Manchester Metropolitan University

National student survey metrics: where is the room for improvement?

A. Mark Langan [1] (iD) · W. Edwin Harris [1]

Abstract

The purpose of this study is to use machine learning and exploratory data analysis to interrogate patterns of metrics from a national-level student survey. Analysis of over 1.8 million returns detected long-term stability of the predictors of student satisfaction, with survey items relating to course management and teaching being consistently most influential. All metric outputs increased over the survey period; however, the rates of increase of several dimensions including Overall Satisfaction decreased markedly in the most recent years to a point of levelling off. There was also a growing similarity in an institution of outcomes at a national level. This study contributes new insights into the influential survey instrument, through rigorous determination of the most influential survey items, descriptions of the changes in variability between institutions, and exploration of the importance of patterns of outliers at the extremes of the metric outputs. We also identify a rapidly growing spike in total satisfaction at a broad course level and highlight how this is inconsistent with a customer satisfaction model. We conclude by considering the challenges of the use of national-level student surveys for the management of student satisfaction in higher education.

Keywords Student satisfaction · Output metrics · Customer · National Student Survey

Introduction

The sustained, growing impetus worldwide to measure performance in higher education (HE) using comparative metrics has been driven by economic pressures and consumer agendas (Hazelkorn 2015). Changes in funding, student numbers, costs of study and political drivers have resulted in a re-prioritisation of HE practices and adoption of business values in the sector

✉ A. Mark Langan
  m.langan@mmu.ac.uk

  W. Edwin Harris
  e.harris@mmu.ac.uk

[1] School of Science and Environment, Manchester Metropolitan University, Chester Street, Manchester M1 5GD, UK

(Birnbaum 2000). This has led to increased public accountability and intensified the emphasis placed on measures of institutional performance, founded on consideration of students as customers, or consumers, of education (see Molesworth et al. 2009). These factors are thought to have contributed to an international trend in using student satisfaction ratings as surrogate measures of educational quality, despite considerable debate surrounding their value and validity in comparing higher education institutions (Ball 2017; Hazelkorn 2015).

National-level student survey instruments propagated rapidly, relying on self-reporting by participants (Yorke 2015). This trend has given a powerful collective voice to students, whilst avoiding costs and heterogeneity associated with alternative observational methods (Richardson 2005). It has been argued that the measurement of student satisfaction could be valuable to higher education institutions as a tool to inform quality enhancement (O'Neill and Palmer 2004). However, evidence suggests that surveys of selected aspects of the student experience have had limited impact in improving levels of satisfaction at a large scale (Shah 2012).

There have been several approaches to systematically measure student perceptions of the university experience at a national level. For example, the US National Survey of Student Engagement (NSSE) requires respondents to reflect on their level of engagement with their educational experience (Kuh 2003). Behaviourist approaches like this quantify students' perceptions of their own activities associated with engagement with a variety of aspects of higher education (Kahu 2013). Within this framework, student engagement may be considered to be the time and effort students devote to 'educationally purposeful activities' (Radloff and Coates 2010). This approach avoids positioning students as consumers merely rating a product (Hazelkorn 2015), and there is a wealth of literature to support their use (e.g. Kuh 2009). There are critics of applications of behaviourist approaches in general (Zepke 2014) and this type of survey can lead to potential confusion and disagreement by both staff and students (Payne et al. 2005) and, in a wider sense, the use of educational learning outcomes in general (Murtonen et al. 2017).

The Australian Course Experience Questionnaire (CEQ) was an early example of a survey designed to capture student perceptions of course quality at a national level. The instrument was founded on the tenet of a strong association between the quality of student learning and student perceptions of teaching quality (Ramsden 1991). However, there is equivocal evidence about how well learners are able to evaluate the quality of their courses (e.g. Marsh 2007; Clayson 2009). Wilson et al. (1997) comment that the CEQ was 'not designed or intended to provide specific or fine-grained diagnostic feedback regarding individual subjects or teachers'. Despite some evidence of a relationship between measures of satisfaction of learners and their learning successes (e.g. Richardson 2005), there remains no clear consensus on a definition of student satisfaction or an accepted measure for satisfaction with higher education experiences (see García-Aracil 2008).

The UK adopted the approach of the CEQ in the design of its National Student Survey (NSS) instrument. The survey is purported to have a robust design (Cheng and Marsh 2010), in that it measures what it claims to measure (based on recommendations in Ramsden 1991). However, 'validity' in the context of survey design is a multivalent concept and there have been challenges to the effectiveness of the questionnaire as a means of assessing quality in higher education. For example, reducing the student experience into a few composite metrics for comparative purposes has been widely criticised (e.g. Elliott and Shin 2002). Despite limitations inherent in this type of approach, student self-reported metrics are widespread and influential in higher education and there is significant interest in the relationship between ratings of individual aspects of the student experience and holistic 'overall satisfaction' metrics (Langan et al. 2013; Burgess et al. 2018).

There is strong criticism of viewing students as consumers or customers to satisfy, rather than as active participants of education (Hazelkorn 2007). The 'consumer ethos' has been associated with survey instruments like the NSS, and other perceived weaknesses of the survey instrument include differences in the appropriateness and interpretation of survey items for different subject areas and learner typologies (Yorke et al. 2014); lack of ability to discriminate between universities (Cheng and Marsh 2010); acquiescence bias and 'yea saying' (Yorke 2009); overemphasis of recent experiences (Douglas et al. 2006); effects of factors not associated with the quality of the experience (Bennett and Kane 2014); lack of representation of the 'whole experience' (Eley 2001); metric differences unrelated to quality of educational provision (Langan et al. 2013); and 'gaming' survey instruments to influence the outcomes (Yorke 2013; Grove 2013). Despite the broad literature highlighting deficiencies in survey assessments of higher education quality, they are widely used internationally.

The UK's NSS was designed to acquire feedback from undergraduate students in their third year of study in order to inform choices of future students and contribute to public accountability of HE institutions (see Surridge 2008). The output metrics are available publicly at a level pertaining to a broad 'subject' level (termed 'JACS3' level). The survey is now firmly embedded into the national psyche for UK HE, and outcomes exert significant pressure on institutions to improve their metrics (Locke 2014). Its importance has been heightened by the inclusion of selected elements in the creation of league tables (Hazelkorn 2015) and also its presence in the UK's Teaching Excellence Framework (TEF) which is used to classify HEIs in terms of quality and educational provision (Gunn 2018). The importance of institutional reputation, coupled with significant 'top down' messages to enhance institutional metrics, places value on a long-term, national overview of the survey instrument.

After 10 years of usage, the NSS was modified in 2017. The review identified perceptions of conceptual weaknesses concerning what it measured and methodological weaknesses related to what it covered (Callender et al. 2014). This highlighted the instrument's narrow scope in terms of capturing the student experience, including their engagement in learning and teaching activities. Subsequently, new survey dimensions were added; however, many of the original survey items remained similar or unchanged. The modification provides an opportunity for reflection on the survey outcomes over the original instrument's lifetime. Recently, Burgess et al. (2018) explored factors that shaped the first 10 years of NSS metric outcomes. Their study describes the increase in overall metric outcomes and observes that survey dimensions of Organisation & Management and Teaching were the strongest predictors of student Overall Satisfaction (in agreement with Langan et al. 2013). The study did not explore the specific survey items that predicted student satisfaction over time, but did map institutional and subject differences as well as highlighting large-scale geographical differences in outcomes. They went on to suggest that the survey has helped prospective students to enrol on suitable courses and found no evidence that the introduction of higher fees had any influence on survey output metrics.

Here, we use machine learning to rank the predictive power of the specific survey items and to evaluate how their influences have changed over time. We will also describe trends of metrics over a decade of usage and highlight the capability of outcomes that has been achieved by the highest-performing universities, suggesting areas where there is room for improvement elsewhere. This will simultaneously identify if institutions are becoming more similar or different in their outcomes and whether this is related to the year of the survey. Finally, we compare survey results with a conceptual model of customer satisfaction to investigate if the outcomes are comparable. Specifically, our main objectives are to (1) describe long-term

patterns and distributions of the survey outcomes and (2) identify how the relative importance of survey items in predicting overall satisfaction varies over time. We discuss our results in the context of the considerable challenge of using student survey outcomes for informing the management of student satisfaction in higher education.

Methods

The UK's National Student Survey (NSS)

The NSS database represents over 2.5 million respondents for the ten-year lifetime of the first version of the survey instrument (Burgess et al. 2018). The survey instrument comprised six dimensions: Teaching; Assessment & Feedback; Academic Support; Organisation & Management; Learning Resources; and Personal Development (Table 1). Survey participants provide ratings of agreement for positively worded statements about university experiences (1–5; strong disagreement to strong agreement). We adhere to several conventions for the UK NSS established in the literature (e.g. Fielding et al. 2010). Briefly, survey items are commonly reported as suffixed by 'Q' (e.g. 'Q1' for survey item 1), implying that they are questions rather than statements. There is an option of neutrality in the response scale (a rating of 3), meaning that an expression of 'dissatisfaction' (ratings of 1 or 2) is not simply an inverse of 'satisfaction' (ratings of 4 or 5). Item Q22 in the original survey enquired directly about respondent 'Overall Satisfaction' with their course, and this holistic item may be considered to be a survey dimension on its own. Reporting of results for the NSS has most often referred to students being 'satisfied' when the Q22 response is 4 or 5 and is often summarised as a percentage of such responses. It is noteworthy that the term 'satisfaction' is often applied to other survey dimensions as well. For example, students may be considered to be 'satisfied' with teaching in reference to survey items in the teaching dimension (Q1–Q4). However, this usage is not an exact representation of the wording of the items, which are ratings of the perceived experience and not an explicit expression of respondent satisfaction (with the exception of Q22).

Data extraction and analysis

Nationally available data for the NSS (2006–2015) were extracted for all subjects and institutions (http://www.hefce.ac.uk/learning/nss/data/). It is notable that text comments are captured by the survey, but provided (anonymously) only to the originating institutions making them difficult to analyse at a national scale (Langan et al. 2017). This is despite increasingly sophisticated semi-automated computational techniques that explore large textual datasets (Neuendorf 2016). We selected institutions with more than 500 respondents for every survey in the period for analysis. A small number of institutions were excluded if they changed their affiliation or name in this period to avoid the possibility that new incarnations were not contiguous comparisons. This resulted in a set of 154 institutions and 1,841,711 returns across 10 years.

Machine learning analysis

We used the machine learning technique 'Random Forest' (Breiman 2001) to identify the importance for specific survey instrument questions (Q1 through Q21) in predicting student

Table 1 Importance values to show relative rankings of survey items in predicting Q22 Overall Satisfaction

Item number	Survey dimensions and item wording	Relative importance
	Teaching	*36.06*
Q1	Staff are good at explaining things	62.51
Q2	Staff have made the subject interesting	14.85
Q3	Staff are enthusiastic about what they are teaching	11.59
Q4	The course is intellectually stimulating	55.28
	Assessment & Feedback	*1.83*
Q5	The criteria used in marking have been made clear in advance	1.38
Q6	Assessment arrangements and marking have been fair	5.66
Q7	Feedback on my work has been prompt	0.93
Q8	I have received detailed comments on my work	0.48
Q9	Feedback on my work has helped me clarify things I did not understand	0.71
	Academic Support	*13.07*
Q10	I have received sufficient advice and support with my studies	23.42
Q11	I have been able to contact staff when I needed to	8.55
Q12	Good advice was available when I needed to make study choices	7.24
	Organisation & Management	*38.75*
Q13	The timetable works effectively as far as my activities are concerned	1.89
Q14	Any changes in the course or teaching have been communicated effectively	15.09
Q15	The course is well organised and is running smoothly	99.27
	Learning Resources	*0.65*
Q16	The library resources and services are good enough for my needs	0.58
Q17	I have been able to access general IT resources when I needed to	0.38
Q18	I have been able to access specialised equipment and facilities when I needed to	1.00
	Personal Development	*8.38*
Q19	The course has helped me to present myself with confidence	10.76
Q20	My communication skills have improved	3.39
Q21	As a result of the course, I feel confident in tackling unfamiliar problems	10.99

The survey dimensions of 'Organisation & Management' and 'Teaching' were most influential, followed by 'Academic Support', each with one or two particularly important survey items. The dimensions of 'Assessment & Feedback' and 'Learning Resources' were considered extremely weak predictors of the satisfaction metric

'Overall Satisfaction' (Q22). Machine learning methods are increasingly being used to analyse complex, large datasets in diverse fields of study, including psychology (Strobl et al. 2009), ecology (Prasad et al. 2006) and higher education research (Langan et al. 2016; Hardman et al. 2013). Random Forest analysis ultimately makes predictions based on variable associations, ranking the predictive importance of each variable (Grömping 2009; Genuer et al. 2010). Random Forest classifiers performed well in comparisons made by Lee et al. (2005), and the algorithm is considered robust with data distributions that do not conform to the same probability distribution restrictions (e.g. normality) that apply to techniques such as the least squares regression analysis and other generalised linear models (Breiman 2001; Strobl et al. 2009).

The approach uses resampling of variables and cases to fit models with many explanatory variables and individual records. In other words, it runs iterations with each of the model omitting a sample of both the predictors (i.e. survey items) and survey outcomes (i.e. particular participant responses). Therefore, each iteration can be viewed as a unique hierarchical tree. The final outcome of the analysis ranks the predictors hierarchically by considering all the trees in the 'forest' and attributes the level of variance that each factor (in this case, each survey item) explains.

The model used included both 'subject' (at JACS3 level) and 'year' (i.e. when the annual survey was completed) as categorical predictors. Their inclusion incorporates these factors into the analysis, for example considering the dataset as many years of entries, rather than a single block of data for the whole survey period (and similarly for the subject areas). This also allows the relative influence of both year and subject to be investigated, in combination with the survey items. It is valuable to know if these factors outperformed the survey items, as it would mean subjects or the year of survey was more important predictors of student satisfaction than the survey items. See Fielding et al. (2010) for outcomes using models that included different factor combinations (including year and subject).

We quantify the relative importance of each predictor in the model (i.e the survey items) as the equally weighted mean percentage of increase in MSE (mean squared error) explained and also, its Gini score (the reduction in accuracy of models excluding the target variable). Variables ranking high in importance indicate they are more influential in predicting the dependent variable (see Genuer et al. 2010), which in this case is the satisfaction value reported by Q22. Analyses were conducted using 'random forests' in the statistical package 'R' (R Core Team 2018), following the methods outlined by Liaw and Wiener (2002).

Results

Descriptive results

Overall, we found that institution-level data shows an increase in levels of agreement with all survey dimensions from 2006 to 2015 (Fig. 1) and survey items that started with lowest values gained most in the 10-year period. Overall Satisfaction (Q22) recorded the smallest gain (5.5%), moving from 80.8 to 86.3%. Overall Satisfaction and Teaching had the highest initial levels of agreement (within 0.2% of each other), and both dimensions increased only comparatively slowly and exhibited little or no increase in the most recent years (with the Teaching median moving from 81.0 to 87.3% over the decade). These dimensions appear to exhibit a 'ceiling effect' just below the 90% median agreement level by the end of the survey period. The median level of agreement for Assessment & Feedback was lowest in magnitude at the start of the survey, but showed one of the greatest dimensional increases in median agreement over the survey lifetime (2006, 62.1%; to 2015, 73.4%). In this period, median values of the dimensions of Assessment & Feedback, Support and Organisation & Management increased by similar amounts, 11.3%, 11.8% and 9.8%, respectively.

There was wide variability in the levels of agreement for institutions within all survey dimensions (Fig. 1). Particularly, high- or low-institutional performance is highlighted by outliers (values more than 1.5 times the interquartile range) and these illustrate the full range of outcomes. We note that there are no upper outliers for the Teaching, Organisation & Management or Overall Satisfaction dimensions, whilst there are lower outliers for all dimensions. The Assessment & Feedback dimension was also interesting as it not only exhibited strong, monotonic improvement over the study period but also supported the largest amount of upper outliers, highlighting the gap in performance for some institutions.

Organisation & Management showed the largest range in agreement across all years (from 36.7 to 92.9%), accompanied by the largest interquartile range (indicated by the height of the boxes in Fig. 1). Teaching exhibited comparatively minimal within-year variability (range from 64.5 to 94.8%) and once again, its pattern of change over time was very similar to that of
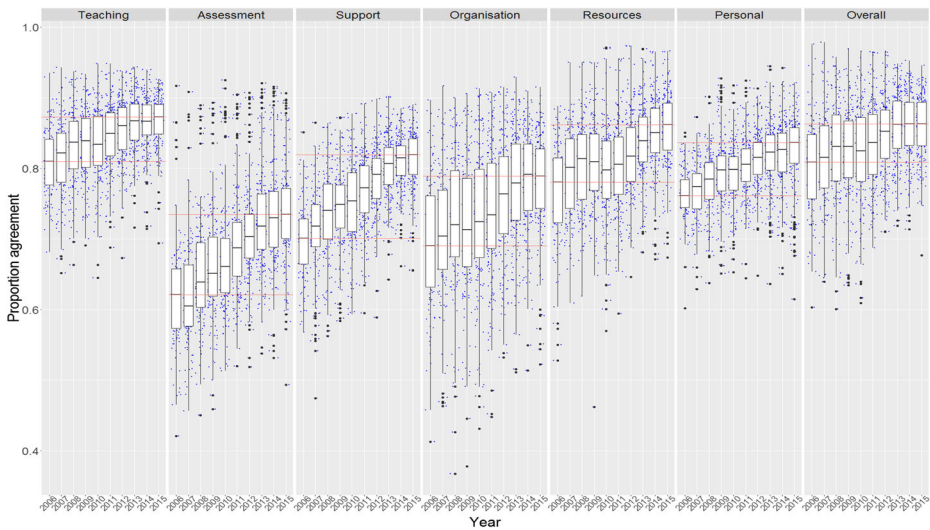
Fig. 1 Trends in 'Overall Satisfaction' (Q22) metrics for agreement (scores of 4 and 5) for institutions that returned over 500 students annually. The survey 'dimensions' of the NSS are displayed as an average of the original survey items, for example 'Teaching' is represented by four survey items. Red lines indicate the first year and last year median values during the survey lifetime, boxes contain 50% of the data, whiskers 90% (with small dots for institutions in the whiskers) and outliers (values more than 1.5 times the interquartile range) are shown as larger dots

the Overall Satisfaction metric. Overall Satisfaction, Teaching and Support all exhibited gradual reductions in variation, indicating greater consistency of institutional outcomes. Other dimensions, notably Assessment & Feedback and Organisation & Management, showed much larger within-year variation. Personal Support had the smallest interquartile range across years and was the only dimension with a trend of increasing institutional variability across the survey lifetime. All the other dimensions showed a trend of becoming less variable over the survey's lifetime (arguably the Assessment & Feedback dimension remained stable in this context). These patterns demonstrate the consistent rise in metric outcomes during the survey period, with a tendency for high-performing metrics to flatten (arguably reaching ceilings) and in most cases, a reduction in the variability between institutions. Overall Satisfaction essentially mimicked the outcomes of the Teaching dimension but with slightly higher variability in responses.

Predictors of satisfaction

A stand-out result is our finding of the domination in importance of the survey item 'The course is well organised and is running smoothly' (Q15) in the holistic judgement of satisfaction (Table 1; Fig. 2). This is followed in emphasis by ratings of staff being good at explaining things (Q1) and the course being intellectually stimulating (Q4). The predictors show a reasonable stability over time for many survey items, (Fig. 2), particularly Q15 as the survey item of highest importance in predicting Overall Satisfaction. However, Q4 jumped over 50% in importance for several years in the middle of the NSS, whilst Q1 showed year-to-year variation of about 10% importance. Other survey items exhibited relatively low importance in predicting Q22. Arguably, Q10, a survey item from the Support dimension and another Teaching item (Q2) had influences on the student satisfaction rating to a level that
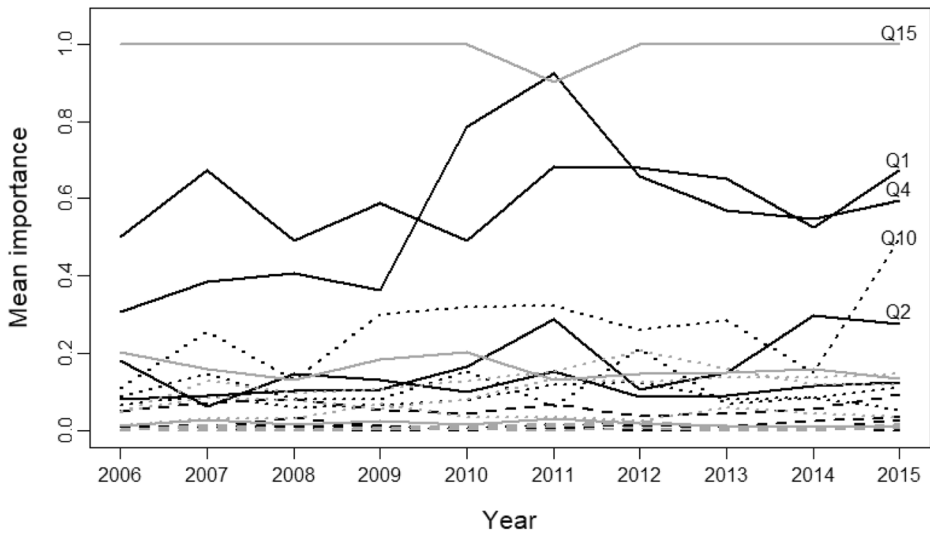
Fig. 2 Trends in the effectiveness of the NSS questionnaire items (Q1–Q21) as predictors of Overall Satisfaction (Q22). Data are shown for all subjects and institutions combined. >>Description here of the line types<< Individual predictors tended to be relatively stable in importance through the survey duration

set them apart from all the other survey items (which can be considered to be little more than noise in the model).

Satisfaction metric distributions

Data distributions of the Overall Satisfaction metric (Q22) provide another view of NSS long-term outcomes. Figure 3 shows the distribution of data grouped at the institution level, representing a relatively coarse view of satisfaction from many courses averaged across
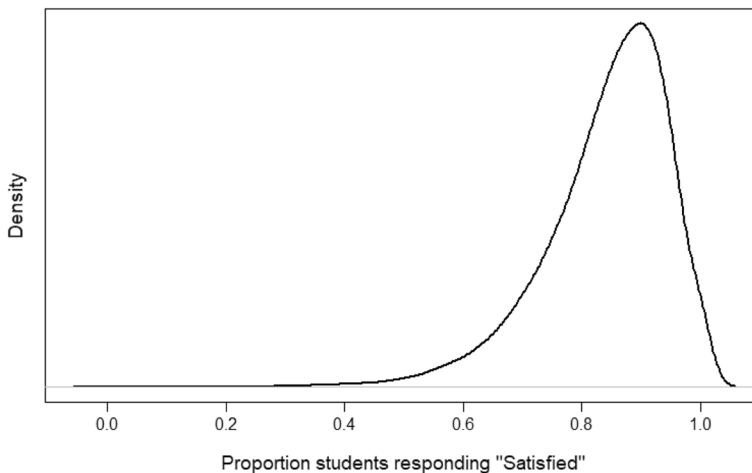

Fig. 3 Satisfaction metric data distributions for institutional level Q22 'Overall Satisfaction' metrics (ratings of 4 or 5 only). Survey outcomes for 154 institutions from 2006 to 2015 are shown. The right-skewed distribution follows that of the expected distribution of satisfaction from self-reporting customer surveys (Peterson and Wilson 1992)

individual institutions. This clearly shows the distribution is negatively skewed with median response values tending towards satisfaction. This pattern has been documented for customer self-reporting satisfaction surveys, for example as used in retail and hospitality arenas (Peterson and Wilson 1992). It would be highly unlikely that a whole institution would receive unanimous satisfaction for all courses, particularly in our group of larger institutions returning over 500 participants annually. As expected, we observed no institution in any year of the survey with a unanimous satisfaction rating (i.e. 100%). It is worth noting that we emphasise the frequency of agreement here, and, whilst low frequencies of agreement represent the absence of satisfaction, they also represent both dissatisfaction and neutrality.

Figure 4 shows the distribution of data grouped at the individual 'subject' level, representing a relatively fine view of satisfaction. Our broad subject level grouping follows the JACS3 convention of grouping courses that have similar content, for example called 'Subjects Allied to Medicine'. Whilst this fine view resembles the conceptual distribution for customer satisfaction predicted by Peterson and Wilson (1992); mirrored by the institutional distribution in Fig. 3), there is an obvious difference due to the large spike in 100% satisfaction results. We note that this prominence of unanimous agreement at subject level results tends to increase during the lifetime of the NSS. This is particularly notable in the final survey (2015/2016), when for the first time, 100% satisfaction represented the most common outcome at the course level.

Discussion

Our analysis extends other explorations of the UK NSS dataset in several ways (e.g. Burgess et al. 2018; Fielding et al. 2010; Langan et al. 2013). We highlight the importance of student perceptions of course organisation, academic staff explanations and the perceived intellectual
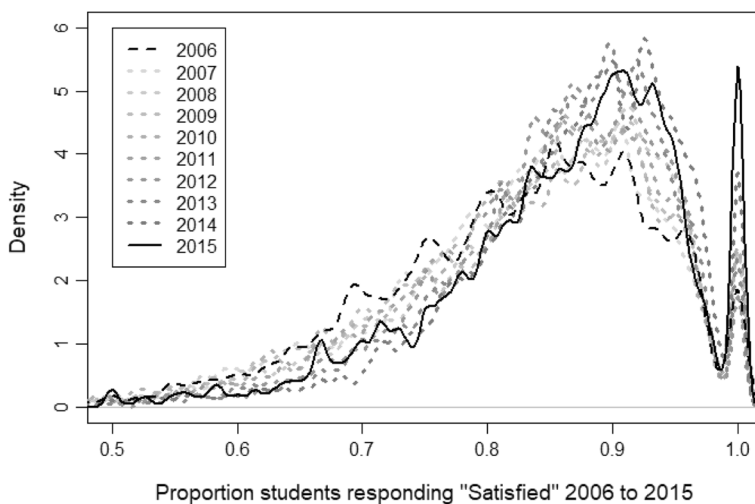

Fig. 4 Satisfaction metric distribution for JACS3 level Q22 (Overall Satisfaction; ratings of 4 or 5) showing median outcomes from 2006 to 2015. The right-skewed distribution shows consensus with self-reporting customer satisfaction surveys in a wide range of applications and industries (see Peterson and Wilson 1992). Between 2006 and 2015, the distribution of the proportion of 'Satisfied' responses tended to become more right skewed, whilst the overall proportion of 100% satisfied responses greatly increased in the final year

stimulation as the areas in strongest association with Overall Satisfaction. We clearly show that very few other NSS survey items have any notable value in the prediction of student satisfaction. We also show evidence of a ceiling effect for some dimensions including Overall Satisfaction and Teaching, which showed relatively high levels of agreement and relatively little in improvement for the duration of the NSS. We highlight the value in identifying outlier results to target areas for managing improvement at individual institutions. Finally, we present evidence that, overall, institution-level student satisfaction in the NSS is consistent with the predicted distribution arising from the consumer model for product satisfaction. However, we point out a striking deviation from this at a broad 'course' level, which had far more unanimous satisfaction results than expected for customer satisfaction models. A trend that intensified in the final year of the survey instrument to the extent that the most common outcome was total satisfaction.

Contextualising metric outcomes

There has been a significant shift towards incorporating NSS metrics into UK Higher Education thinking and schema (Sharpe 2019). Increases in the NSS metrics, particularly those that started with low-entry values, are sometimes interpreted as an indication that institutions have responded to survey outcomes and improved the student experience (Burgess et al. 2018). Survey dimensions that started with the highest ratings, such as Overall Satisfaction and Teaching, went on to exhibit signs of ceiling effects at levels of around 90% agreement. Many universities may now be working towards metric targets (for example levels of 90%+ agreement) that we observe have been extremely rare in the lifetime of this survey instrument. At the same time, institutions have become more similar in their survey outcomes. This suggests that increasing institutional inputs would be needed to make further metric gains in these upper quartiles, due to the difficulty of 'pleasing almost all of the people, almost all of the time' at the institutional level.

Ceiling effects and growing similarity in institutional outcomes reduce the differentiating value of the survey as a surrogate measure of institutional quality. We agree with Cheng and Marsh (2010) that the ability to use the NSS to measure quality in HE at the institutional level is limited by survey instrument sensitivity. They argue that this is the product of variation between (rather than within) institutions, making differences between institutions small in magnitude and difficult to detect, other than at the extremes. This lack of discriminatory power, or 'spurious precision', is known on a global scale (Soh 2017) and is reflected in the narrative around NSS outputs being poorly designed for ranked outputs (Surridge 2008). League tables, such as 'Good University' rankings created by third parties (e.g. newspapers), compile metrics in many different ways (Dill and Soo 2005). As long as stakeholders continue to compare institutional output metrics and respond to university rankings that incorporate NSS outcomes, ceiling effects and bunching of metric outcomes will likely serve to intensify competition between higher education providers if they manage quality improvement targets based on ranked metric outcomes (see Rienties 2014).

We found that metrics relating to aspects of course organisation and teaching were the main predictors of Overall Satisfaction metric (Q22), which is consistent with previous studies (Burgess et al., 2018; Langan et al. 2013; Fielding et al., 2010). Variation in the Teaching dimension closely tracked Overall Satisfaction, both in its trends and absolute values. Survey item Q15 (Organisation & Management; 'The course is well organised and is running smoothly') had the strongest association with the Overall Satisfaction for every year of the

survey. The overarching value placed on courses running smoothly could be viewed as reflecting a more 'customer-focused' viewpoint of students. Arguably, the dominance of a course organisation metric in shaping the holistic view of the overall satisfaction metric suggests that students valuing organisational components of education over aspects of the educational experience associated with learning gains. This is highlighted by items associated with Assessment & Feedback being amongst the poorest predictors of Overall Satisfaction throughout the decade under scrutiny.

This feeds into the debate about what learners 'want and need'. This is epitomised by an interesting small study by Boehler et al. (2006) who found that medical students perceived 'feedback' on assessments as more satisfactory when they were praised. This was in contrast to when they received unflattering, constructive feedback that led to better performance in future assessments which was perceived as less satisfactory. This highlights how interpretations of survey outcomes inform quality enhancement should consider relationships between what students perceive the want and which they rate highly, compared with what is needed to drive educational gains but which may not be as popular (see Sharpe 2019). It is well documented that, in a general sense, students 'liking' their tutors tend to rate the quality of the teaching experience more highly regardless of subsequent learning gain (e.g. Nisbett and Wilson 1977). Our findings suggest that the experience of HE in general is influenced by 'in-class' experiences as three of the four items (two in particular) in the Teaching dimension had predictive capacity that separated them from the bulk of survey items that had little predictive capacity. Langan et al. (2017) showed that this area of the student experience also attracts most free-text comments. This was to the extent that a high proportion of positive comments about teaching and a limited proportion of negative comments about course organisation was a profile associated with high levels of overall satisfaction.

Our findings suggest that the biggest gains in quality enhancement in higher education at a national level might be to position focus on only lower outliers in survey items. For example, Academic Support and Organisation & Management have potential for enhancement nationally, both exhibiting a significant number of institutions that have underperformed in the sector. Whilst targeting improvements in course organisation may be a way for many institutions to enhance Overall Satisfaction metrics, it is not a measure that is utilised in the UK's Teaching Excellence Framework that classifies institutions into Bronze, Silver and Gold categories (Gunn 2018). Likewise, high outliers may indicate areas of best practice, which may be underutilised at the national level to inform quality improvement in the context of a somewhat adversarial ranking system.

Challenges of responding to student surveys

Student surveys are of course unable to capture every aspect of satisfaction of a course experience (Wilson et al. 1997). Burgess et al. (2018) noted that there is a great deal of variation in Overall Satisfaction that is not explained by core survey items. Our results show that the holistic Overall Satisfaction metric outperforms all other dimensions in the survey, also suggesting that other, unmeasured factors are at play when students rate their satisfaction. These results are consistent with a large number of studies that suggest a complicated range of inter-related factors influence student satisfaction metrics (e.g. Bell and Brooks 2018; Douglas et al. 2015; Lenton 2015; Letcher and Neves 2010; Shah et al. 2017).

We acknowledge that there are many complex influences on survey output metrics that are needed to contextualise ratings of student satisfaction. Individuals are known to vary in how

they interact with the survey instruments for many reasons, such as their course of study (Yorke 2013) and demographic or personal attributes (Dean and Gibbs 2015). There is also evidence that certain typologies of institution perform unequally with respect to NSS outcomes. Universities with higher-research rankings, such as the 'Russell Group' (together with an old classification called '1994 Group') in the UK, tend to outperform others in the relative increase in Overall Satisfaction when compared with the value of the constituent survey items (Langan et al. 2013). There is a regional variation, such as differences between responses in the different countries in the UK which have differing student-funding approaches (Bell and Brooks 2018; Burgess et al. 2018). Student survey output metrics have been shown to vary due to factors such as student typology (e.g. ethnicity, gender, mode of study, age, and registered disability; Surridge 2008), which can to some extent be accounted for by reporting them in 'split metrics' (as used in the UK's Teaching Excellence Framework). However, satisfaction of educational experiences, such as student counselling services, learning environments and student communities, are considered influential but not evaluated as part of the NSS and difficult to measure in simplistic surveys (Hartman and Schmidt 1995).

It is interesting that the conceptual distribution of customer satisfaction (Peterson and Wilson 1992) fitted the NSS 'Overall Satisfaction' metric distribution at an institutional level, but not at a broad subject level. The growing spike in 100% agreeable returns at this more detailed level did not follow the expected consumer satisfaction pattern. It has been noted that 'yea saying', where respondents report they strongly agree for every survey item, has been reported previously and could represent around 5% of all responses (HEFCE 2013). This may have contributed to the pattern we report, but the recent increase appears too great to be a result of this behaviour in isolation. The presence of this rapidly growing 'satisfaction spike' suggests that the discriminatory potential of the survey instrument at the high end of the scale is being diminished. Whilst improvements in institutional educational practice at the individual course level (perhaps as a result of the NSS) could potentially explain the rapid increase in 100% satisfaction we observed (e.g. Brown 2011). Of course, other factors may have also contributed. For example, Palacio et al. (2002) describe how satisfaction is driven by perceived performance and individual expectations. Sampling bias is also a challenge if survey participation favours those that were more satisfied. One aspect of bias includes students that may be less likely to participate if they are neutral in their views (or unsatisfied). Alternatively, targeted actions in institutions or even courses could encourage satisfied students to respond. Such effects would hamper interpretation of educational quality when based on metric responses alone and perhaps warrant further research on this phenomenon. Other approaches to explore student satisfaction may yield different outcomes. For example, Douglas et al. (2015) found that student satisfaction is strongly related to how 'attentive' the teaching staff was to students during their studies, particularly in the final year. This possibly suggests student preference for a supportive, individualised experience and suggests that dissatisfaction was most acute when communication with students was perceived as poor. However, as long as national student survey metrics remain prominent in the HE landscape, their behaviours and predictors will remain of significant interest within the sector.

Conclusions

We used machine learning to evaluate and rank the consistency of predictors of student satisfaction. The dominant factors we identified that impact student satisfaction were in survey items relating to respondent perceptions of a smooth-running course and stimulating, well-

explained teaching. Student support had some impact, but items relating to Learning Resources, Assessment & Feedback and Personal Development had minimal influence on student satisfaction. We report a ceiling effect with an increasing trend in overall student satisfaction ratings coupled with reduced variability amongst institutions, and we suggest this could intensify competition amongst institutions. We suggest that the use of statistical outliers have great potential for national quality improvement. Finally, we also detected a rapid increase in total satisfaction at an individual course level that does not conform to the expected pattern of conceptual consumer satisfaction models.

References

Ball, S. J. (2017). *The education debate* (third ed.). Bristol: Policy Press.

Bell, A. R., & Brooks, C. (2018). What makes students satisfied? A discussion and analysis of the UK's national student survey. *Journal of Further and Higher Education, 42*(8), 1118–1142.

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. https://doi.org/10.1080/13538322.2014.924786.

Birnbaum, R. (2000). The life cycle of academic management fads. *The Journal of Higher Education, 71*(1), 1–16. https://doi.org/10.1080/00221546.2000.11780813.

Boehler, M. L., Rogers, D. A., Schwind, C. J., Mayforth, R., Quin, J., Williams, R. G., & Dunnington, G. (2006). An investigation of medical student reactions to feedback: a randomised controlled trial. *Medical Education, 40*(8), 746–749. https://doi.org/10.1111/j.1365-2929.2006.02503.x.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Brown, S. (2011). Bringing about positive change in the higher education student experience: a case study. *Quality Assurance in Education, 19*(3), 195–207. https://doi.org/10.1108/09684881111158027.

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, 13*(2), e0192976. https://doi.org/10.1371/journal.pone.0192976.

Callender, C., Ramsden, P., & Griggs, J. (2014). Review of the "National Student Survey": report to the UK higher education funding bodies by NatCen Social Research, the Institute of Education, University of London and the Institute for Employment Studies.

Cheng, J. H. S., & Marsh, H. W. (2010). National Student Survey: are differences between universities and courses reliable and meaningful? *Oxford Review of Education, 36*(6), 693–712. https://doi.org/10.1080/03054985.2010.491179.

Clayson, D. E. (2009). Student evaluations of teaching: are they related to what students learn?: a meta-analysis and review of the literature. *Journal of Marketing Education, 31*(1), 16–30. https://doi.org/10.1177/0273475308324086.

Dean, A., & Gibbs, P. (2015). Student satisfaction or happiness? A preliminary rethink of what is important in the student experience. *Quality Assurance in Education, 23*(1), 5–19. https://doi.org/10.1108/QAE-10-2013-0044.

Dill, D. D., & Soo, M. (2005). Academic quality, league tables, and public policy: a cross-national analysis of university ranking systems. *Higher Education, 49*(4), 495–533. https://doi.org/10.1007/s10734-004-1746-8.

Douglas, A., Douglas, J., & Barnes, B. (2006). Measuring student satisfaction at a UK university. *Quality Assurance in Education, 14*(3), 251–267. https://doi.org/10.1108/09684880610678568.

Douglas, J. A., Douglas, A., McClelland, R. J., & Davies, J. (2015). Understanding student satisfaction and dissatisfaction: an interpretive study in the UK higher education context. *Studies in Higher Education, 40*(2), 329–349. https://doi.org/10.1080/03075079.2013.842217.

Eley, M. (2001). The course experience questionnaire: altering question format and phrasing could improve the CEQ's effectiveness. *Higher Education Research and Development, 20*(3), 293–312. https://doi.org/10.1080/07294360127208.

Elliott, K. M., & Shin, D. (2002). Student satisfaction: an alternative approach to assessing this important concept. *Journal of Higher Education Policy and Management, 24*(2), 197–209. https://doi.org/10.1080/1360080022000013518.

Fielding, A., Dunleavy, P. J., & Langan, A. M. (2010). Interpreting context to the UK's National Student (satisfaction) Survey data for science subjects. *Journal of Further and Higher Education, 34*(3), 347–368. https://doi.org/10.1080/0309877X.2010.484054.

García-Aracil, A. (2008). European graduates' level of satisfaction with higher education. *Higher Education, 57*(1), 1. https://doi.org/10.1007/s10734-008-9121-9.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters, 31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician, 63*, 308–319.

Grove, J. (2013). *Global gender index. Times Higher Education*. http://www.timeshighereducation.co.uk/features/globalgender-index-2013/2003517.article. Accessed 15 April 2019.

Gunn, A. (2018). Metrics and methodologies for measuring teaching quality in higher education: developing the Teaching Excellence Framework (TEF). *Educational Review, 70*(2), 129–148. https://doi.org/10.1080/00131911.2017.1410106.

Hardman, J., Paucar-Caceres, A., & Fielding, A. (2013). Predicting students' progression in higher education by using the random forest algorithm. *Systems Research and Behavioral Science, 30*, 194–203. https://doi.org/10.1002/sres.2130.

Hartman, D. E., & Schmidt, S. L. (1995). Understanding student/alumni satisfaction from a consumer's perspective: the effects of institutional performance and program outcomes. *Research in Higher Education, 36*(2), 197–217. https://doi.org/10.1007/BF02207788.

Hazelkorn, E. (2007). The impact of league tables and ranking systems on higher education decision making. *Higher Education Management and Policy, 19*(2), 1–24. https://doi.org/10.1787/hemp-v19-art12-en.

Hazelkorn, E. (2015). *Rankings and the reshaping of higher education: the battle for world-class excellence*. Springer.

HEFCE. (2013). *UK review of the provision of information about higher education: National Student Survey results and trends analysis 2005–2013* (p. 60).

Kahu, E. R. (2013). Framing student engagement in higher education. *Studies in Higher Education, 38*, 758–773. https://doi.org/10.1080/03075079.2011.598505.

Kuh, G. D. (2003). What we're learning about student engagement from NSSE: benchmarks for effective educational practices. *Change: The Magazine of Higher Learning, 35*(2), 24–32. https://doi.org/10.1080/00091380309604090.

Kuh, G. D. (2009). The national survey of student engagement: conceptual and empirical foundations. *New Directions for Institutional Research, 141*, 5–20. https://doi.org/10.1002/ir.283.

Langan, A. M., Dunleavy, P., & Fielding, A. (2013). Applying models to national surveys of undergraduate science students: what affects ratings of satisfaction? *Educational Sciences, 3*(2), 193–207. https://doi.org/10.3390/educsci3020193.

Langan, A. M., Harris, W. E., Barrett, N., Hamshire, C. H., & Wibberley, C. (2016). Benchmarking factor selection and sensitivity: a case study with nursing courses. *Studies in Higher Education, 43*(9), 1586–1596. https://doi.org/10.1080/03075079.2016.1266610.

Langan, A. M., Scott, N., Partington, S., & Oczujda, A. (2017). Coherence between text comments and the quantitative ratings in the UK's National Student Survey. *Journal of Further and Higher Education, 41*(1), 16–29. https://doi.org/10.1080/0309877X.2014.1000281.

Lee, J. W., Lee, J. B., Park, M., & Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis, 48*(4), 869–885.

Lenton, P. (2015). Determining student satisfaction: an economic analysis of the National Student Survey. *Economics of Education Review, 47*, 118–127. https://doi.org/10.1016/j.econedurev.2015.05.001.

Letcher, D. W., & Neves, J. S. (2010). Determinants of undergraduate business student satisfaction. *Research in Higher Education Journal, 6*, 1.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2/3*(December), 18–22.

Locke, W. (2014). *Shifting academic careers: implications for enhancing professionalism in teaching and supporting learning (Report)*. York: Higher Education Academy https://www.heacademy.ac.uk/resource/shifting-academic-careers-implications-enhancing-professionalism-teaching-and-supported. Accessed 13 September 2018.

Marsh, H. W. (2007). Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: an evidence-based perspective* (pp. 319–383). Dordrecht: Springer Netherlands. https://doi.org/10.1007/1-4020-5742-3_9.

Molesworth, M., Nixon, E., & Scullion, R. (2009). Having, being and higher education: the marketisation of the university and the transformation of the student into consumer. *Teaching in Higher Education, 14*(3), 277–287. https://doi.org/10.1080/13562510902898841.

Murtonen, M., Gruber, H., & Lehtinen, E. (2017). The return of behaviourist epistemology: a review of learning outcomes studies. *Educational Research Review, 22*, 114–128. https://doi.org/10.1016/j.edurev.2017.08.001.

Neuendorf, K. A. (2016). *The content analysis guidebook*. Sage.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*, 250–256. https://doi.org/10.1037/0022-3514.35.4.250.

O'Neill, M. A., & Palmer, A. (2004). Importance-performance analysis: a useful tool for directing continuous quality improvement in higher education. *Quality Assurance in Education, 12*(1), 39–52. https://doi.org/10.1108/09684880410517423.

Palacio, M. A., Díaz Meneses, G., & Pérez Pérez, P. J. (2002). The configuration of the university image and its relationship with the satisfaction of students. *Journal of Educational Administration, 40*, 486–505. https://doi.org/10.1108/09578230210440311.

Payne, S. L., Kleine, K. L. M., Purcell, J., & Carter, G. R. (2005). Evaluating academic challenge beyond the NSSE. *Innovation Higher Education, 30*, 129–146. https://doi.org/10.1007/s10755-005-5015-2.

Peterson, R. A., & Wilson, W. R. (1992). Measuring customer satisfaction: fact and artifact. *Journal of the Academy of Marketing Science, 20*(1), 61. https://doi.org/10.1007/BF02723476.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems, 9*, 181–199. https://doi.org/10.1007/s10021-005-0054-1.

R Core Team. (2018). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing URL https://www.R-project.org/. Accessed 15 April 2019.

Radloff, A., & Coates, H. (2010). Doing more for learning: Enhancing engagement and outcomes - Australasian Student Engagement Report. *Australian Council for Educational Research*. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1011&context=ausse.

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. *Studies in Higher Education, 16*(2), 129–150. https://doi.org/10.1080/03075079112331382944.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment & Evaluation in Higher Education, 30*(4), 387–415. https://doi.org/10.1080/02602930500099193.

Rienties, B. (2014). Understanding academics' resistance towards (online) student evaluation. *Assessment & Evaluation in Higher Education, 39*, 987–1001. https://doi.org/10.1080/02602938.2014.880777.

Shah, M. (2012). Ten years of external quality audit in Australia: evaluating its effectiveness and success. *Assessment & Evaluation in Higher Education, 37*, 761–772. https://doi.org/10.1080/02602938.2011.572154.

Shah, M., Cheng, M., & Fitzgerald, R. (2017). Closing the loop on student feedback: the case of Australian and Scottish universities. *Higher Education, 74*(1), 115–129.

Sharpe, R. (2019). Evaluating the student experience: a critical review of the use of surveys to enhance the student experience. In K. Trimmer, T. Newman, & F. F. Padró (Eds.), *Ensuring quality in professional education volume II: engineering pedagogy and international knowledge structures* (pp. 29–45). Cham: Springer International Publishing.

Soh, K. (2017). The seven deadly sins of world university ranking: a summary from several papers. *Journal of Higher Education Policy and Management, 39*(1), 104–115. https://doi.org/10.1080/1360080X.2016.1254431.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*, 323–348. https://doi.org/10.1037/a0016973.

Surridge, P. (2008). *The National Student Survey 2005–2007: findings and trends (No. HEFCW/08/79 Annex A)*. Higher Education Funding Council for England.

Wilson, K. L., Lizzio, A., & Ramsden, P. (1997). The development, validation and application of the course experience questionnaire. *Studies in Higher Education, 22*(1), 33–53. https://doi.org/10.1080/03075079712331381121.

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

Yorke, M. (2013). Surveys of 'the student experience' and the politics of feedback. In *Reconceptualising feedback in higher education: developing dialogue with students* (pp. 6–18). London: Routledge.

Yorke, M. (2015). "Student experience" surveys: theoretical, political and empirical dimensions. In *Researching higher education: international perspectives on theory, policy and practice* (pp. 209–226). London: Routledge.

Yorke, M., Orr, S., & Blair, B. (2014). Hit by a perfect storm? Art & design in the National Student Survey. *Studies in Higher Education, 39*(10), 1788–1810. https://doi.org/10.1080/03075079.2013.806465.

Zepke, N. (2014) Student engagement research in higher education: questioning an academic orthodoxy. *Teaching in Higher Education, 19*(6), 697–708. https://doi.org/10.1080/13562517.2014.901956.