Vol 7 | Issue 1 (2019) | pp. 57-64

JOURNAL OF
Perspectives in Applied Academic Practice

www.jpaap.napier.ac.uk

On the Horizon: Making the Best Use of Free Text Data With Shareable Text Mining Analyses

Jill R D MacKay, University of Edinburgh

ABSTRACT

The current sector-wide Enhancement Theme of 'optimising the use of existing evidence' encourages the sector to identify what evidence exists, and to explore associated opportunities for best practice. Across the higher education sector, there is a prevalence of free text datasets which are generated through annual surveys and rarely explored across institutions, partly because of the privacy concerns that exist due to the nature of the data. In a recent project exploring secondary analyses of National Student Survey data, the University of Edinburgh also explored text mining approaches to offer fast and repeatable analyses of free text data that can be adopted by other institutions and researchers, without sharing sensitive data. This method has been trialed on institutional level data from the 2016 National Student Survey simultaneously with an in-depth open coding approach to the same data. This 'on the horizon' paper demonstrates the usefulness of the data mining approach, but also shows it must be accompanied by some qualitative examination of the data to understand the results in context. Alongside this paper is the shareable code for other groups to replicate this approach on their own datasets, to contribute to the optimisation of existing evidence use.

Keywords: surveys, open science, enhancement themes, text mining, quality assurance

Introduction

Education institutions throughout the world rely heavily on surveys to evaluate the student experience and engagement. In the United Kingdom, the National Student Survey was developed to collect and evaluate student feedback to monitor the quality of teaching throughout the higher education sector (Richardson, Slater, & Wilson, 2007). In Australia there is the Student Experience Survey (SES, (QILT, 2018)), and in the US and Canada there is the National Survey of Student Engagement, or NSSE (NSSE, 2017). These surveys represent longitudinal datasets with which to monitor and evaluate teaching quality. In addition, individual institutions will likely survey at various points such as the end of courses or to consult on proposed changes. The low response rate to many in-house institutional surveys has been attributed to 'survey fatigue' (Adams & Umbach, 2012) due to the number of surveys students are asked to complete within a year.

There are many criticisms of this approach. For example, the NSS often receives criticism for vague or poorly worded questions (Bennett & Kane, 2014; Blair, Orr, & Yorke, 2012; Cocksedge & Taylor, 2013) and the reliability of the quantitative data was questioned by Yorke (2009). Despite this, NSS data is often use in impactful evaluations, such as the Teaching Excellence Framework (Neary, 2016; Shattock, 2018)

At present, the Quality Assurance Agency Scotland has a sector-level activity on enhancement themes for higher education, including optimizing the use of existing evidence and responding to the student voice. Students are regularly surveyed in higher education contributing to survey fatigue (Adams & Umbach, 2012), and yet student satisfaction is still challenging to measure, requiring multiple attributes and data (Bell & Brooks, 2017). In my opinion, we can make better use of the survey data available. Surveys were previously identified by Baepler & Murdoch, (2010) as a possible area for targeting with data mining as the higher education sector grows more comfortable with academic analytics. The NSS as a dataset has been mined for comparisons across the quantitative results at an institutional and disciplinary level (Burgess, Senior, & Moores, 2018), and the free-text data has also been mined for patterns through exploring word frequencies to identify common themes (Buckley, 2012).

Datamining large open-source datasets like the NSS is becoming more feasible due to the open source software developments in data science. The standardisation of data processing through 'tidy data' principles (codified by Wickham, 2014, tidy data has each variable in a column, each observation in a row, and each type of observational unit in a table) an observation in each row has enabled easier replication of data workflows by providing a common grammar for data analysts to use. For example, Silge & Robinson, (2016) produced a free package for the statistical programming language R called 'tidytext', which enables access to rapid text mining on a standard desktop computer.

One strategy for making better use of pre-existing data is the use of shareable code (Nosek et al., 2015) to promote standardised approaches to free text survey data across institutions. While these types of analysis cannot replace an in-depth qualitative analysis, their shareability and open principles provide additional benefits that could be of great use to the quality enhancement of the sector.

In this Horizon paper I will demonstrate a workflow for interpreting National Student Survey data in this way, and compare with an open-coding analysis conducted on the same data to inform on the strengths and weaknesses of a text mining approach.

Example Analysis

Example Data

The 2016 NSS was sent to 3908 final year students at the University of Edinburgh between January and April 2016. There were 3080 participants, a response rate of 78.8%. The free-text responses were broken down into three questions 'Looking back on the experience, are there any particularly positive or negative aspects of your course you would like to highlight?' (Positive and Negative), and 'What one thing would do most to improve the quality of your student experience?' (University Specific Question). The institutional structure is organised by college, school, and programme. The free-text responses can identify responses to the level of programme where there are more than ten students enrolled, however this analysis was performed at the school and college level, and all reporting is done at the college level.

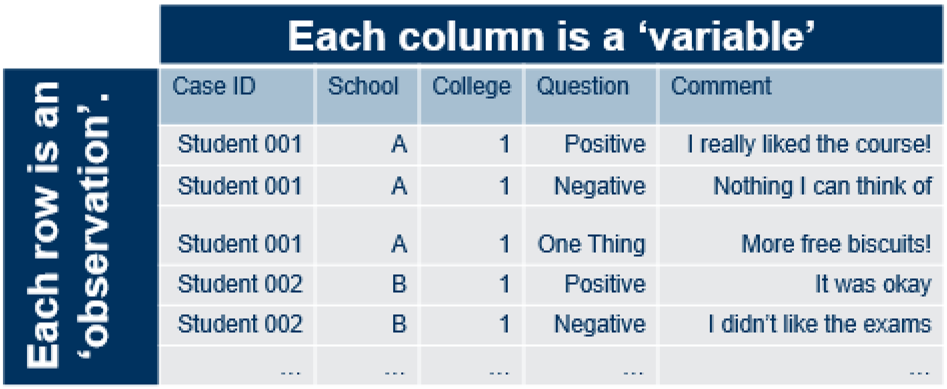
Basic Tools

In this example I use the R programming language, an open-source computer language designed for statistical computing, and a selection of freely available packages (user-created collections of code to expand functionality) to create the analysis. In this paper, I will not describe their use in detail, but instead demonstrate how a single code file can be used to replicate the analysis from our institution in any other institution.

Workflow

One of the most important principles of a tidy data approach is that the data processing steps become standardised, and so an analysis can be re-run on new data with relative ease. This is termed a 'workflow' (Grolemund & Wickham, 2017). In order to share our workflow, we must begin with data in the same format. Ideally, the processing steps to generate tidy data are part of the workflow, however due to the proprietary nature of NSS free text data, and the requirement to protect student anonymity (IPSOS Mori, 2017), I will start with the data already in a tidy format.

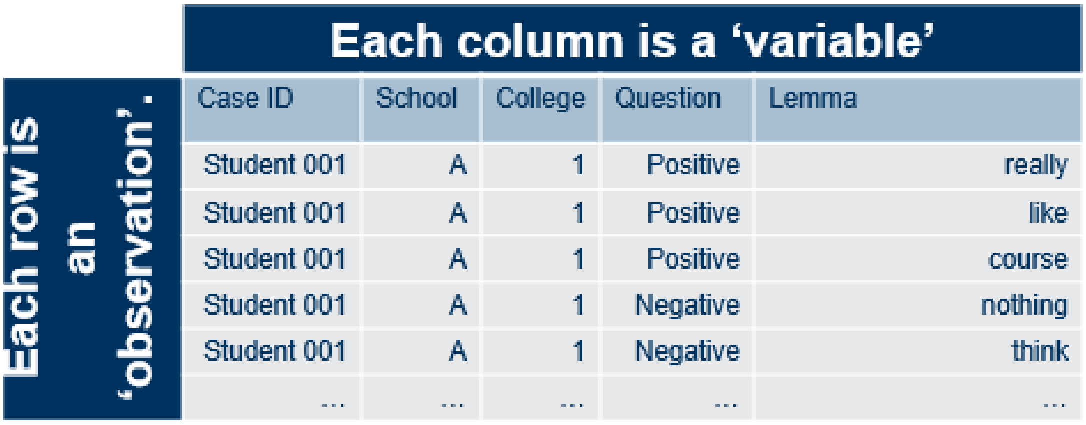
The key principle of tidy data is to recognise the observational unit of interest. In the case of survey data, each comment is its own observation (or row). This may seem counter-intuitive at first, as we expect the observational unit to be the individual student. However, in the case of tidy text mining, the individual respondent, or school, or question, is all the information attached to the comment. Therefore a student who provided a free text answer for all three questions would expect to have three rows in the dataset (Figure 1).


Each column is a 'variable'
Each row is an 'observation'.

Case ID	School	College	Question	Comment
Student 001	A	1	Positive	I really liked the course!
Student 001	A	1	Negative	Nothing I can think of
Student 001	A	1	One Thing	More free biscuits!
Student 002	B	1	Positive	It was okay
Student 002	B	1	Negative	I didn't like the exams
...	...	...	...	...

Figure 1: A pictorial representation of National Student Survey data in a tidy format.

The comment variable is then 'tokenised' or broken down into its constituent parts. In this case we are interested in the 'word' as the constituent part, but we may also be interested in word pairs (bigrams) or triads (trigrams). The words are then lemmatized, which simplifies words into their dictionary form over their many variants and improves pattern recognition for this type of analysis (Bergmanis & Goldwater, 2018). For example 'run', 'ran', 'runs', 'running' would be simplified to 'run', but 'runner' would be retained. Common words are removed from the dataset and so only meaningful words are left, associated with important information such as question answered, college, school and participant ID (Figure 2).



Each column is a 'variable'

Each row is an 'observation'.

Case ID	School	College	Question	Lemma
Student 001	A	1	Positive	really
Student 001	A	1	Positive	like
Student 001	A	1	Positive	course
Student 001	A	1	Negative	nothing
Student 001	A	1	Negative	think
...	...	...	...	...

Figure 2: A pictorial representation of National Student Survey comments tokenized and lemmatized

Example Results

With the data tokenized in this manner, it is possible to make comparisons in word frequency and word uniqueness across different groupings, and to visualize them in different manners.

For example, one can calculate a 'term frequency – inverse document frequency' across the three different questions asked of Edinburgh students, and explore what terms are unique to the response to the negative question, the positive question and the 'one thing' question (Figure 3).

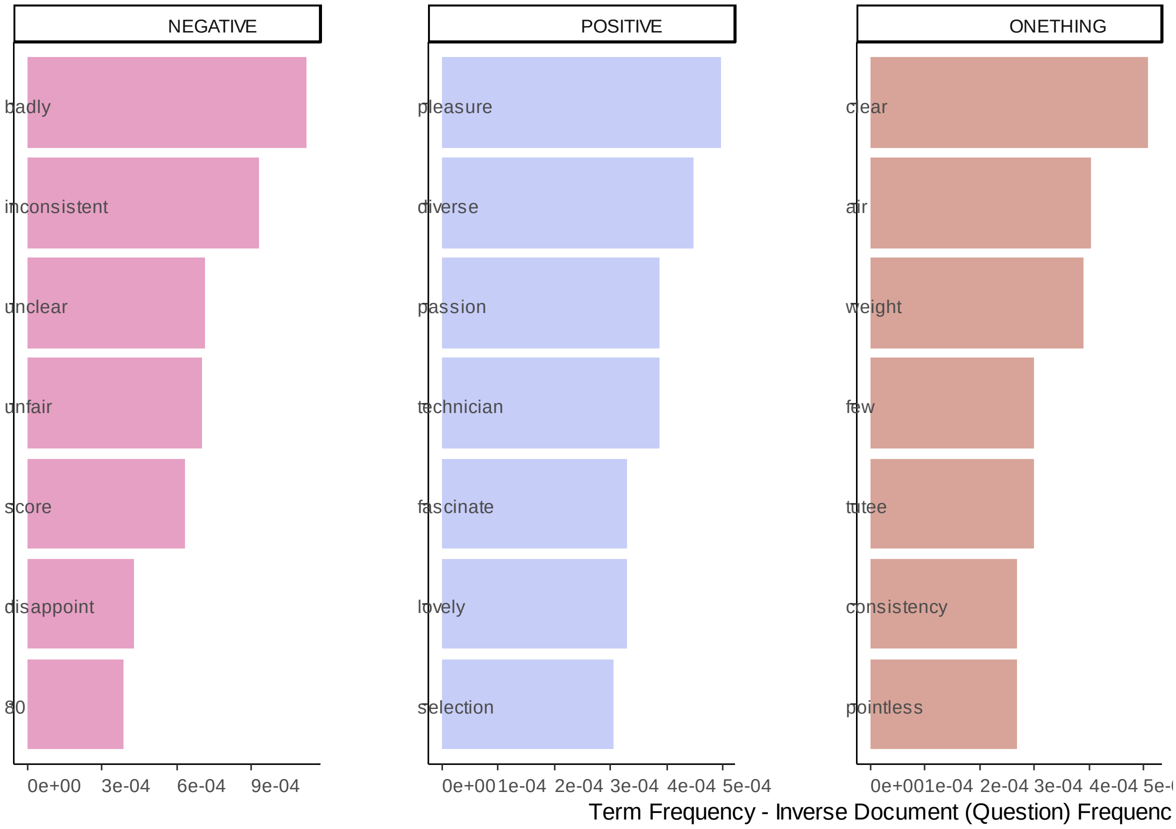


Figure 3: Measures of relative uniqueness of words across the three questions asked of University of Edinburgh final year students in the 2016 National Student Survey

In this figure, there are some ‘obvious’ conclusions to draw. For example, when asked to highlight particularly positive experiences, students focus on diversity, passion, friendliness of staff, and certain roles such as technicians. There are some results which are more confusing without comparisons made to in-depth qualitative studies of the same dataset (see MacKay, Hughes, Lent, Marzetti, & Rhind, 2018a; MacKay, Hughes, Marzetti, Lent, & Rhind, 2018b). Students repeatedly use the number ‘80’ when talking about a particularly negative experience and one thing they would change would be ‘air’. The other studies demonstrate there is a strong theme of students perceiving an ‘80%’ threshold within marking, and feel it is impossible to attain a higher grade in assessments (MacKay, Hughes, Lent, et al., 2018). Even so, further exploration of text mining can also be valuable. The meaning of ‘air’ was not immediately clear to me until further exploration of the mined data showed that ‘air’ was only ever mentioned within the responses to the ‘one thing’ question. The word was therefore unique to that grouping, even though it only occurred 5 times in total, and was used to highlight the air quality in labs and ‘an air of condescension’. These very unusual words can help show where small, potentially easy fixes to student experience can be made.

Correlations between terms can also be explored. The 5 most correlated words with two of the most unusual positive terms is shown in Figure 4. The correlations are calculated at the level of the individual, e.g. how often does each student use the same words, although they do not necessarily have to be next to one another in the text. As we have seen, words that occur a small number of times can bias the results, so to reduce this, I set an arbitrary threshold where a word had to occur 7 times or more. This is shown in Figure 4. In these correlations we see more evidence of certain schools and programmes, which may suggest that these programmes are doing something very well which should be explored in more detail. One could track these correlations over time, and explore how relationships change pre and post an intervention.

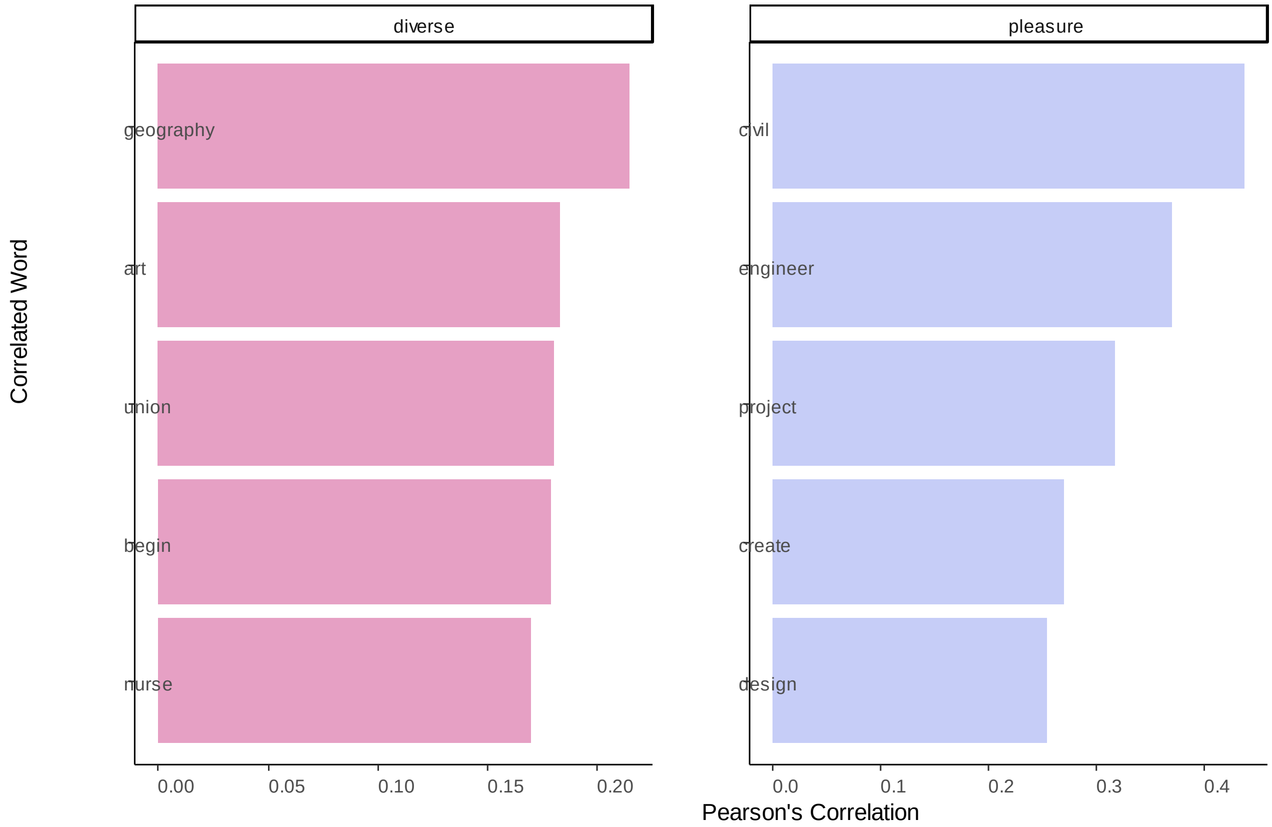


Figure 4: Pearson correlations with 'diverse' and 'pleasure' in response to the positive question asked of University of Edinburgh final year students in the 2016 National Student Survey (top 5 correlations shown).

Considerations of This Approach

Ethics

While the open science movement encourages the sharing of data and methods for reproducibility, there is a contesting drive to protect the identity of individuals and particularly limit the sharing of personal data through the General Data Protection Regulation in the EU (Cornock, 2018; Mourby et al., 2018). Sharing data workflows satisfies both requirements, but researchers and institutions should be very careful not to share data inadvertently. Although these datasets may seem large with thousands of comments, Figure 4 demonstrates these methods are quickly able to pull out information which identifies groups. With knowledge of the institution, it is possible to identify subject areas quickly in these types of analyses. Information identifying participants can still be an output in these types of analysis, although the risk is reduced. The statement of data use which participants agreed to should be carefully scrutinized prior to commencing any form of secondary analysis. For example, in this case NSS data should not be used to promote an institution, and the sharing of individual quotes for research purposes may be questionable given the exact wording of the Ipsos Mori consent statement (IPSOS Mori, 2017). However, there can be important and useful information within NSS data that would be valuable to compare across institutions, and shareable workflows may facilitate a deeper understanding of student experience across institutions. The higher education landscape is becoming more data-driven with discussion of how analytics and data-mining can be used to improve student experience and attainment (Baepler & Murdoch, 2010), but it is vital that we take a responsible approach and consider the ethical implications and possible privacy concerns of all methodologies. With care, these types of shareable workflows may be a useful method of providing repeatable and replicable analyses across confidential datasets.

No Short Cuts to Understanding

While this approach has many benefits, from its open principles to its speed and reproducibility, it should not replace a theory-led interrogation of such datasets. In this particular case, the text mining approach can help illustrate points, such as highlighting the 80% threshold for assignments, but this may not have been recognized without the more comprehensive analysis accompanying this paper.

Discussion and Future Directions

The QAA Enhancement Theme of engagement and evaluation has highlighted the importance of sharing practice across institutions and making better use of free text data (QAA, 2017), but we need to consider how feasible it is to conduct in-depth analyses of every survey. The adoption of text mining, with proper recognition of the importance of a more in-depth approach, may help the sector listen to the voice of staff and of students. While there are limitations, some of which have been suggested in this brief analysis, there are also opportunities for institutions to begin moving constructively towards a shared approach to analysis, if not a shared approach to data.

Given the levels of survey fatigue in students (Adams & Umbach, 2012; Porter, Whitcomb, & Weitzer, 2004), and the criticisms of questions in large surveys such as the NSS (Bennett & Kane, 2014), the higher education sector should consider why we survey so frequently, and question whether this is necessary. In theory, any number of surveys measuring student experience should not differ hugely in their results, if they are measuring the same underlying construct (Artino, La Rochelle, Dezee, & Gehlbach, 2014). The quantity of surveys in higher education suggests that the sector has little faith in the measurement of these constructs and it may be that the current surveys are not be specific enough, or may not be robustly designed and able to truly measure the underlying construct of interest. However, we will not be able to explore the meaningful differences across surveys until we adopt an approach that allows for more rapid analysis, particularly of free-text data.

At Edinburgh, we are exploring potential tools which could be developed to make this approach more feasible. We have provided the code alongside publications (MacKay et al., 2018a was accompanied by a link to a GitHub repository, freely available), and are considering the development of tools to make the workflow more accessible to those unfamiliar with programming languages. As we move forward with secondary analyses of datasets, we will incorporate these methodologies, and hope to provide guidance on how these types of analyses can inform institutional practice. The wealth of data available in higher education can be used to promote institutional resilience (King & Brennan, 2018; QAA, 2018), which is perhaps more important than ever in the current political climate. Making best use of existing data sources includes exploring how institutes can work together, with economy of resources and respecting their students' privacy, to tackle challenges in higher education.

Biographies

Jill MacKay is a research fellow in veterinary education at the Royal (Dick) School of Veterinary Studies at the University of Edinburgh. She teaches research methodology and professional skills to a range of undergraduate and postgraduate programmes within the school, and is interested in the methodologies we use to evaluate educational research. She is also found on Twitter @jilly_mackay and reposits her teaching code on github at https://github.com/jillymackay.

The associated code for this paper is available at: https://gist.github.com/jillymackay/7f360b5e57feecf833f23b0188eb64e7

Acknowledgments

The author would like to acknowledge the helpful comments from Susan Rhind, Kirsty Hughes, Neil Lent and Hazel Marzetti in the production of this manuscript.

References

Adams, M. J. D., & Umbach, P. D. (2012). Nonresponse and Online Student Evaluations of Teaching: Understanding the Influence of Salience, Fatigue, and Academic Environments. *Research in Higher Education*, *53*(5), 576–591. http://doi.org/10.1007/s11162-011-9240-5

Artino, A. R., La Rochelle, J. S., Dezee, K. J., & Gehlbach, H. (2014). Developing questionnaires for educational research: AMEE Guide No. 87. *Medical Teacher*, *36*(6), 463–74. http://doi.org/10.3109/0142159X.2014.889814

Baepler, P., & Murdoch, C. J. (2010). Academic Analytics and Data Mining in Higher Education. *International Journal for the Scholarship of Teaching and Learning*, *4*(2), Article 17. http://doi.org/10.20429/ijsotl.2010.040217

Bell, A. R., & Brooks, C. (2017). What makes students satisfied? A discussion and analysis of the UK's national student survey. *Journal of Further and Higher Education*, (September), 1–25. http://doi.org/10.1080/0309877X.2017.1349886

Bennett, R., & Kane, S. (2014). Students' interpretations of the meanings of questionnaire items in the National Student Survey.

Quality in Higher Education, *20*(2), 129–164. http://doi.org/10.1080/13538322.2014.924786

Bergmanis, T., & Goldwater, S. (2018). Context Sensitive Neural Lemmatization with Lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1391–1400).

Blair, B., Orr, S., & Yorke, M. (2012). *Erm, That Question... I Think I Probably Would've Just Put Something in the Middle and Sort of Moved on to the Next One, Because I Think It's Really Unclear': How Art and Design Students Understand and Interpret the National Student Survey. Group for Learning in Art and Design.*

Buckley, A. (2012). *Making it count: Refecting on the National Student Survey in the process of enhancement.*

Burgess, A., Senior, C., & Moores, E. (2018). A 10-year case study on the changing determinants of university student satisfaction in the UK. *PLoS ONE*, 1–15. http://doi.org/10.1371/journal.pone.0192976

Cocksedge, S. T., & Taylor, D. C. M. (2013). The National Student Survey: Is it just a bad DREEM? *Medical Teacher*, *35*(12), e1638–e1643. http://doi.org/10.3109/0142159X.2013.835388

Cornock, M. (2018). General Data Protection Regulation (GDPR) and implications for research. *Maturitas*, *111*, A1–A2. http://doi.org/10.1016/j.maturitas.2018.01.017

Grolemund, G., & Wickham, H. (2017). *R for Data Science* (1st ed.). O'Reilly.

IPSOS Mori. (2017). The National Student Survey Privacy Statement. Retrieved 8 January 2018, from http://www.thestudentsurvey.com/privacy-statement.php

King, R., & Brennan, J. (2018). *Data-driven risk-based quality regulation.*

MacKay, J. R. D. (2018). GitHub: NLPforNSS. Retrieved 23 July 2018, from https://github.com/jillymackay/NLPforNSS

MacKay, J. R. D., Hughes, K., Lent, N., Marzetti, H., & Rhind, S. M. (2018a). What do Edinburgh Students Want? A mixed methods analysis of NSS 2016 Free Text Data. In *The University of Edinburgh Learning and Teaching Conference: Inspiring Learning* (p. 19). Edinburgh.

MacKay, J. R. D., Hughes, K., Marzetti, H., Lent, N., & Rhind, S. (2018b). *Using National Student Survey (NSS) Qualitative Data to Explore Disciplinary Cultures Around Assessment and Feedback.*

Mourby, M., Mackey, E., Elliot, M., Gowans, H., Wallace, S. E., Bell, J., … Kaye, J. (2018). Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK. *Computer Law and Security Review*, *34*(2), 222–233. http://doi.org/10.1016/j.clsr.2018.01.002

Neary, M. (2016). Teaching excellence framework: A critical response and an alternative future. *Journal of Contemporary European Research*, *12*(3), 690–695.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. http://doi.org/10.1126/science.aab2734

NSSE. (2017). *Engagment Insights: Survey Findings on the Quality of Undergraduate Education. National Survey of Student Engagement.*

Porter, S. R., Whitcomb, M. E., & Weitzer, W. H. (2004). Multiple Surveys of Students and Survey Fatigue. *New Directions for Institutional Research*, (121).

QAA. (2017). *Evidence for Enhancement: Improving the Student Experience. Enhancement Theme 2017-2020.*

QAA. (2018). QAA Briefing: Helping providers get the most from their data. QAA.

QILT. (2018). *2017 Student Experience Survey National Report*.

Richardson, J. T. E., Slater, J. B., & Wilson, J. (2007). The National Student Survey: development, findings and implications. *Studies in Higher Education*, *32*(5), 557–580. http://doi.org/10.1080/03075070701573757

Shattock, M. (2018). Better Informing the Market ? The Teaching Excellence Framework in British Higher Education. *International Higher Education*, *92*, 21–22.

Silge, J., & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *The Journal of Open Source Software*, *1*(3), 37. http://doi.org/10.21105/joss.00037

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10). http://doi.org/10.18637/jss.v059.i10

Yorke, M. (2009). 'Student experience' surveys: Some methodological considerations and an empirical investigation. *Assessment and Evaluation in Higher Education*, *34*(6), 721–739. http://doi.org/10.1080/02602930802474219