QAA

The Wisdom of Students:

Monitoring Quality through Student Reviews

Alex Griffiths, Meghan Leaver and Roger King

June 2018

Contents

Foreword

A hallmark of our continuing digital revolution is the proliferation of online avenues for sharing ideas and opinions. Online it's easy to review services and products; express an opinion and share a view.

Now that all these views and opinions are available to us online, an obvious question is what can be done with the data? Are individual views just that - the discrete and disconnected views of individuals with little to tell us in aggregate about the things they focus on? Or do they reveal themes or patterns, and thus provide more obvious pointers towards changes which might improve services or products?

For higher education, this new research suggests that the views of students posted online do reveal patterns and themes. More specifically, it finds a positive association between 'social media sentiment' - students' unsolicited views of universities and colleges - and the results of existing measures of quality such as the Teaching Excellence and Student Outcomes Framework, and National Student Survey. As the authors note, this is just the beginning and more work is needed to really understand how it could be applied in a regulatory setting, but the potential is compelling.

The idea that social media sentiment might predict the outcomes of more traditional review or inspection outcomes is not a new concept. Alex Griffiths' earlier research in the healthcare sector showed a strong correlation between unsolicited online feedback and the Care Quality Commission's (CQC) hospital inspection findings.

What does this mean for how we assess higher education? It's unlikely that the use of this type of data will supersede the need for formal measures to assess quality and standards in UK higher education. Independent human expertise will always be needed to put data into context.

It also raises a few questions. At a time when people are increasingly wary about how their data is used by technology companies, would students accept their feedback being used by providers and regulators? Would making it a formal part of the regulatory system lead to providers incentivising students, or an observer effect - tainting the authenticity of the reviews? These are some of the risks which, along with the benefits, are thoughtfully considered by the authors in their work.

At QAA, we think there is merit in exploring this idea even further. That is why, from autumn 2018, we intend to pilot this approach to quality management with 10 providers. The pilot will not form part of formal quality assessment, but will test how this pattern holds over time. Universities and colleges might also find value in this sort of analysis for their own internal monitoring, spotting the early warning signs and putting right issues before they warrant regulatory intervention.

In a fast-developing higher education sector, new sources of information about students' learning and teaching experiences will continually emerge. I would like to extend my thanks to Alex Griffiths, Meghan Leaver and Roger King, for producing this report. I hope it will spark conversations about how we can capture and act on feedback from students wherever it can be found.

Will Naylor
Director, Colleges and Alternative Providers, QAA

Executive summary

> 'Our lecturers are amazing and have a great deal of experience not only in the university but also in industry. Because of our small class sizes, we can get a lot of contact time with the staff. My course is very practical and we spend a lot of time in the lab. A lot of emphasis is put on lab skills and good practise, which makes us highly employable. I have rated the facilities very highly because we have a large library and computer areas which are open 24/7. There are also drop-in facilities for help with key skills such as maths and writing essays.'
>
> 'Not very good, timetabling is very poor, there is mass room problems and a shortage of lecturers. The only good thing about my course was the year's placement that I did and also the fact I am assured a job. It is very average here.'
>
> An example of unsolicited student feedback on social media

Over the past two decades, the policies of successive governments have introduced more market-like characteristics to the higher education sector in England. This is best exemplified by the current governmental aim to place students at the heart of a consumer-led system, where the exercise of better-informed student choice among universities is intended to drive up competition, quality and learning innovation. At the heart of the new consumer-led system, will be the new regulator - the Office for Students (OfS) - mandated to operate a risk-based approach and to reduce the regulatory burden on (the great majority of) providers.

Risk-based and low-burden oversight in monitoring quality, standards and broad institutional compliance is to be achieved via data-driven self and external regulation. An important requirement for the new regulatory system is ensuring that data is timely, robust, low-burden, and representative of student views and experiences. However, several problems stand in the way of achieving these objectives, including cost, the need for configuring existing or designing new data collections, and avoidance of 'gaming the system' by providers.

One alternative source of information that may overcome many of the longstanding issues with regulatory data, is the unsolicited feedback of service users. Such data can be gathered in near real-time reports at various stages of the student experience (rather than simply at the end of the third-year of an undergraduate course, as with the National Student Survey), and bypasses providers' administrative systems. Moreover, in England, the automated scraping and analysis of millions of items of patient feedback using machine learning has been empirically proven to be an effective predictor of the outcome of in-depth hospital inspections by the CQC[1].

This research explores whether the ever-growing volume of online student feedback can provide insight into the quality of higher education provision - in short, it can. Over 200,000 (overwhelmingly positive) reviews were examined and a collective-judgement score comprising a time-limited moving average of the review scores for each provider was derived. This collective-judgement score had a positive association with Annual Provider Review (APR) outcomes, Teaching Excellence Framework (TEF) outcomes, and the overall satisfaction scores from the National Student Survey (NSS). Sixty days prior to the announcement of TEF ratings, the average collective-judgement score of providers subsequently judged to be 'bronze' was lower than that for providers subsequently judged to be 'silver', which was in turn lower than that of providers subsequently judged to be 'gold'. Similarly, the average collective-judgement score on the final data submission date for the APR was lower for those providers that required an action plan (demonstrating the need for

improvement) than those that did not. Finally, there was a positive, albeit weak, correlation between a provider's collective-judgement score on the day the NSS survey opened and their overall satisfaction score. Interestingly, where the collective-judgement score was only a weak predictor, the Whatuni.com and StudentCrowd.com data both individually demonstrated a strong positive correlation with the outcome of the NSS overall satisfaction question.

These findings suggest that the use of unsolicited student feedback could have significant benefits for regulators, providers and students. It also raises a number of challenges around the use of the data were it to be monitored more systematically, including its acceptance by all parties, coverage, and the changes in behaviour from both students and providers. Such an analysis has only been performed in one other sector to date, and no regulator has yet operationalised the systematic aggregation and review of service users' feedback in such a manner. This research, therefore, offers the chance for higher education to lead the way, but also means that there is additional research to be done.

Introduction

Government policy for higher education in England over the last two decades has focused on the marketisation of the sector. Current policy places students at the heart of a consumer-led system, where the exercise of better-informed student choice between universities is intended to drive up competition, quality and learning innovation. This move towards a more market-based system has resulted in a number of changes to the sector.

The most high-profile transformation in recent years has been the introduction and expansion of tuition-fee payments by students to providers sustained by an income-contingent loan repayment system. This increase in fees has been accompanied by a reduction in grants paid to providers by the Government. These funding changes have ostensibly generated a stronger student-consumer interest in learning processes and employment outcomes.

To support choice and improve quality in the more market-based higher education system, there has been an increase in the collection and availability of course, and provider-level, performance data. The results of large national data collections, such as the National Student Survey (NSS) and Destinations of Leavers of Higher Education (DLHE) survey, are widely reported and form the basis of multiple league tables that aim to rank providers and inform student choice. Regulators too have made use of this outcome data to support student choice and promote good-quality provision. The TEF framework, for example, uses data on student entry qualifications, student satisfaction, and employment, to broadly grade providers according to student performance.

Regulators' use of data has not been limited to teaching excellence however. In line with broader changes to regulatory practices, external oversight bodies have aimed to reduce the burden on (the great majority of) providers through data-driven, risk-based approaches. The rise of risk-based approaches has been accompanied by a gradual reduction in the remit of QAA in assessing and assuring teaching and learning quality in providers, and the rise in the perception of student experience and employment outcomes' data as more accurately indicating excellence in higher education provision. Most recently, the OfS has sought to regulate using lead indicators and periodic, but not cyclical, assurance judgements based on risk analyses of these data.

A key challenge to the success of this new regulatory environment is therefore ensuring that data is timely, robust, low-burden and representative of student views and experiences.

Several problems stand in the way of obtaining such data. First, shaped by the cyclical nature of higher education, existing data collections are annual and retrospective, providing a 'lagged' view on provision that makes timely interventions more challenging. Second, data must be collected by the providers whose performance will be judged on that data. This leaves the risk of providers 'gaming the system' in ways that may impact on the accuracy of the data submissions and sector-wide comparisons. Third, it is costly, time consuming and burdensome to develop new data collections geared to the specifics of a new system. Consequently, existing data collections tend to be adapted to a purpose quite different from their original design. Fourth, existing data collections measure what the designers of those measures deemed important, rather than what students in their growing role as consumers deem important.

One potential solution to the challenge of securing timely, robust, low-burden and insightful data may come from the student body. Francis Galton stumbled upon the 'wisdom of crowds' phenomenon after coming across a competition to guess the butchered weight of a live ox at a local fair. There were 800 competitors, most of whom were not experts in cattle or

butchery, submitting their guesses on numbered cards. Curious about the entrants' guesses, Galton borrowed the entry slips once the competition was over and the weight of the butchered ox determined to be 1,198 lbs. Much to his surprise, the average of the entrants' guesses was 1,197 lbs, essentially perfect. The 'Wisdom of Crowds' phenomenon that Galton had discovered means that, under the right circumstances, groups can be remarkably insightful. This can be the case even if the majority of people within a group are not especially well informed or rational[2]. Whilst we as individuals seldom have all the necessary facts to make an accurate assessment, and are subject to numerous heuristics and biases, when our individual assessments are aggregated in the right way, our collective assessment is often highly accurate.

This phenomenon has already been shown to be effective in healthcare where the collected online feedback and social media postings of patients and their families has been proven to be a statistically-significant predictor of the outcome of subsequent in-depth quality inspections. Not only was the collective judgement of patients an effective predictor of quality outcomes, it was available in a more timely manner than existing data sets, at a more granular-level, offered new insights at different stages of the care journey, and added no burden to providers' existing duties. Despite its proven value in healthcare, the use of unsolicited service user feedback as a regulatory tool has been the subject of little research and, to date, no regulator has incorporated the systematic gathering and analysis of such feedback into their oversight approach. This research explores, for the first time, whether student reviews might be used to identify the quality of higher education provision. The data and statistical methods used for this analysis are described below.

Data and methods

This section describes the general trends found in the student review data, the individual sources of the student reviews used for this study and the other quality measures that they have been compared to, and the methods used to explore the relationship between the student reviews and other measures of the quality of higher education provision.

Data landscape

There is a significant volume of student feedback available online. For this research, over 210,000 reviews were gathered from 165 higher education institutions, 211 further education colleges and 12 alternative providers in the UK considered in scope. These reviews came from three sources: Facebook, Whatuni.com and Studentcrowd.com. This may, however, be only a fraction of the data available. In the past seven months, over 2.6 million tweets have mentioned providers' main Twitter accounts. If just two per cent of those tweets relate to the student experience - as is the comparable rate in healthcare - that would represent an additional 90,000 items of feedback a year even before departments, career services, students' unions and other accounts are considered. Some data sources require more work than others, however, and significant investment in resource required to identify and score relevant tweets amongst the millions available have meant its inclusion is beyond the scope of this research. Twitter data will be included in future research.

The significant volume of reviews available online is a relatively new phenomenon. As Figure 1 shows, the number of student reviews made each year, using the three data sources considered as part of this study, has grown significantly since 2012. By the end of March 2018, over 15,000 further reviews had already been posted.

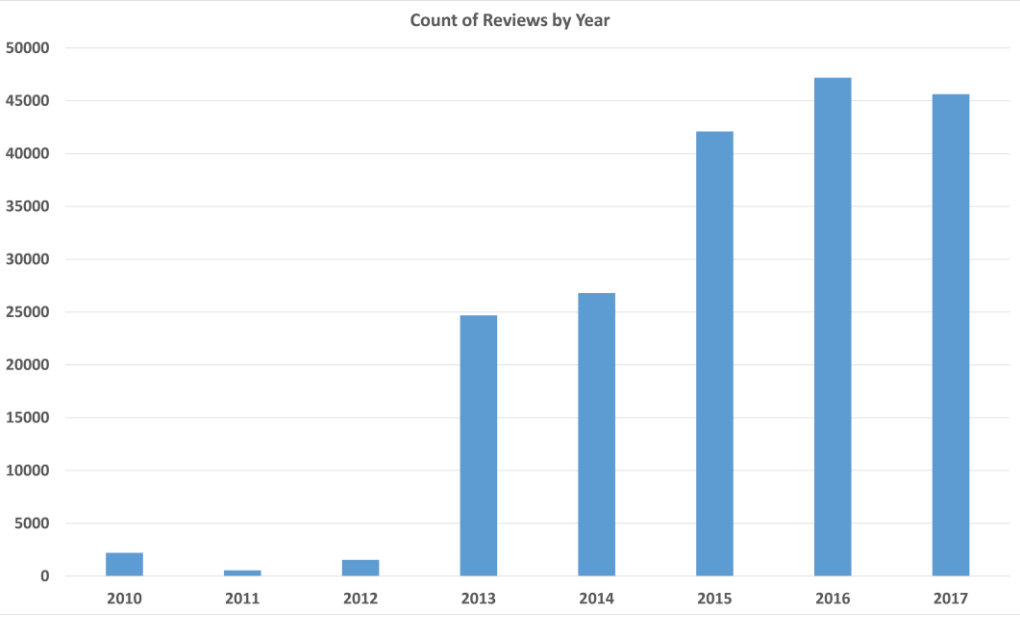

Figure 1: The number of reviews available by year from 2010-2017.

Unsurprisingly, the majority of student reviews concern Higher Education Institutions (HEIs). When the number of students by provider type is taken into account, Alternative Providers (APs) are underrepresented by the student reviews. As shown in Table 1 below, the largest source of student reviews for both HEIs and APs is Whatuni.com, with Facebook being the largest source for Further Education Colleges (FECs). However, it should be noted that, as with some official data sources[3], it is difficult to differentiate reviews of HE and FE provision

on FEC's Facebook pages and it is certain that some reviews of a colleges' FE provision will have been included.

	HEI	FEC	AP	Total
Whatuni.com	120,080	678	898	121,656
Facebook	58,108	14,880	436	73,424
StudentCrowd.com	14,960	21	35	15,016
Total	193,148	15,579	1,369	210,096

Table 1: The distribution of student reviews by source and provider type.

For each of the three data sources that have been used in this research, users are required to rate their overall experience on a scale of one (worst) to five (best) stars (Table 2). The scores can relate to the reviewer's overall experience of the provider, or their opinion of a specific course, department or set of facilities within a provider. Reassuringly for the sector, the average score over the 210,000 reviews is 4.18 stars, suggesting a high level of overall satisfaction with UK higher education.

	2014	2015	2016	2017
HEI	4.22	4.24	4.21	4.18
FEC	4.20	4.12	4.06	4.18
AP	4.06	4.42	4.07	3.91
Combined	4.22	4.23	4.21	4.18

Table 2: The average review score by provider type and year.

This high level of satisfaction seems to have remained consistent over the past four years and across the three provider types. Caution should be taken, however, when considering the fall in the average review score for APs given the limited number of reviews - just 212 in 2017.

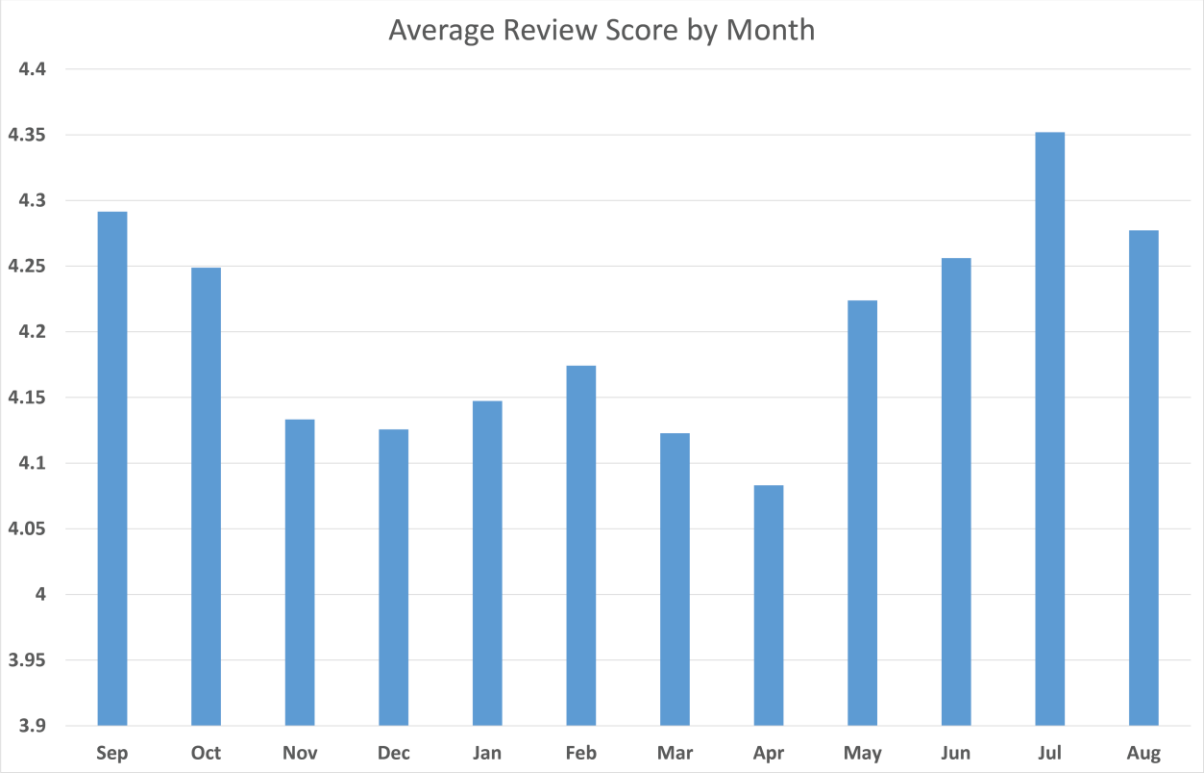

Figure 2: The average review score by month across all provider types.

When we look at satisfaction across the year (Figure 2), we see that the reviews are most positive at the start and the end of the academic year with a small lift in February. The higher scores at each end of the academic year can most likely be explained by student excitement and relief, and the associated optimism and hindsight biases, at starting and finishing an academic year. The slight increase in average review scores in February may be inherent natural variation, a result of Whatuni.com successfully gathering a greater number of reviews than normal in that month, or the possible effect of providers making an effort in those months to ensure a good National Student Survey (NSS) score. It is likely to be a result of all three.

Individual data sources

Despite the common rating system, each data source has its own unique challenges. These challenges are explored in turn below after a brief comparative overview of the data sources.

Student reviews

Whatuni.com

With over 121,000, Whatuni.com had the greatest number of reviews. Although Whatuni.com included reviews for 79 FECs and 8 APs, there were only just over 1,500 reviews for these providers in total compared to over 120,000 for HEIs. Each review offered more in-depth ratings covering topics such as 'course and lecturers' and the 'city life'; however, these were considered to be out of scope for this project. With an average overall review score of 4.11 the Whatuni.com reviews were overwhelmingly positive.

The majority of Whatuni.com reviews detail not only the provider, but the course that reviewers are studying, and tend to be well written. One disadvantage of the Whatuni.com data is that it tends to appear in batches. For example, there were no reviews of King's College London in April, May or June 2017, but over 100 reviews between 22 January and 7

February 2018. The reasons for the periodic nature of comments is not clear, but it may be the result of promotional pushes for reviews, Whatuni.com checking that comments do not identify individuals or constitute libel in batches before posting, or some other reason.

Facebook
Only 13 of the 378 providers considered in this research did not have a Facebook page. Of the 365 that did, 231 had the 'Reviews' function enabled, giving people the chance to rate them and optionally leave a comment. Further to provider-level pages, we systematically searched for pages relating to departments, schools, institutes, faculties, students' unions, and career services at each provider. This search yielded an additional 731 pages with the 'Reviews' function enabled. Provider-level reviews accounted for almost 54,000 of the 73,424 reviews on Facebook, with just fewer than 20,000 reviews identified at the subprovider-level.

Unlike the Whatuni.com and StudentCrowd.com data, the Facebook data did suffer from what appeared to be irrelevant comments that we attempted to clean from the data. These reviews were accompanied by text from people who: were looking for accommodation or flatmates, had never visited the provider but were looking for advice, were promoting their businesses, or were advertising jobs. Approximately 600 such comments were removed.

The majority of the reviews on Facebook appeared to be from current or recent students, there were also reviews from parents having taken their children to open days, job applicants, former staff and alumni from decades ago. With an average review score of 4.33, Facebook had the most positive reviews of the three data sources.

StudentCrowd.com
StudentCrowd.com data reviews are nearly exclusively focused on HEIs. Reviews cover providers, specific courses and halls of residence. For this research, we identified over 6,000 reviews at provider-level, and over 8,000 reviews at course-level. As with Whatuni.com, each StudentCrowd.com review offered more in-depth ratings covering topics such as campus/facilities and the quality of internet/Wi-Fi. Again, as with Whatuni.com reviews, these more in-depth ratings were considered out of scope for this study. With an average overall review score of 4.08, the StudentCrowd.com reviews were more negative than those from Whatuni.com and Facebook, but were still overwhelmingly positive.

Other measures of quality

Annual Provider Reviews
As part of its new operating model for quality assessment in 2016-17[4], HEFCE conducted its first Annual Provider Review (APR) in 2016-17 of the HEIs and FECs it funded. The APR considers existing data that is used in a 'contextualised and rounded way' with judgements on quality and standards matters reached through peer review. Of the 376 HEIs and FECs included in this study, 330 had an HEFCE APR rating as of 20 March 2018. Over 90 per cent of those providers, 302 to be exact, were judged as 'meets requirements' and warranted no further action, and 20 were judged as 'meets requirements with an action plan', meaning an action plan was required to address areas of immediate concern. At the time of writing, eight providers remain as 'pending', where a final outcome is yet to be decided. As the outcome of the 'pending' reviews is still not clear, these providers were not formally included in the analysis of the relationship between student reviews and APR outcomes[4].

The collective-judgement score for each provider (see Statistical Methods section below for a full description of the collective-judgement score) with an APR outcome of 'meets requirements' or 'meets requirements with an action plan' was taken for 1 December 2016, chosen as it was the deadline date for the submission of data for the APR[5], and compared with the APR outcomes published as of 20 March 2018.

Teaching Excellence Framework
The Teaching Excellence and Student Outcomes Framework (TEF) was designed by the Department for Education to recognise teaching quality in UK higher education providers by rating them as Gold, Silver or Bronze. TEF year two ratings were awarded in 2017 and were judged based on benchmarked metrics - taking into account the context of each provider and differences in students' backgrounds, entry qualifications and subjects studied - and a qualitative provider submission. Both the data and qualitative submission were considered by a panel of experts before the final judgement was made. The benchmarked metrics concerned students' views on 'the teaching on my course', 'assessment and feedback', and 'academic support' as captured by the NSS; non-continuation rates; and rates of employment or further study six months after graduation. TEF awards for Year Two were at provider level and concerned undergraduate students only[6].

Participation in TEF Year Two was voluntary; however, it was believed that participation would be necessary for English providers if they wished to raise their fees in 2017/18. Subsequently, the Government decreed that English fees would be frozen regardless of TEF participation or outcome. In total, 295 providers participated in Year Two TEF of which 56 received a Bronze award, 116 a Silver award, and 59 a Gold award. A further 64 providers received a 'provisional' award where there was insufficient data to fully assess the provider.

Following discussions with QAA, each provider's collective-judgement score taken on 22 April 2017, 60 days prior to the announcement of the TEF awards, was compared with their TEF award.

National Student Survey
The National Student Survey is primarily aimed at final-year undergraduate students and covers a number of aspects of the student experience including 'overall satisfaction'. To date, nearly three million students have taken the NSS and return rates are consistently around the 70 per cent mark. In 2017, students at a number of alternative providers participated in the survey for the first time[7].

As we do not differentiate the student reviews considered in this analysis by the focus of the review - for example, reviews of the overall experience are treated no differently to those where the qualitative comments relate specifically on one aspect of provision - we have focused solely on the 'overall satisfaction' question at the end of the NSS. NSS performance can be assessed by absolute performance - what was your score - or by benchmarked performance - how far did your score differ from what would be expected taking into account the courses you teach, the characteristics of your student population, and so on. Both the absolute and benchmarked performance of providers have been considered as part of this research[8].

The results of the 2015, 2016 and 2017 NSS were considered with the collective-judgement score of each provider on the day the NSS opened compared with the results.

Statistical methods

For each review, we knew the one to five-star rating awarded by the student, the date the review was made and the provider it concerned. This allowed us to calculate a moving average from each data source for each provider on any given date. Moreover, by combining the reviews from the three data sources, we were able to create a collective-judgement score for each provider on any given date. By increasing the volume and diversity of reviews, the 'wisdom of crowds'[2] effect suggests that this collective-judgement score will offer a better reflection of the quality of providers and align with other measures of quality.

In developing the collective-judgement score, two key decisions needed to be made. First, what time-period should the scores cover? Only including reviews from the past 30 days would result in a collective-judgement score very responsive to new reviews, but would also be so volatile and based on too few reviews to be usable. Conversely, if the collective-judgement scores include all data from the past three years then there will be enough reviews included in each score to make nearly all of them robust, but at the expense of a lack of responsiveness and with it, an ability to account for recent improvements or falls in performance. Second, what should be the minimum number of comments comprising the collective-judgement score for it to be considered robust? Few would argue that a single comment over the period of a year is sufficient to indicate the quality of a provider, but there is a trade-off to be made between robustness and inclusiveness; the greater the number of reviews required, the lower the number of providers, especially FECs and APs, that will have a robust collective-judgement score.

To determine the answer to these questions, the collective-judgement scores were calculated using a period of 180 and 365 days, both with a minimum number of 5 and 10 reviews required for them to be considered robust. When comparing providers' collective-judgement scores to their APR, TEF and NSS outcomes, it was clear that a 365-day collective-judgement score was marginally better than a 180-day score, as was a minimum count of 10 reviews in that time-period rather than 5. A 365-day period comprising a minimum of 10 reviews was therefore chosen as the requirement for the collective-judgement score to be considered as part of the analyses. In opting for 10, rather than 5, as the minimum number of comments required for a collective-judgement score to be considered robust, the number of additional providers for which no collective-judgement score could be generated was minimal.

Results

The overall finding of this research is that providers' collective-judgement scores are positively associated with their subsequent APR, TEF and NSS outcomes. This is true of each individual data source, each is positively associated with the outcomes of the APR, TEF and NSS. The collective-judgement score created by combining all the reviews proves an even more effective predictor than the individual data sources for APR and TEF outcomes, whereas the opposite is true for the NSS. A poor collective-judgement score does not guarantee a provider will perform poorly on other quality measures; however, a provider with a poor collective-judgement score has a greater probability of performing poorly on other quality measures. Likewise, a provider with a good collective-judgement score is not guaranteed to be performing well on other quality measures but has a greater likelihood of doing so.

There are any number of analyses that could be performed with such a rich data set. For this research, we have focused on two. First, the relationship between the student reviews and other measures of quality. Second, the utility of large and sudden moves in the collective-judgement score to identify concerns for both providers themselves and oversight bodies.

The relationship between collective-judgement scores and other measures of quality

Annual Provider Review outcomes

Figure 3 below is a box plot of the collective-judgement scores for providers that were judged as either 'meets requirements with an action plan' or 'meets requirements' and for which the collective-judgement score comprised 10 or more reviews.

A box plot can be interpreted as follows. The thick black line in the middle of each box represents the median collective-judgement score - that is, the score halfway through the list when they are ordered by size, for providers with the specified APR outcome. The top and bottom edges of each white box represent the third and first quartile collective-judgement scores - that is, the score three-quarters and one-quarter of the way through the list when ordered by size, respectively. Finally, the top and bottom lines (or whiskers) indicate the collective-judgement score that is 1.5 times greater and less respectively than the interquartile range (the difference between the third and first quartile). Alternatively, if no collective-judgement scores exceed this value, the top and bottom lines (or whiskers) represent the highest or lowest collective-judgement score. Where outlying values exceed 1.5 times the interquartile range, they are shown as individual dots. The number of providers in each category and their mean collective-judgement score is written at the top of the plot.

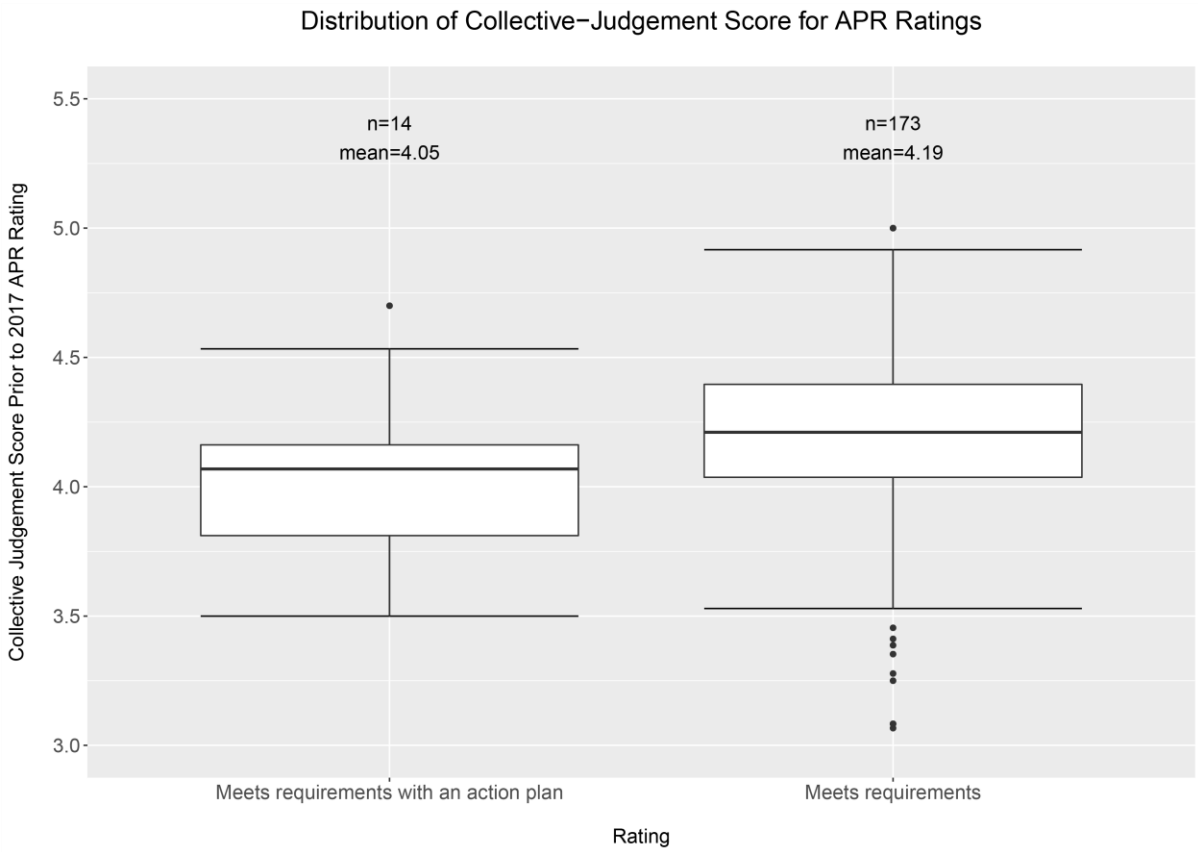

Figure 3: A box plot of the collective-judgement scores on the data submission deadline for the 2016-17 APR.

The mean and median collective-judgement score prior to the APR data submission deadline is lower for the 14 providers judged 'meets requirements with an action plan' and that had a robust collective-judgement score available than for the 173 providers judged 'meets requirements' with a robust collective-judgement score available. Some caution must be taken with these results given the limited number of providers judged 'meets requirements with an action plan'; however, it is encouraging to see that there is limited overlap in the interquartile ranges, represented by the white boxes in Figure 3, suggesting differentiation between providers' subsequent APR outcomes based on their collective-judgement scores.

	Collective judgement		Whatuni.com		Facebook		StudentCrowd.com	
APR outcome	Count	Average score	Count	Average score	Count	Average score	Count	Average score
Meets requirements	173	4.19	91	4.19	154	4.36	42	4.09
Meets requirements with an action plan	14	4.05	6	4.13	14	4.23	4	3.08

Table 3: A breakdown of both the average collective-judgement score and moving average score for individual data sources, and count of providers with a sufficient number of reviews (a minimum of 10 in the preceding 365 days) to be included in the analyses.

Table 3 shows that, as would be expected, the collective-judgement score provides a robust score - that is, a score containing 10 or more reviews in the preceding 365 days, for a greater number of providers than the three data sources considered in isolation. Unfortunately, only 187 out of the 322 providers judged as either 'meets requirements' or 'meets requirements with an action plan' had a sufficient number of reviews to form a collective-judgement score. Of these 187 providers, 118 were HEIs and 69 were FECs. Alternative providers were not part of HEFCE's APR process, and therefore were not included in this analysis.

The individual data sources all demonstrated the same association between their moving average scores and the subsequent APR outcome. StudentCrowd.com data provided the clearest distinction between the two APR outcomes; however, it also had the lowest coverage. Facebook data had by far the best coverage, especially for providers that required an action plan, and the second-best differentiation between providers based on their APR outcomes.

TEF ratings

Figure 4 below shows that the relationship between the collective-judgement scores and TEF outcomes is similar to the relationship between the collective-judgement scores and APR outcomes. Both the mean and median collective-judgement score for Bronze-rated providers are lower than the mean and median collective-judgement scores for Silver-rated providers, which in turn are lower than the mean and median collective-judgement scores for Gold-rated providers.

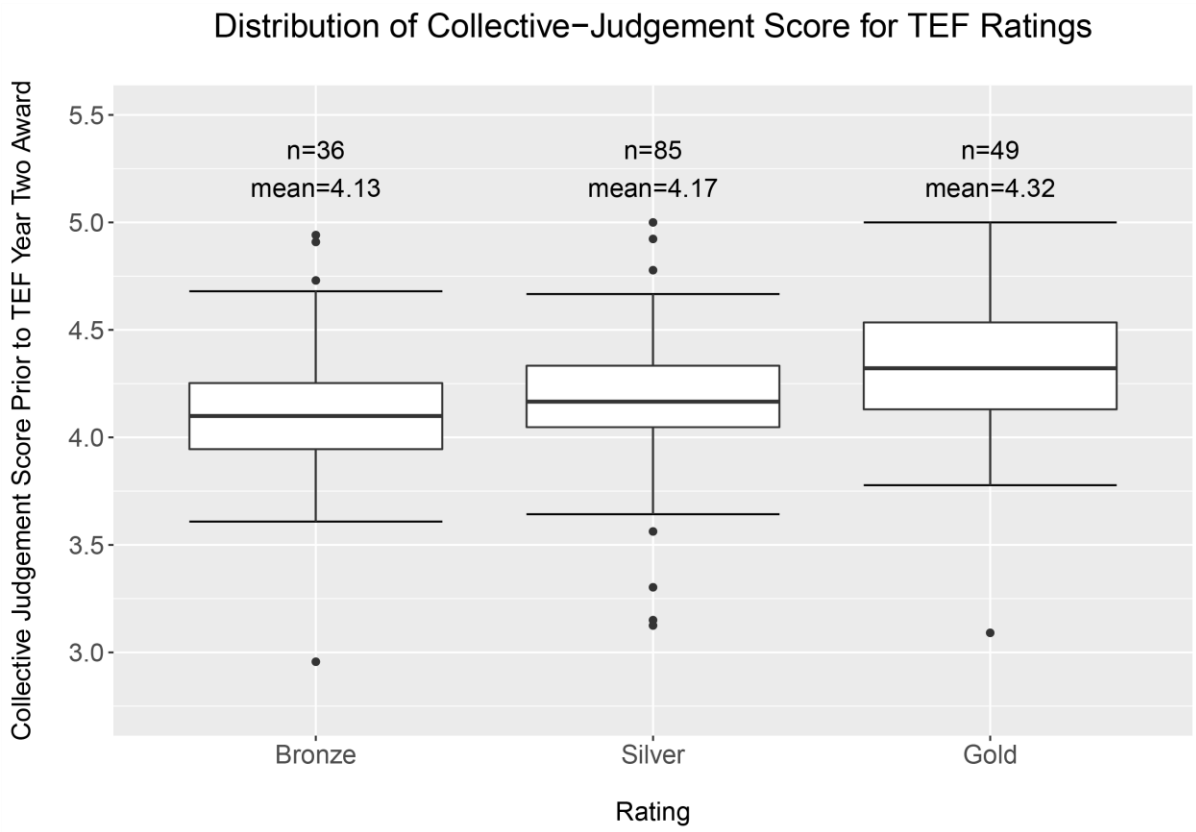

Figure 4: A box plot of the collective-judgement scores on 22 April 2017, 60 days prior to the announcement of the TEF awards.

For the TEF analysis, there is greater overlap between the interquartile ranges and the whiskers of the box plot for the Bronze, Silver and Gold-rated providers. This suggests that, whilst the collective-judgement score provides a statistically-significant predictor of a provider's TEF outcome overall, there is a significant variation of scores within the Bronze, Silver and Gold-rated providers and a number of outliers that significantly buck the general trend. It is interesting to note that the distribution of the collective-judgements scores was greater for Gold providers than for Bronze or Silver providers, suggesting that it may be more challenging for students to identify Gold provision than Silver or Bronze provision.

	Collective judgement		Whatuni.com		Facebook		StudentCrowd.com	
TEF outcome	Count	Average score	Count	Average score	Count	Average score	Count	Average score
Gold	49	4.32	33	4.19	43	4.59	22	3.91
Silver	85	4.17	61	4.03	79	4.43	30	3.61
Bronze	36	4.13	18	3.99	35	4.41	9	3.57

Table 4: A breakdown of both the average collective-judgement score and moving average score for individual data sources, and count of providers with a sufficient number of reviews to be included in the analyses.

Out of the 231 providers with a Gold, Silver or Bronze TEF rating, 170 had a sufficiently robust collective-judgement score to be included in this analysis. Of the 170 providers, 127 were HEIs, 40 were FECs and 3 were APs. For those providers where the collective-judgement score was sufficiently robust, we can again see that each individual data source displays the same positive association as the collective-judgement score. In each case the difference between the average score for Bronze and Silver-rated providers is notably smaller than the difference between Silver and Gold-rated providers. Once again, we see that the StudentCrowd.com data demonstrates the greatest difference between providers with different ratings; however, it also has the lowest coverage. Moreover, we again see that Facebook has the greatest coverage with remarkably similar differentiation between providers as the Whatuni.com data.

NSS overall satisfaction scores

The analysis of the relationship between student reviews and the outcome of the NSS is necessarily different from the analysis for the APR and TEF outcomes, as the NSS result for each provider is a continuous score out of 100, rather than a discrete rating. Instead, the relationship between the collective-judgement score and NSS outcomes is assessed by the correlation between the two variables. Here we have used the Pearson correlation coefficient which is a measure of the relationship between two variables that ranges from -1, indicating a strong negative relationship, to +1, indicating a strong positive relationship. A coefficient of 0 indicates no relationship between the two variables. The closer the correlation coefficient gets to ±1 (and thus the further it gets from 0), the stronger the relationship between the two variables. A strong correlation coefficient does not imply a causative relationship - for instance, sales of ice cream and sunglasses are strongly correlated, but buying ice cream does not make people buy sunglasses or vice versa.

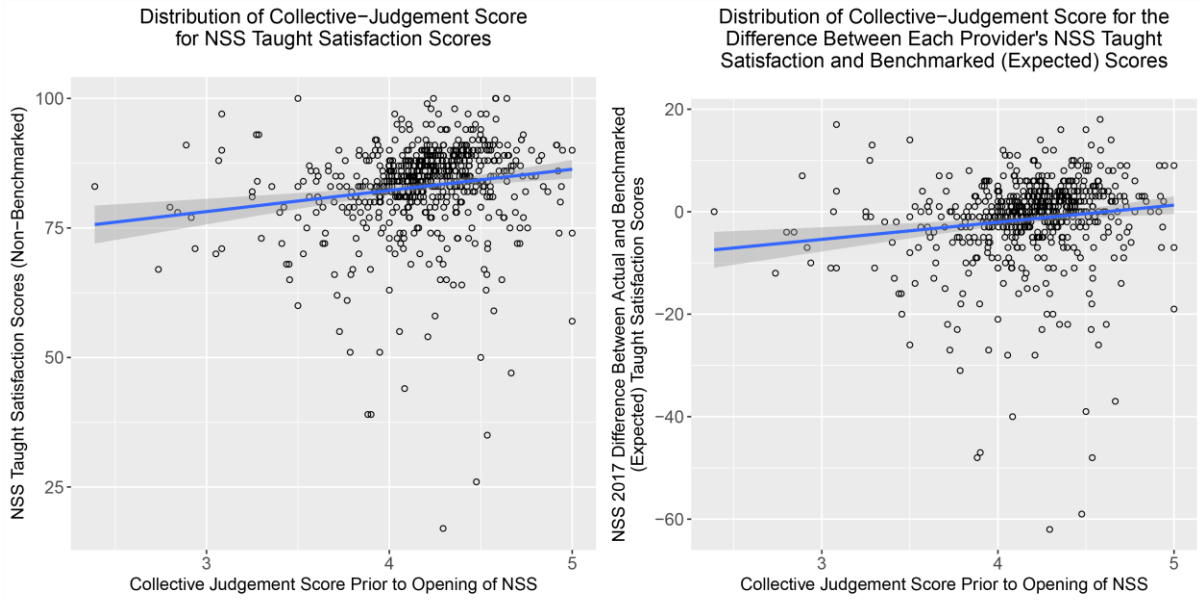

Figure 5: Two scatterplots of the collective-judgement scores on the opening day of the 2015, 2016 and 2017 NSS and, respectively, the non-benchmarked and benchmarked NSS overall satisfaction scores for taught students. Both plots feature a linear line of best fit with confidence intervals shown by the dark grey banding around the line.

As Figure 5 shows, there is a weak positive correlation between the collective-judgement score and both the non-benchmarked and benchmarked 'overall satisfaction' score (Pearson's correlation coefficients of 0.16 and 0.13 respectively). This suggests that, the higher a provider's collective-judgement score on the opening day of the NSS, the greater

the likelihood that they will do well on the NSS. This is far from a robust relationship, however, with a significant number of providers doing well on the NSS with a poor collective-judgement score and vice versa.

	Collective Judgement			Whatuni.com			Facebook			StudentCrowd.com		
Year	n	Corr	Corr (Bmrked)	n	Corr	Corr (Bmrked)	n	Corr	Corr (Bmrked)	n	Corr	Corr (Bmrked)
2015-2017	629	0.16	0.13	327	0.46	0.41	563	0.21	0.18	164	0.41	0.28

Table 5: A breakdown of correlations between the collective-judgement score and individual data sources, and NSS scores where there was a sufficient number of reviews to be included in the analyses.

Unlike the analyses of the APR and TEF outcomes, here we can see that aggregating the student reviews from the three distinct data sources weakens the relationship between the student reviews and the NSS scores. Why this is the case is not immediately obvious. Looking at the performance of the individual data sources, we can see that Whatuni.com this time demonstrates the strongest relationship with the NSS non-benchmarked satisfaction score. For all three data sources, the correlation with the student reviews is slightly stronger for the non-benchmarked NSS score than for the benchmarked score. Why the correlation is notably weaker for the Facebook data is not clear. The obvious reason, that Facebook contains more FEC data some of which will relate to non-HE activity, does not appear to explain the results as the correlation coefficient drops to 0.14 and 0.15 for the non-benchmarked and benchmarked scores respectively when only HEIs are considered.

If we break the analysis down over the three years of NSS data, we can see that the relationship between both the collective-judgement score and the individual data sources, and both the non-benchmarked and benchmarked NSS 'overall satisfaction' scores has remained consistent. This is despite a significant increase in the number of alternative providers entering the NSS in 2017 following the requirement that all designated providers do so[9]. Of the 629 instances where a provider has had 10 or more student reviews in the 365 days prior to the opening of the NSS between 2015 and 2017, 403 have concerned HEIs, 218 have concerned FECs, and just eight have concerned APs.

	Collective Judgement			Whatuni.com			Facebook			StudentCrowd.com		
Year	n	Corr	Corr (Bmrked)	n	Corr	Corr (Bmrked)	n	Corr	Corr (Bmrked)	n	Corr	Corr (Bmrked)
2015	207	0.14	0.13	95	0.52	0.48	192	0.21	0.20	0	-	-
2016	211	0.16	0.14	119	0.45	0.37	186	0.24	0.22	100	0.41	0.36
2017	211	0.16	0.15	113	0.53	0.46	185	0.19	0.16	64	0.44	0.35

Table 6: A three-year breakdown of correlations between the collective-judgement score and individual data sources, and NSS scores where there was a sufficient number of reviews to be included in the analyses.

Plotting the relationship between the Whatuni.com data and the 2017 NSS, the strongest relationship detailed above with correlation coefficients of 0.53 and 0.46 for the non-benchmarked and benchmarked data respectively. Figure 6 below shows that the trend is far more pronounced than for the collective-judgement score across the three years shown in Figure 5 above.

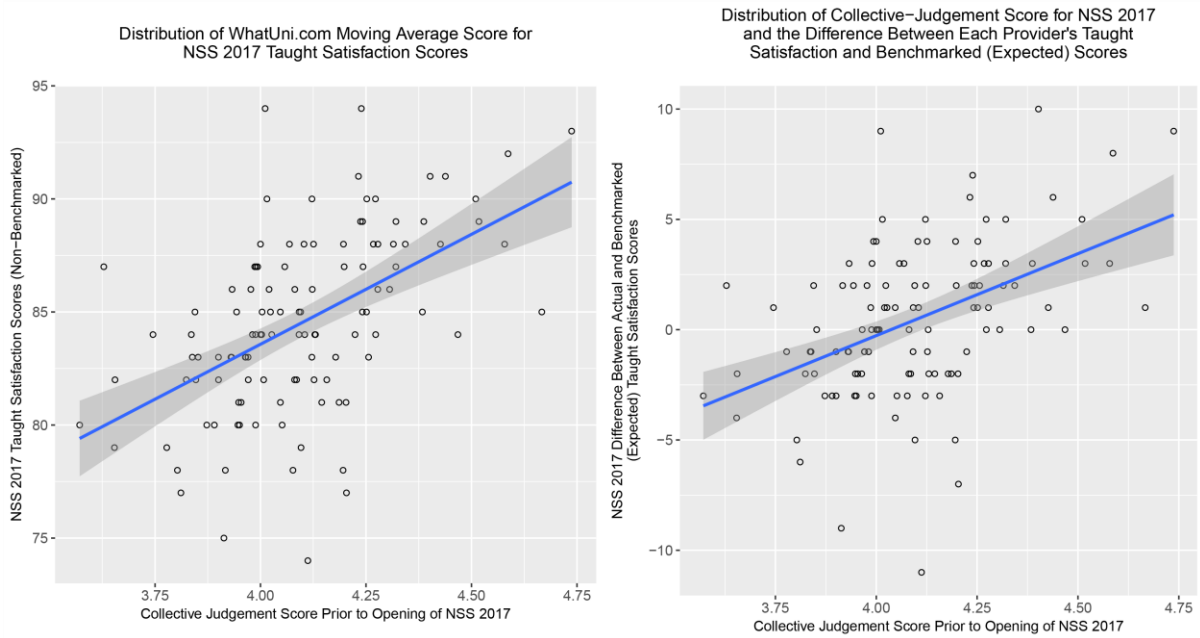

Figure 6: Two scatterplots of the Whatuni.com moving average scores on the opening day of the 2017 NSS and, respectively, the non-benchmarked and benchmarked NSS overall satisfaction scores for taught students.

The above results show that there is a clear association between the collective judgement of students and the subsequent outcome of the APR and TEF, and, to a lesser extent, the NSS. Whether rapid changes in this collective judgement can be used to identify significant concerns is explored below in the second of the two analyses in this study.

Marked changes in the collective-judgement score

So far, we have seen that the collective-judgement score formed by aggregated student reviews from multiple sources, has a positive association with, and can be an effective predictor of, the outcome of other quality measures. This in itself can be useful both as a quality improvement tool for providers and as a risk prioritisation tool for oversight bodies. It may be the case that users do not wish to focus solely on whether a provider collective-judgement score is good or bad, but also on its direction of travel. In this section, therefore, we analyse the cases where there have been significant shifts in the collective-judgement score of a provider over a short period of time.

As with constructing a collective-judgement score, when monitoring sudden changes, we must decide upon a minimum number of student reviews as a baseline and the time period over which changes should be considered. Building on the work above, we determined that in order to prevent too many unnecessary alerts and those alerts not being considered robust, any significant changes flagged for users should comprise at least ten reviews at both the point in time the change is being recorded from and to. This prevents smaller providers raising alerts every time a new review is posted, or an old review is discarded from the moving average. Deciding on the time period requires balancing immediacy of the alert -

and with it, the ability to act quickly to student praise or concerns - against there being sufficient time for new reviews to impact upon a moving average, and do so in a way that is not overly sensitive. It was decided that looking for changes over a 60-day period provided a good balance between these two competing priorities.

Large movements in a provider's collective-judgement score were often the result of new data sources becoming available. All three data sources considered in this study have different characteristics with Facebook reviews being more positive than Whatuni.com reviews, which in turn are more positive than StudentCrowd.com reviews. When a provider was first reviewed using one of these three data sources, frequently by multiple reviewers in short succession, their collective-judgement score would often shift significantly. Moreover, a department or careers team, for example, starting a Facebook page could lead to an initial surge in targeted reviews affecting the collective-judgement score. Once sudden changes resulting from new data sources had been discounted, it became clear that there were two sets of circumstances resulting in fast and significant changes to a provider's collective-judgement score.

The first cause of sudden changes was concerning specific issues motivating reviewers. In the most extreme example, one FEC saw its collective-judgement score drop from 4.66 to 3.21 within 60 days in late 2017. This was the result of five one-star reviews in the space of three days. Of these five reviews, three were accompanied by review text. One review appeared to be a genuine grievance left by a student who had left the college after a short period; however, the remaining two reviews voiced their annoyance at what appears to have been the FEC banning a meme or removing what they deemed inappropriate posts from their Facebook site:

> 'Let the poor kiddos enjoy their memes or bad ratings will continue #threats'

A second, more serious, example saw an HEI's collective-judgement score fall from 4.45 to 3.99 over 60 days at the end of 2017, despite the HEI's collective-judgement score comprising over 175 reviews in this period. The fall was the result of over 20 negative Facebook reviews in the space of four days angrily reacting to the action taken by the provider following an alleged racist incident.

> 'Racism is impossible in the school! A [Nationality A] student was beaten by a [Nationality B] who was born in the UK, because [Nationality B] did not want a discussion with [Nationality A] and said no-one of [Nationality A] can stand in front of him! This school sent the email to suspend the victim!"
>
> 'We all know that school is the place for learning, not for international students being racist. Why [Nationality A] student was beaten and racist in [provider redacted].'

This surge in reviews leading to significant changes was not confined to a provider's collective-judgement score getting worse. In one instance, an FEC saw its collective-judgement score rise from 2.88 to 3.85 over a 60-day period due to a collection of eight, five-star reviews over two days, all from international students praising the college and its international office following a visit.

> 'Excellent College with Professional and Competent Staff. The international Office is very specialist and available with student. Very good experience. Thank you!!'

Not all the sudden shifts in collective-judgement scores were the result of specific issues. Amongst the significant short-term changes to providers' collective-judgement scores were

instances of broad quality matters that should be of concern to any provider or oversight body. One example saw an HEI's collective-judgement score fall from 4.00 to 2.90 over a period of 60 days in early 2016. In this case, the fall was the result of four, one-star reviews, three on Whatuni.com and one on Facebook, and no positive reviews in little over a month.

> 'My overall uni experience is an bad one I thought I would really enjoy uni and [provider redacted] really stood out and was bigged up to me but since being there, it's been awful! Not one good thing I could say about the place!'
>
> 'Terrible. Really, don't go to [provider redacted], don't do it! You are wasting so much money just for being in a nice building. 'nice' until you find that having class in an open space along with 2 other courses is a hell. No skilled people, no real teaching, low level of competence. Don't do that! Just go on the website and think why there are no details about the courses: it's because they don't have a structure! The only good thing is that you have quite lot of equipment that you can borrow for free (but I guess all the other uni have this). Really, I'm doing a master and the course are like 'introduction' to things, the level is like high school class. Really, don't do this.'
>
> '… Teachers are not skilled, and I wonder how the uni is allowed to provide MSc title. They shouldn't as I (undergrad) have more tech skills than the tutors. Really: don't do that. Unless you are a very beginner, do not expect any serious quality of things in that place. Your application will be accepted regardless of your background: the result is that no-one in my course have any experience in this filed other than me and I'm basically doing nothing while they discover the basics of the field….'

A second example occurred at an FEC which saw its collective-judgement score fall from 4.08 to 2.94 in the space of 60 days during the summer of 2016. In this instance, seven, one-star reviews and one, three-star review, were left in the space of a few months whilst the previous steady flow of positive comments stopped entirely. Those that were accompanied by text described concerns over the quality of provision, including:

> 'The last year I was there I loved it with great staff and fun in the courses but this year it's fallen because since staff have left the new staff just expect students to get on when they don't even know what they're doing, plus the courses mostly have exams now too, whereas before it was just done on work and talent.'
>
> 'Don't go to this college full stop, the way they treat students is a disgrace.'

In these instances, it appears genuine concerns have been highlighted and, as they are not being offset by positive reviews, they have the ability to add to quality monitoring efforts by both providers and oversight bodies. It should be remembered that, while several reviews contain concerns that either do not relate to the quality of provision or relate to a very specific issue, such reviews are not the norm and the 'wisdom of the crowd' effect enhanced by combining multiple sources of reviews will negate the effects of such comments. A discussion of the results of this study, what they might mean for the sector and future research questions are explored below.

Discussion

This research is a first examination of whether, as has been shown to be the case in healthcare, aggregated user feedback might be able to identify the quality of higher education provision. The association between the collective judgement of students and existing quality measures suggests that it can. Moreover, it can do so in a more timely and less burdensome way than existing measures. The use of collective judgement however is reliant on a sufficient volume of reviews that, at present, is generally only available for larger providers.

This initial research raises a multitude of questions and we discuss the most pressing ones below. This discussion is divided into three parts: the potential use of collective judgement as an oversight tool, the potential use of collective judgement as a quality improvement tool, and wider points including ethical concerns and future research questions.

Student feedback as an oversight tool

Our results suggest that unsolicited student reviews are advantageous for measuring public perception and response, and therefore may be of particular utility for regulatory bodies who currently rely on more traditional surveys and metrics, and also have a strong new steer to act as a student champion[10 11]. Whilst no one would suggest monitoring unsolicited student reviews should supersede other methods of oversight, such an approach offers a number of advantages over existing methods and may be a powerful new weapon in the regulatory armoury. These advantages include:

- year-round feedback providing insight on different aspects of provision as they happen, rather than a reflective judgement at the end of a module or academic year
- more diverse feedback not just originating from third-year undergraduates or recent leavers
- near real-time feedback allowing intelligence to be gathered and, where necessary, acted upon quicker than existing measures
- feedback is based on what students deem to be important to them, rather than on what the creator of surveys or evaluation forms would like to know about
- no additional burden is placed on students to generate the data (they are already doing so), and collection burden is a fraction of that of existing approaches
- providers do not act as a middle-man in the collection of the data, minimising opportunities to influence or sanitise the feedback
- the ability to explore sector-wide issues, such as feedback relating to free speech, contact hours, or vice-chancellor pay.

Further to these advantages, there is more that can be done with the data available. For example, time series can be developed to predict future movements accounting for seasonality and other factors, and the tool could focus on specific dimensions of quality such as teaching and learning or pastoral care. Alongside these current and future benefits there are challenges that must be considered when looking to make use of student reviews for oversight purposes. These challenges include:

- The focus and perceived reliability of the data. Understandably, many being judged by collective student feedback may be sceptical of its accuracy and value. Such scepticism may be heightened by considering individual comments, such as the dispute over memes discussed in 'Marked changes in the collective-judgement score', which do not give a rounded assessment of a student's higher education experience. Such concerns should be allayed by the association demonstrated between existing (if disputed) measures of quality - the APR, TEF and NSS - and

the collective-judgement score, and the sheer volume of reviews leading to a 'wisdom of the crowds' effect whereby what some may perceive as less insightful comments are balanced out by others and an accurate reflection of quality prevails.

- Goodhart's law states that 'any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes'[12]. To what extent might review sites and social media, which have hitherto received limited attention from providers with more pressing performance metrics to consider, become the focus of providers attention and weaken or break the association between collective-judgement scores and other quality measures? Providers may seek to influence scores either by entering their own reviews directly, or by seeking to influence students' feedback. Steps can be taken to limit such behaviour, including monitoring patterns in reviews to detect suspicious comments and the already planned step of adding additional data sources to 'drown out' the effects of manipulation, but it is unlikely the effects could be eliminated entirely. Similarly, students, aware of the impact of a negative feedback, may use the threat of a bad review as leverage in discussions with academics or the provider again likely weakening the relationship between collective judgement and other quality measures.
- At present, the utility of the tool is far greater for larger providers, mostly HEIs, as they have a larger of reviews. The majority of higher education provision takes place in HEIs, so most provision may still be covered, but the utility of the tool is currently limited at smaller providers, especially alternative providers.

Our results also suggest that monitoring student reviews may provide some benefit in spotting, and hopefully preventing, declining quality before the consequences become too great. The examples seen in 'Marked changes in the collective-judgement score', however, raise a number of points for consideration. First, there is a clear need for interpretation and contextualisation of any sudden shifts in the collective-judgement score before any action is taken. Nuances in the data and the phenomenon of students focusing their attention on specific issues such as racist incidents or a provider's decision not to divest from fossil fuels, means that any sharp changes in collective-judgement scores must be contextualised. Second, despite significant effort taken to remove Facebook reviews that are clearly not relevant to the quality of provision, agreement needs to be reached over what is not deemed to be relevant, and additional action taken to implement the identification and removal of such comments. Third, some reviews are not accompanied by text explaining the reason for the review score. In some instances, this makes it difficult to know whether the review relates to broad quality concerns or a specific issue, relevant or otherwise. Finally, as with day-to-day monitoring of the collective-judgement score, the impact of specific issues affecting scores is far more prevalent in, although by no means exclusive to, smaller providers simply due to the size of their student population, and hence their number of reviews.

This tool therefore offers the regulator an opportunity to broaden their current set of mechanisms capturing information on organisational performance and respond in a more timely manner. The benefits of aggregated student feedback however need not be limited to oversight bodies, there are significant opportunities for providers themselves to embrace the feedback as a quality improvement tool. Below we discuss the possible benefits offered to providers by student feedback and the conditions necessary for providers to capitalise on it.

Student feedback as a quality improvement tool

Metrics aimed at improving the quality of higher education have traditionally failed to capture the student experience in a contemporaneous way. The timely and reliable extraction of the student collective-judgement is an important method to facilitate quality improvement in higher education. In fact, our study reveals that students have a key role to play in helping to

reach an accurate assessment of the overall quality of the education provider and in identifying adverse events or failures within the delivery of their education. This work indicates that there is considerable scope for students to play an active part in ensuring that their experience is efficient and appropriate. Facilitating the analysis of student voice has important implications for the quality improvement in this domain. The use of student voice can promote organisational learning; identify system strengths and weaknesses and can be used to help prioritise or measure the efficacy of quality interventions within specific providers or industry wide.

Organisational learning

The collective judgement of students presents a unique opportunity for organisational learning to improve quality. From an organisational learning perspective, it is well understood that institutional effectiveness and adaptation to change relies on the analysis of appropriate data and that modern technologies enable providers to gain access to insights from data that were previously unachievable. As technologies continue to penetrate all facets of higher education, students are generating valuable information. For example, the information generated from students presents an important opportunity for the process of self-reflection and promotes speaking up about key system issues as well as sharing the lessons learned from failures or missteps in the delivery of provider services. Students are well situated to 'hold the mirror' for the organisation to reflect on its performance - the following example is a keen observation of these phenomena:

> 'Many problems to begin with, which got worse each semester. Some lecturers had disgraceful attitudes, were not helpful and just made me stressed. Two lecturers were nice but that was it unfortunately. I remember doing a presentation and some staff member coming in half way through, to argue with someone in the room. So I had to stop my presentation until they stopped, I wouldn't recommend studying IT there what so ever. I also brought up an issue once with a staff member in confidence only to find he announced it in class to everyone. But you can't do anything because the management stick up for everyone but students to cover themselves as always. You can get in there and sign up and they all seem nice at first, but weeks down the line it completely changes … I could go on for quite some time but [provider redacted] was absolutely rubbish for IT, and anyone looking to study the subject I would recommend you look elsewhere. If you do go here you will experience constant problems. It's not worth paying the tuition fees and getting in debt when the level of education they provide is terrible. They may get good Ofsted results, but If the assessors came unplanned it would be a different story.'

Identifying and synthesising the student experience presents a unique opportunity for the organisation to reflect on its performance. Reflection is a key organisational process, and has been found to be an essential feature of successful organisations across several domains[13]. Crucially, follow-up on collective feedback may be encouraged and disseminated across the stakeholders within and belonging to the provider in order to foster a positive culture of speaking up. Meaning that the direct and timely follow up to problems must be much more than a mechanistic technical exercise and organisational analysis of problems, it should adopt a systems way of thinking that takes into account the social, political and cultural considerations that underpin the critical feedback and be shared inter- and intra-organisationally to promote the best learning opportunities. This is to say that in order for the successful implementation of quality improvement based on the student feedback, it requires organisations with cultures that support student-centred data to provide quality improvement capacity and host a leadership that is receptive to technical expertise[14 15].

Peer comparison
Nationally established procedures for analysis of contemporaneous student feedback would allow providers to see, comment on and spread relevant lessons learned from the data at the organisation and specialty level[16]. In turn, this fosters the ability of regulators to create comparator groups across the industry to monitor progress and performance as well as promoting transparency amongst peers. In fact, in healthcare the promotion of transparent peer comparisons (for example, reputational effects) has revealed itself as a motivating factor that improves performance[17].

Furthermore, the use of peer comparisons can reveal the qualities that sustain well-performing organisations. For example, extracting information on the 'best performers' can be used to illustrate 'best practice' in response to organisational failures or missteps and be used as a guide for organisations working towards excellence.

Prioritise quality interventions
In other domains, the voice of the consumer is used to prioritise and measure quality interventions. For example, in the healthcare domain, patient complaints have been used to reliably identify problems such as medical errors, breach of clinical standards and poor communication. This information has been used to target interventions in the organisations such as improved handovers, system modifications and team communication[18 19]. Similarly in healthcare, patient complaints have been known to precede incidents and therefore can be regarded as an early warning system[20 21]. However, quality intervention is not a 'one-size-fits-all' approach and the availability of reliable and timely feedback from students can be used as an indicator of the most appropriate quality interventions that are needed within the specific provider. For example, the following excerpt from one student review illustrates an opportunity for improvement from the perspective of the student:

> 'I have been very disappointed with the courses that I've been enrolled on, each worse than the last, the distance learning is not well supported to managed with staff changes and sickness having a major effect on organisation and communication, something is going wrong at [redacted provider] all the tutors are leaving or on sick.'

Drawing on this information, the provider is equipped with enough information to tailor an intervention strategy that would be best directed at this specific problem.

Furthermore, providers may monitor the student feedback to measure the performance or efficacy of on-going or past quality interventions - for example, in the post-implementation phase. For instance, students' collective judgement can be used to monitor the impact or relevance of a change in management within a provider or the performance of a specific course/module modification. This scenario is exemplified in the following student comment:

> 'After [redacted organisational change] the [provider] has worsened in regards to management and lectures. The lack of information given is not acceptable and a lot of students on my course have struggled with the workload and lack of help. However things seem to have been done and changes made'

This type of comment provides an important opportunity for the organisation to measure the efficacy and appropriateness of the quality improvement actions they have taken and can be used to illustrate performance.

Wider considerations

We have seen that student feedback has the potential to support regulatory oversight and organisational improvement and the specific challenges associated with using the data in

this way, but there remains two wider issues for consideration.

First, some may question how reassuring it is that the collective judgement of students is positively associated with other quality measures? The APR, TEF and NSS processes are important to providers and the focus of significant attention, but they are not without their critics. For example, the TEF measures educational outcomes, but it does not directly measure teaching excellence. Is there a danger, therefore, that a new measure with similar - although far from identical - findings is simply replicating the same issues present in existing metrics?

Second, and salient given the ongoing concerns over the use of Facebook data by Cambridge Analytica and others, are the privacy and ethical concerns that arise from monitoring student feedback[23]. For this research we have only used reviews that are publicly available. Specifically, for the Facebook data, we have only used reviews posted publicly on providers' review pages. We have not examined general posts made by students, even those posted on any other part of a provider's page, nor have we gathered any information about the users. At the very least, future research and any operationalisation of student feedback, must maintain strict ethical and privacy rules in order to be accepted.

Conclusion and future research questions

This research has explored whether the use of online student reviews can provide insight into the quality of higher education provision. After considering over 200,000 reviews and calculating a collective-judgement score from them, we found a positive association between the collective judgement of students and APR, TEF and NSS outcomes. The 'wisdom of crowds' phenomenon, which premises that when our often-flawed individual assessments are aggregated in the right way, our collective assessment is often highly accurate, seems to hold true in this instance. As a result, the use of the collective judgement of students could have significant benefits for regulators, providers and students.

A key requirement that is generated by this study is the necessity for assessing student, staff and organisational leaders' willingness to accept student feedback data, of the kind used here, as a valuable and helpful tool for improving quality, the student experience and organisational learning. Moreover, given such willingness (and its undoubted variability and distribution in the sector), what advisory and consultancy programmes could be designed for providers and others that would ensure more effective organisational attention to student feedback data and how it is best used for improvement in meeting both internal and external objectives?

This study represents only an introduction to the topic and a beginning of research into the use of the collective judgement of students. There is far more to be explored including, but not limited to:

- assessing whether students and providers accept the monitoring of student feedback as an oversight and or quality improvement tool
- considering whether more can be done to encourage student feedback in underrepresented areas, most notably FECs and APs
- automatically distinguishing higher education feedback from further education feedback in for FECs and APs
- examining Twitter and other sources of data
- considering the ability of student feedback to identify department or course-level quality and/or the different dimensions of quality
- sector-wide thematic analysis of student feedback to identify the overarching issues that are exercising students.

The good news is that as these questions are explored, the volume of student feedback will continue to grow, and, the more it does, the more accurate the algorithms developed to classify the data and derive meaning from it will become.

Acknowledgements

We are extremely grateful to staff at the regulatory and other agencies for the provision of additional useful information, and for the data providers, especially StudentCrowd.com, for their support during the research.

Alex Griffiths, Meghan Leaver and Roger King
June 2018

References

1. Griffiths A and Leaver M P (2018) Wisdom of patients: predicting the quality of care using aggregated patient feedback, *BMJ Quality & Safety*, 27(2), pp 110-18

2. Surowiecki J (2004) *The wisdom of crowds*, London: Little, Brown

3. Parry G, Callender C, Temple P, et al (2012) *Understanding higher education in further education colleges*

4. HEFCE (2016) *Revised operating model for quality assessment*, available at: www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/2016/201603/HEFCE2016_03.pdf

5. HEFCE (2016) *Annual Provider Review Guidance October 2016/29*, available at: www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/2016/201629/HEFCE2016_29.pdf

6. OFS (2018) *What is the TEF?* Available at: www.officeforstudents.org.uk/advice-and-guidance/teaching/what-is-the-tef/

7. NSS (2018) *The National Student Survey 2018*, available at: www.thestudentsurvey.com/about.php

8. HEFCE (2017) *National Student Survey 2017*, available at: www.hefce.ac.uk/lt/nss/faq/#Q

9. BIS (2016) *Specific course designation for alternative higher education providers. Guidance for providers: Criteria and Conditions July 2016*, available at: www.gov.uk/government/uploads/system/uploads/attachment_data/file/631131/withdrawn_specific-course-designation-2016.pdf

10. Gyimah S (2018) *Strategic Guidance to the Office for Students - Priorities for Financial Year 2018/19 - Letter from Sam Gyimah, Minister of State for Universities, Science, Research and Innovation, to Sir Michael Barber, Chair of the Office for Students 2018*, available at: www.officeforstudents.org.uk/media/1111/strategicguidancetotheofs.pdf

11. *Higher Education and Research Act 2017*

12. Goodhart C A E (1984) *Monetary Theory and Practice: The UK Experience*, Macmillan Publishers Limited

13. Reynolds M (2017) *Organizing reflection*, Routledge

14. Davies E A, Meterko M M, Charns M P, et al (2011) Factors affecting the use of patient survey data for quality improvement in the Veterans Health Administration, *BMC health services research*;11(1), pp 334

15. Hildebrand T, Lane HW, Research UoWOWBS, et al (1991) *Organization learning: theory to practice*, London: Research and Publications, Western Business School, University of Western Ontario

16. Pučėtaitė R, Novelskaitė A, Lämsä A-M, et al (2016) The relationship between ethical organisational culture and organisational innovativeness: Comparison of findings from Finland and Lithuania, *Journal of business ethics*;139(4), pp 685-700

17. Bevan G and Evans A (2018) Reputations count: why benchmarking performance is improving health care across the world, *Health Economics, Policy and Law*

18. Levtzion-Korach O, Frankel A, Alcalai H, et al (2010) Integrating incident data from five reporting systems to assess patient safety: making sense of the elephant, *Joint Commission Journal on Quality and Patient Safety*; 36(9), AP1-AP18

19. Weingart S N, Pagovich O, Sands D Z, et al (2005) What can hospitalized patients tell us about adverse events? Learning from patient-reported incidents, *Journal of General Internal Medicine*; 20(9), pp 830-36

20. Reader T W, Gillespie A and Roberts J (2014) Patient complaints in healthcare systems: a systematic review and coding taxonomy, *BMJ Quality & Safety*, 23(8), pp 678-89

21. Kroening H L, Kerr B, Bruce J, et al (2015) Patient complaints as predictors of patient safety incidents, *Patient Experience Journal*, 2(1), pp 94-101

22. Griffiths A (2016) *Forecasting Failure: Assessing Risks to Quality Assurance in Higher Education Using Machine Learning*, King's College London

23. Cadwalladr C and Graham-Harrison E (2018) Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach, *The Observer*, Sat 17 March 2018

All links last accessed 21 June 2018

QAA2151 - June 18

Southgate House, Southgate Street, Gloucester GL1 1UB
Registered charity numbers 1062746 and SC037786

Tel: 01452 557000
Web: www.qaa.ac.uk