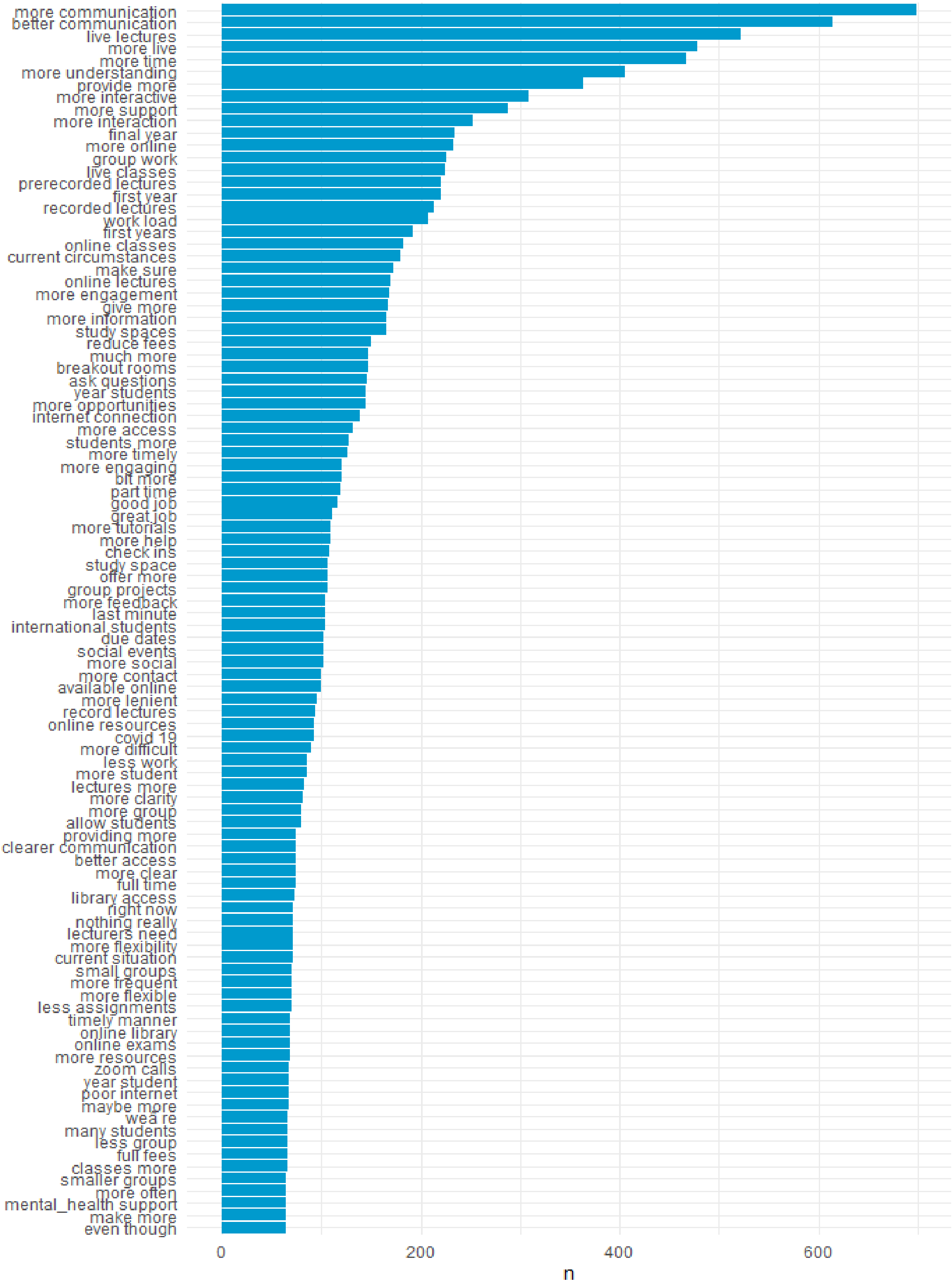
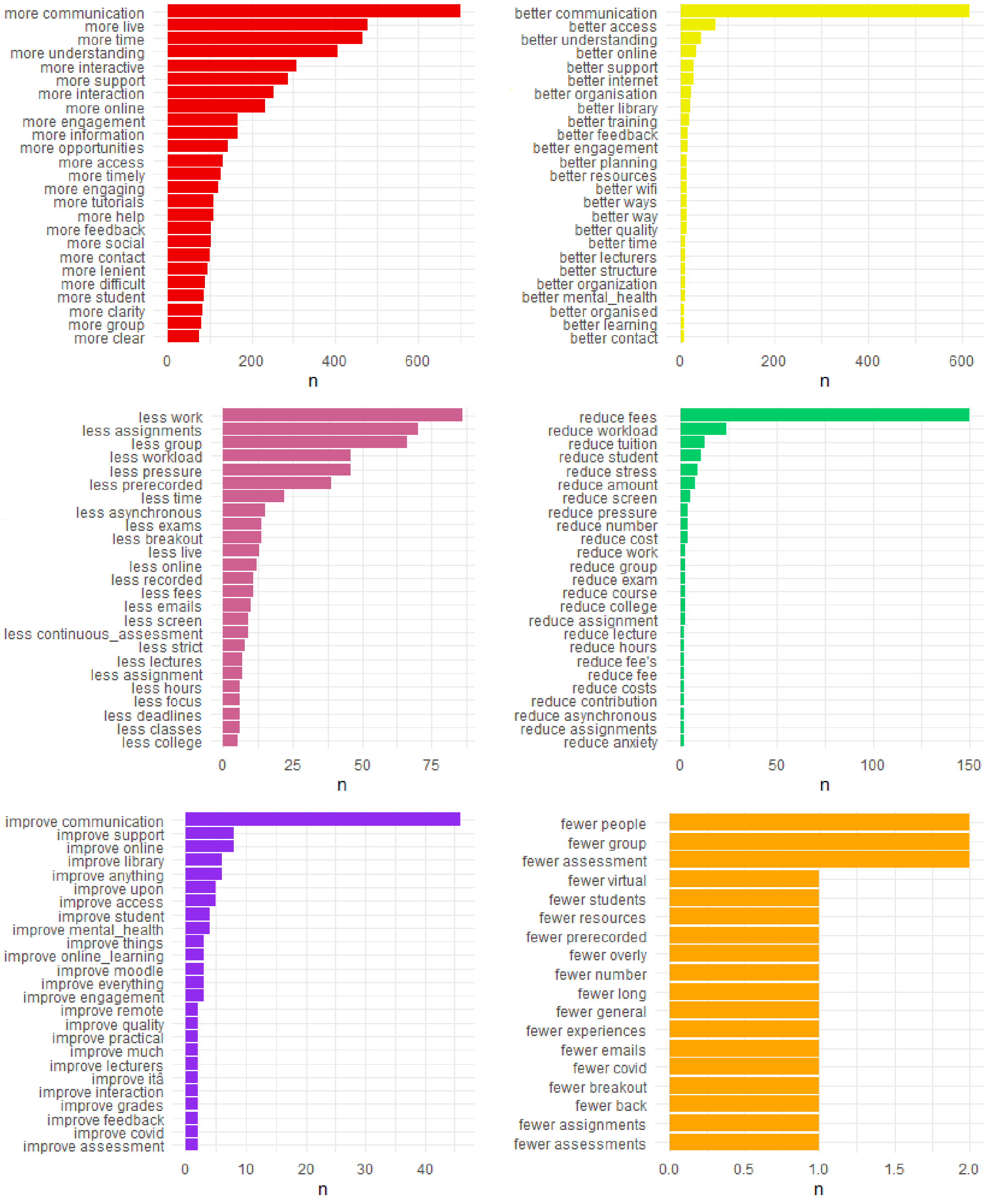


Figure 12 advances this further and presents the most frequently occurring bigrams in the corpus with the most frequently used modifiers as the first word in the set. From this, one can see how often more, better, and improved communication is mentioned by students. Reducing fees and providing students with less work also are frequently used bigrams.

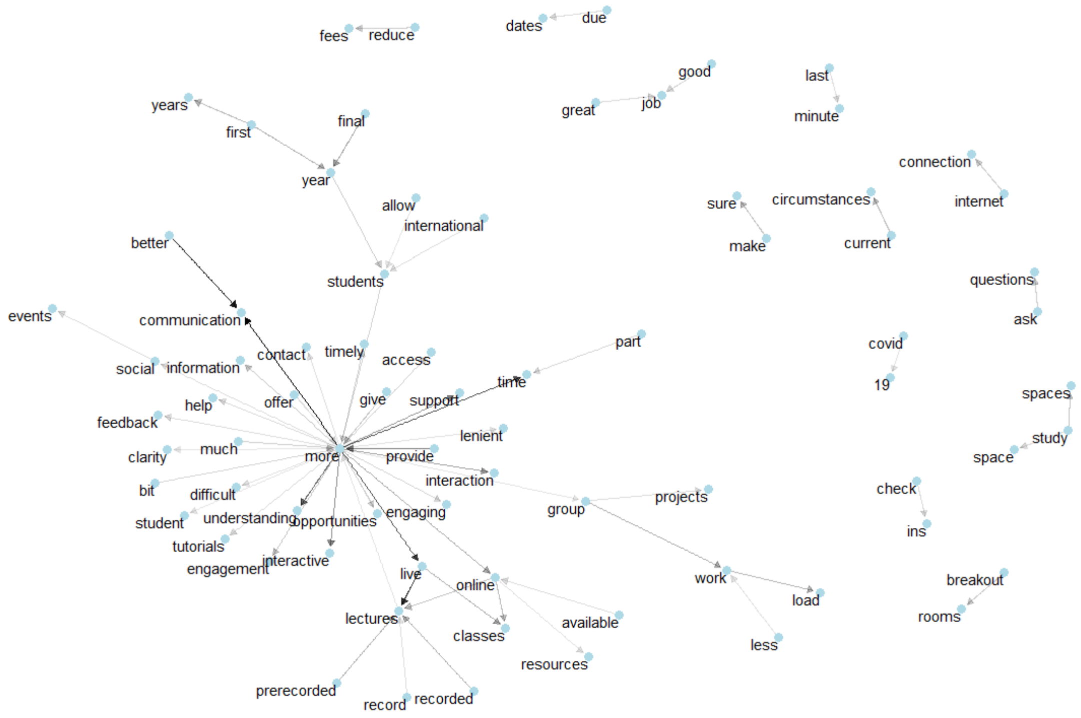
Figure 12: Relative frequency of bigrams containing ‘more/better/less/reduce/improve/fewer’ as first word



However, one has to be wary of painting a distorted picture. Note that reducing fees is mentioned around 150 times whereas more/better/improve[d] communication is mentioned around 1200 times. Other themes emerging from the bigram analysis include HEIs being more understanding of students and engaging with students more, all of which can be seen as a corollary of better communication between students and their HEIs.

The interrelationship between bigrams can be further illustrated through how they relate to one another rather than being viewed in isolation. Figure 13 presents bigrams where there are at least 250 cases and visualise the relationships among words simultaneously, rather than just a selected word. The arrows linking words show the direction of association. The chart shows one large network of bigrams with 'more' being the central node, along with a number of individual bigrams.

Figure 13: Markov chain bigram network (Q2)

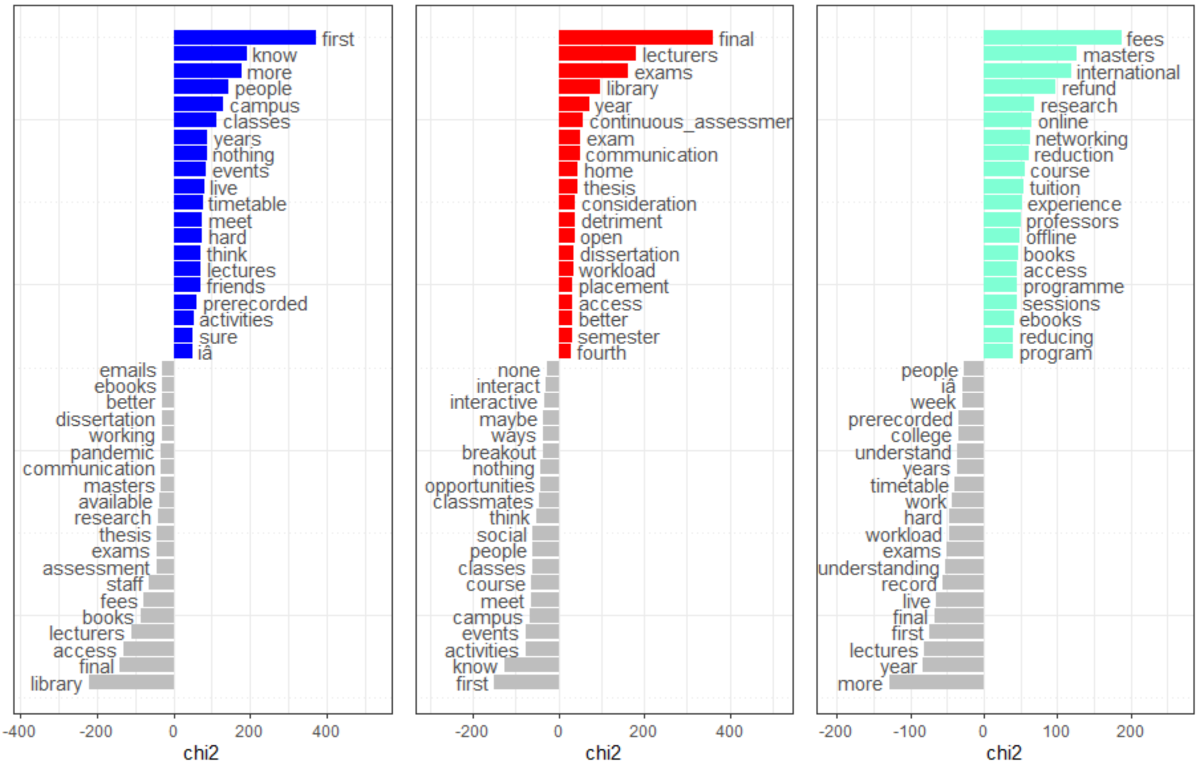


The analysis of Q2, like that of Q1, has shown that at the aggregate level, students' responses to the question have focussed largely on one aspect where support can be improved. This again appears to be a commonality across all student groups. However, it is also worth examining where groups differ in perspective. This has been done by measuring the *keyness* of words used in students' responses across groups to see if there are differential associations of keywords between a target and a reference group.

Figure 14 examines this by comparing each individual group against a reference group (i.e. the other two groups) to see the relative frequencies of keywords. The bars in grey are terms frequently used by the reference groups and those in colour are the target group for that chart.

Figure 14: Relative frequency analysis (keyness) by student status

First year undergraduates	Final year undergraduates	Taught Postgraduates



Much like what was seen for Q1, the word choices across groups give us some indications of how groups differ in their priorities when answering the question. From the first chart on the left, we can see that first year undergraduates tend to mention things that have been missing from their student experience because of the shift to online learning such as ‘people’, ‘events’ and ‘activities’. For final year undergraduates we can see that much like that seen for Q1 at this point, these students are focussed on assessment with ‘exams’, ‘continuous assessment’, thesis’, ‘dissertation’ and so on being used more by final year undergraduates than the other two groups. In contrast, taught postgraduates mention a lot of words around tuition fees and how they would like them reduced.

As a final check for Q2, a sentiment dictionary was run over the corpus. This form of sentiment analysis differs from the sentiment analysis conducted for Q1, in that it examined whole responses provided by students rather than just individual words. The sentiment contained in a comment can be broadly classified as positive, neutral or negative, is then represented on a numeric scale, to better express the degree to which a body of text contains positive or negative sentiments. Higher scores for comments indicate higher levels of positive sentiment, and lower scores indicate higher levels of negative sentiment.

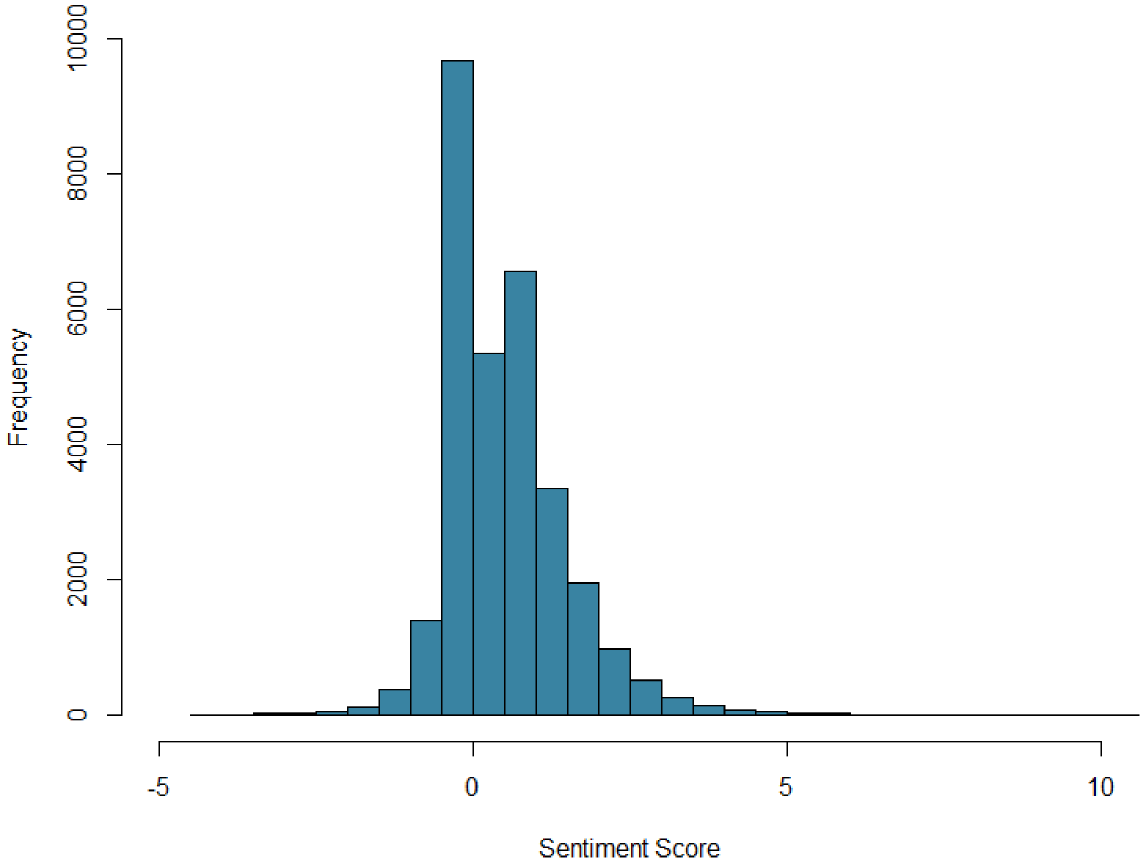
It is worth highlighting that this approach is not without its imperfections, but there is an inherent trade-off between human coding, time, and error. In general, from examining the scores associated with each sentence in the corpus, the sentiment dictionary does a good job of capturing the sentiment contained within comments even if it lacks the ability to parse nuance, irony and sarcasm. It is more than this methodology excels at evaluating the underlying sentiment when the wealth of material is too much for a human coder and would fall prey to biases in human coders. It would also take days for a human coder to read all of the text contained in responses to Q2, whereas the software evaluated the corpus in considerably less time.

A summary of the sentiment scores for the whole corpus is provided in Table 3, and Figure 15 provides a histogram of the distribution of sentiment scores. A random sample was also read to ensure that the sentiment analysis had performed an adequate job of summarising the underlying sentiment contained in student responses. The mean sentiment score for the Q2 corpus was 0.6 and the median was 0.5 and ranged from -4.4 to 11.8. Overall, the sentiment analysis and follow-up checks showed that some students used their responses to Q2 to vent their frustrations, and the sentiment dictionary did a good job of giving these students a very low score, but these were a very small proportion. On the whole, students were realistic in assessing their situation and thought their HEIs were doing a good job of moving to online teaching in the exceptional circumstances, which were then ascribed a marginally positive sentiment score.

Table 3: Summary statistics of the sentiment scores for Q2

Mean	Median	Standard deviation	Interquartile Range	Minimum	Maximum	Total N
0.6	0.5	0.9	1	-4.4	11.8	30,937

Figure 14: Histogram of the distribution of sentiment scores for Q2



In summary, much like that found for Q1, responses provided by students largely focussed on one area, communication. Students desire to have more and better communication from their HEI. Other themes were mentioned but to a much lesser extent, for example, students also mentioned wanting more understanding of the unique situation that students are living through, along with greater efforts through events and activities to mitigate the isolating effects that online teaching can have on students. This was noted especially by first year undergraduates. In contrast, taught postgraduates highlighted the cost of fees and wanted to see these reduced because of their inability to access the full range of resources available to them in normal times.