HIGHER EDUCATION PEDAGOGIES

Higher Education Pedagogies

ISSN: (Print) 2375-2696 (Online) Journal homepage: https://www.tandfonline.com/loi/rhep20

Lessons in learning gain: insights from a pilot project

F. Arico, H. Gillespie, S. Lancaster, N. Ward & A. Ylonen

To cite this article: F. Arico, H. Gillespie, S. Lancaster, N. Ward & A. Ylonen (2018) Lessons in learning gain: insights from a pilot project, Higher Education Pedagogies, 3:1, 249-265, DOI: 10.1080/23752696.2018.1454845

To link to this article: https://doi.org/10.1080/23752696.2018.1454845

Published online: 06 Sep 2018.

HIGHER EDUCATION PEDAGOGIES
2018, VOL. 3, NO. 1, 249–265
https://doi.org/10.1080/23752696.2018.1454845

OPEN ACCESS Check for updates

Lessons in learning gain: insights from a pilot project

F. Arico[a], H. Gillespie[b], S. Lancaster[c], N. Ward[d] and A. Ylonen[e]

[a]School of Economics, University of East Anglia, Norwich, UK; [b]School of Education and Lifelong Learning, University of East Anglia, Norwich, UK; [c]School of Chemistry, University of East Anglia, Norwich, UK; [d]Vice-Chancellor's Office, University of East Anglia, Norwich, UK; [e]School of Education and Lifelong Learning, University of East Anglia, Norwich, UK

ABSTRACT

'Learning gain' has become an increasingly prominent concept in debates about the effectiveness of higher education across OECD countries. In England, interest has been heightened by the Higher Education Funding Council for England (HEFCE)'s major research initiative on learning gain, launched in 2015, and by the new Teaching Excellence Framework which assesses learning and teaching and student outcomes. HEFCE's novel research initiative has funded a set of experimental projects across the English higher education sector for the first time. This paper presents preliminary findings from one such project at the University of East Anglia (UEA). The project trials and evaluates three approaches to identifying and measuring learning gain using data from cohorts of students across different discipline areas during 2015–2016 and 2016–2017. It builds upon previous work carried out at UEA in developing self-efficacy assessments and applying concept inventories. Student marks provide a simple comparator as a third approach to measuring learning gain.

ARTICLE HISTORY
Received 26 July 2017
Revised 1 February 2018
Accepted 16 March 2018

KEYWORDS
Learning gain in higher education; students marks; Grade Point Average (GPA); self-efficacy; concept inventories

Introduction

'Learning gain' has become increasingly prominent in debates about higher education provision and student achievement over recent years. This is an international trend, but has a particular expression in England. The introduction of £9000 tuition fees for English undergraduates in 2012, increasing competition for students and the development of the new Teaching Excellence Framework (TEF) to rate higher education providers according to educational metrics have all focused attention on student outcomes and how these might be measured. The Higher Education Funding Council for England (HEFCE) launched a major programme of new work in February 2015 to explore different approaches to identifying learning gain. The national government, in its Higher Education Green Paper, *Fulfilling Our Potential* (Department of Business, Innovation & Skills, 2015), also signalled the desirability of developing criteria for how students 'get added value from their studies' (p. 33). HEFCE's initiative has meant that the question of learning gain has gone from being

CONTACT A. Ylonen a.ylonen@uea.ac.uk

a little considered aspect of English higher education to a prominent focus of research and debate in a short space of time. This paper presents preliminary findings from one of the HEFCE-funded projects to explore methods of identifying and measuring learning gain across universities in England.[1]

Learning gain has come to international prominence in higher education through debates in the US. The Spellings Commission on the future of US higher education (US Department of Education, 2006) was established as a result of concerns about the clarity of student learning outcomes and anxiety that students' progress was not sufficiently actively and systematically measured. The question became a national public controversy following the publication of a book, *Academically Adrift* (Arum & Roksa, 2011), which presented an empirical study suggesting that as many as 45% of US college students demonstrated no significant improvement in key academic skills such as critical thinking, complex reasoning and writing during their first two years at college.

Internationally, the Organisation for Economic Co-operation and Development (OECD) has also been keen to develop comparative measures of student learning outcomes across different national higher education systems (Schleicher, 2015; Tremblay, Lalancette, & Roseveare, 2012). Tremblay et al. (2012) cite a range of factors which help account for the increasing interest in student learning outcomes and learning gain. These include:

(1) the expansion of national higher education systems
(2) the emergence of new types of higher education providers
(3) increasingly diverse profiles of institutions, programmes and students
(4) the growth of international education
(5) the growing emphasis on competition and use of quality signalling mechanisms such as performance indicator rankings
(6) new modes of higher education governance, which stress performance, quality and accountability.

All of these factors apply in the English context. However, it is the growing emphasis on competition and the concern of politicians to evidence 'value-for-money' from public investment and private tuition fees in university education that has been the strongest impulse in recent years, with 'learning gain' trailed as a potential conceptual key to unlock understanding of value for money. There has been significant transatlantic dialogue on the issue, and the 'Wabash study', in particular, has attracted considerable attention in England, with its authors invited to address audiences of British academics and policy-makers.[2]

So what, in essence, is learning gain? At its simplest, learning gain can be understood as the 'distance travelled' by a student – that is, the learning achieved between two points in time which could be the start and end of a course or programme. Higher Education Funding Council for England (2015) has defined learning gain as: 'the improvement in knowledge, skills, work-readiness and personal development made by students during their time spent in higher education'. This is a relatively open and flexible definition, which begs questions about the specific meaning of its key terms – improvement, knowledge, skills, work-readiness and personal development.

In 2015, HEFCE commissioned a major review of learning gain in higher education by the RAND Corporation (McGrath, Guerin, Harte, Frearson, & Manville, 2015). The study drew a useful distinction between the concept of value-added and learning gain by suggesting that: 'The concept of learning gain is often used interchangeably with the notion of value added. Learning gain is measured based on the difference between two measures of actual student performance, while value added is based on the comparison between performance predicted at the outset of studies and actual performance achieved' (McGrath et al., 2015, p. xi).

'Value-added' has some currency in British higher education as a result of its use as an indicator in one of the most widely used British university league tables – the Guardian University Guide.[3] The league table's measure of value-added is a complex composite indicator derived from a combination of information on the average entry tariff of students and the proportion of students achieving 'good honours' (a First or Upper Second Class degree). It therefore measures the attainment of student cohorts, in the form of degree classifications, while accounting for a measure of the quality of the intake. An institution which is adept at taking in students with relatively low entry qualifications, which are generally more difficult to convert into a 1st or 2:1, will score highly in the value-added measure if the number of students getting a 1st or 2:1 exceeds modelled expectations.

Value-added measures such as this one used in the Guardian are therefore measures of when a student or cohort of students outperforms what might be expected of them on the basis of statistical evidence from a wider population. While learning gain is more usually an *absolute* measure of a student or a cohort's progress, value-added is instead a *relative* measure of progress – relative to the progress that might typically be expected. (To complicate this distinction somewhat, some measures of learning gain which involve student's self-evaluation can also be considered relative rather than absolute measures).

Learning gain is most commonly defined in literature as a simple form of 'distance travelled' rather than value-added relative to expectations. Difficulties start to arise, however, when moving from defining what constitutes learning gain to operationalising these definitions in practice and seeking to measure learning gain. Here key debates centre around the use and advantages of different standardised tests, the relative merits of generic versus subject-specific skills, and the merits of statistical value-added modelling to measure learning gains. Some argue that self-reported learning gain measures are most appropriate, when relying on survey design that carefully accounts for academic engagement, discipline-specific aspects of learning and demographic differences amongst respondents. Empirical evidence from the US shows that self-reported data from student experience surveys display good correlation with student Grade Point Averages (GPAs) and perform better than standardised tests, such as the Collegiate Learning Assessment, at mirroring student learning. Nevertheless, these types of measures appear more useful when considered within, rather than across, higher education institutions. Institutional differences, such as the demographic composition of the student body, as well as specific configurations of learning outcomes, might affect the reliability of self-reported metrics, because they influence the culture of each specific institution and of their respondents, preventing meaningful comparisons. (Douglass, Thomson, & Zhao, 2012; Thomson & Douglass, 2009).

At the University of East Anglia (UEA), emerging interest in learning gain among a small group of academic staff led them to begin experimenting with different pedagogical approaches to deepening student engagement and facilitating real learning. The HEFCE initiative therefore offered an opportunity to formalise and expand this work and so contribute to a national research effort to strengthen understanding of approaches to measuring and supporting learning gain.

UEA is a medium-sized research intensive university in Norwich in the east of England. It has 16,000 students taught in disciplines across four faculties – arts and humanities, medicine and health sciences, science and social sciences. The University has typically ranked in the Top 20 in the three main British university league tables (the Times, the Complete Guide and the Guardian) over recent years and performs relatively strongly in the National Student Survey (NSS), which measures students' satisfaction with their course and their institution among all final year students.

In the sections that follow, the methodology, findings and implications of the ongoing learning gain project at UEA are presented and discussed.

Context

UEA's learning gain project sought to trial and evaluate three different approaches to identifying and measuring learning gain: (i) student marks; (ii) self-efficacy assessment; and (iii) concept inventories. The project took a cross-sectional approach, looking at different cohorts of students for the 2015–2016 academic year in Phase 1 of the project and for the 2016–2017 academic year in Phase 2. (The reason for a cross-sectional approach rather than a longitudinal approach was because, for our self-efficacy assessments, different students will take different modules each year and so tracing the same students over a prolonged period beyond a single academic year would not be possible).

The project built upon previous work carried at the UEA in developing self-efficacy assessments in the School of Economics and in using concept inventories in the School of Chemistry. The inclusion of student marks/GPAs as a further approach to measure learning gain emerged as a result of an initiative by the Higher Education Academy (HEA) which was promoting supplementing final degree classifications with a GPA (Higher Education Academy, [HEA], 2015). (This study was a response to concerns that too great a proportion of UK students were achieving 'good honours' and there was a perceived need to more strongly differentiate degree classifications). Wider interest was sought and in order to trial different learning gain approaches across a wider range of discipline areas and so further Schools of Study were incorporated into the project. Self-efficacy assessment was trialled in the School of Psychology in the Faculty of Social Sciences, in a Foundation Year module in the Faculty of Arts and Humanities and in Nursing in the Faculty of Medicine and Health Sciences. Concept inventories were also trialled in the School of Pharmacy and the School of Biology.

The two main research questions of the project to be considered in this paper are, first, what can each of the three approaches offer as an attempt to measure learning gain in higher education and, second, what are the lessons from how the different approaches are administered? In the following sections, we present the findings from our experimentation with each of the three approaches.

Findings

Student marks, higher degree classification and GPA

The student marks strand of the project explores how student progress is expressed in what might be regarded as the incumbent measure of student learning, the award of marks. At

Table 1. UEA's modification to the HEA Grade Point Average Scale.

Mark	Grade point	Classification
>=80	4.50	I
75–79	4.25	I
71–74	4.0	I
67–70	3.75	2 i (high) or I
64–66	3.50	2 i (med)
61–63	3.25	2 i (low)
57–60	3.0	2 ii (high)
54–56	2.75	2 ii (med)
50–53	2.50	2 ii (low)
48–49	2.25	3
43–47	2.0	3
40–42	1.50	3
38–39	1.0	Fail
35–37	.75	Fail
30–34	.5	Fail
<=29	.0	Fail

Source: Adapted from HEA (2015).

UEA, a percentage system is used for the award of marks on taught programmes, with marks awarded for individual pieces of assessment aggregated first at module level and then by year of study, with an established algorithm used for the conversion into a final percentage which feeds a calculation of higher degree classification (HDC) against the traditional categories of first class, upper second class, lower second class, third class and fail. This way of expressing student progress and achievement is common throughout UK higher education, although there are many differences in the way in which marks are awarded and their overall impact on an individual student's outcome. In addition, we recently decided to add a GPA calculation to the students' final degree transcript as a supplement to the HDC. It seems unlikely that the sector will abandon this traditional method of expressing a final degree outcome in favour of GPA, despite the HEA's extensive project on GPA (HEA, 2015). For the purposes of the learning gain project, we refer to the HEA scale with a small adaption. The scale has been amended by adding one incremental point to the top of the scale, accounting for marks of 80+(GPA 4.50). This was because we found that the existing HEA scale failed to differentiate sufficiently for students in our School of Mathematics. Table 1 shows UEA's modification to the HEA Grade Point Average Scale.

Stage by stage, the process of managing marks and classifying degrees at UEA is as follows:

i. An individual academic awards a provisional first mark to a piece of work, on a scale of 1–100. (The University provides a 'Senate Scale' to guide judgements in common forms of student work).
ii. The mark, before the student sees it, may be adjusted as a result of moderation or double marking, in accordance with the University's policy.
iii. The mark is transferred to the student record system and the student is given a provisional mark. System software calculates the overall module mark based on the marks of each individual component, weighted appropriately.
iv. Marks are approved on a yearly basis by the Boards of Examiners.
v. At the end of a student's period of study, an algorithm and set of University-wide rules are used to convert the marks into an overall percentage normally using 40% of the

Stage Two weighted year average and 60% of the Stage Three weighted year average. This classification mark directly maps to a HDC and GPA (see Table 1).

It is tempting to see this process as a systematic way of converting student performance into an outcome that can be expressed to all, usually for the purposes of comparison between students by employers and by those assessing suitability for further study. However, there are a number of reasons why this system could be regarded as problematic for the purposes of comparing student outcomes, both within and across institutions.

Higher education institutions take different approaches to algorithms to calculate outcomes and regulations vary on issues such as whether a student is required to pass all their modules. Even within institutions, the range of different types of assessment undertaken by students varies. With mark schemes and guidance, there are significant differences in the numbers and types of assessments undertaken by students on courses as diverse as Nursing and History (see, e.g. Gibbs, 2010; Pokorny, 2016). Furthermore, even with institution-wide scales and frameworks to guide the award of marks for assessments, it is inevitable that there will be elements of disciplinary marking cultures which make comparison of marks across an institution complex (Pokorny, 2016; Sambell, McDowell, & Montgomery, 2013).

Calculating learning gain using student marks

Our approach to using student marks to calculate learning gain compared a standard measure of actual percentage marks awarded at two points in time. The study looked at undergraduate student marks across all Schools of study across the University. This approach created 20 groups, classifying Integrated Masters courses and degrees with foundation years in science schools separately and excluding Norwich Medical School, because it does not award marks at the module level on its medical degree. We compared the average mark per student cohort, first by School and then by route (standard, with foundation year, or, with integrated master year). We calculated an average mark using the last 5 years of student cohorts' marks at the end of Year 1 and compared them to the average mark (calculated in the same way) at the end of Year 3. We then converted this to a raw GPA and used an amended form of the HEA GPA scale to give each student a banded GPA.

Figure 1 above shows differences in measurements of learning gain expressed as the difference in average marks in the first column and as the difference expressed as banded GPA in the other. The apparent range of variation in distance travelled is significant. Expressed as marks, the difference between the cohort with the greatest distance travelled (average student mark 5.52% higher in final year than first year) and the cohort with the lowest (average student mark 4.58% lower) is over 10%. It should be noted that when looking at the absolute average, the differences rarely move through degree classification boundaries. For example, an average mark moving from 62 to 66% still produces an upper second class outcome.

The results also show a pattern in the Faculty of Science, with all of the cohorts showing a lower average mark in the final year as compared to the first. More work needs to be done on the causes of this difference as there is no evidence of 'unlearning' or other quality problems. It seems more likely that the differences are due to variations in the assessment norms and expectations around progression in different disciplines rather than differences in

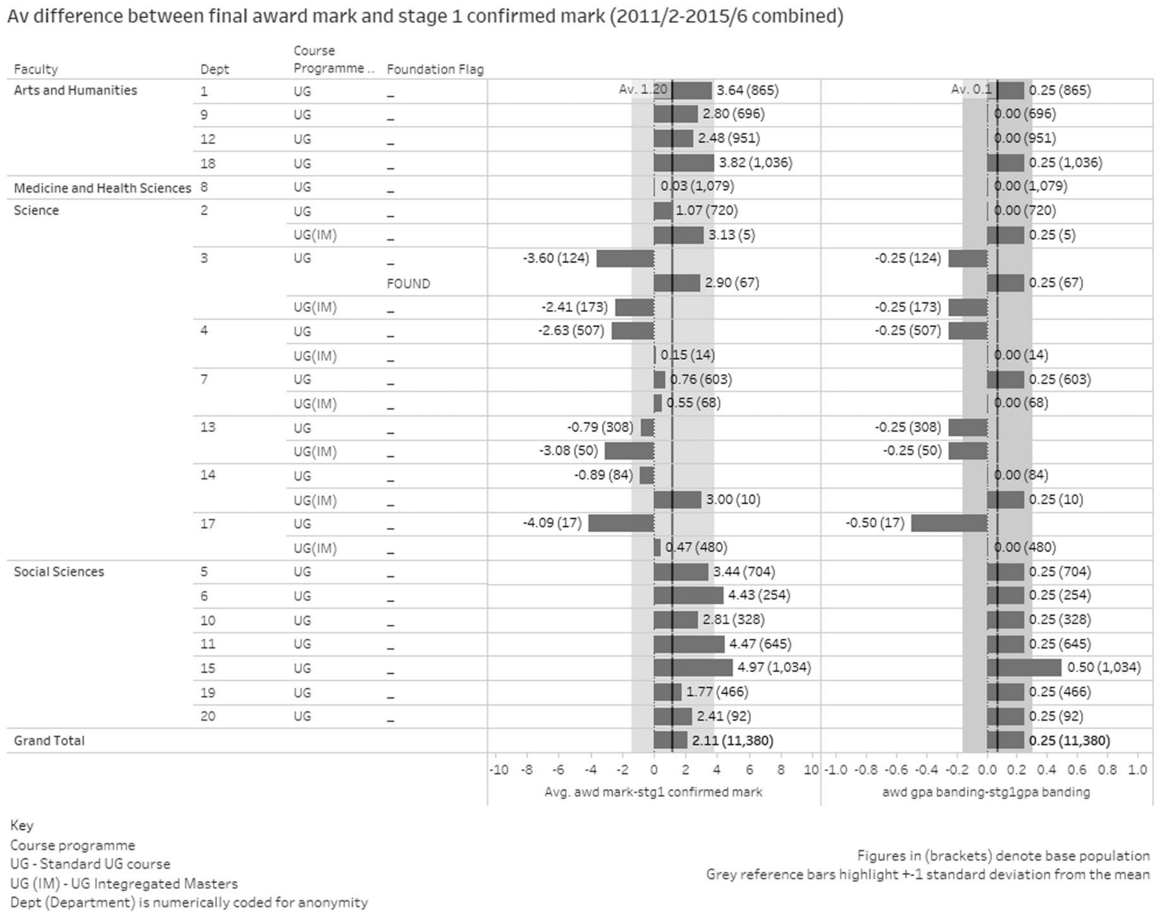

Figure 1. Average difference between final award mark and stage 1 mark across the four main faculties at UEA.

student attainment. Consequently, we conducted interviews with academics from different disciples to begin to examine the reasons behind these differences.

The interviews were carried out on a semi structured basis and covered a range of topics related to the assessment process. The participants were volunteers from disciplines across the University in the subjects of biology, linguistics, business studies, speech and language therapy and environmental sciences. The participants discussed what they thought might be the source of inconsistencies in the assessment process, and from the responses it is clear that while all subjects use a 0–100 percentage scale to award marks at undergraduate level, the practices behind the award of marks are not consistent across disciplines, even though they are all working within university policies and procedures. The emerging findings, which highlight some of the reasons behind these inconsistencies, include:

- While a generic marking scale is applied across the University, some academics develop more subject-based marking rubrics.
- The nature of subjects give different marking profiles, with mathematical subjects producing a different (bimodal) distribution of marks when compared to essay-based subjects which tend to be more clustered.
- There is an acceptance of the subjectivity of the marking process in some subjects, especially when it comes to small differences (for example 2%) in marks awarded.
- The nature of the assessment design varies from course to course with some students having to produce different numbers of assessments for modules of the same credit size.
- Opportunities to discuss marking and assessment approaches between schools are limited.

Further work is needed to test the hypothesis that different subjects produce different marking distributions as well as to illuminate the reasons behind this. Understanding the impact of generic marking scales also requires further work and we are currently planning a research strategy around this investigation

Self-efficacy assessment

The second strand to the project examines academic self-efficacy (ASE). This is defined as students' confidence in their ability to accomplish specific tasks or attain specific goals (Bandura, 1977), and is a core learning skill developed alongside the academic curriculum. In his comprehensive survey of empirical contributions, Pajares (1996) highlights that when self-efficacy beliefs are carefully aligned with specific academic tasks, ASE measures are a powerful predictor of student achievement. This trend is reflected by more recent evidence showing that students displaying high ASE are also high achievers (e.g. Chemers, Hu, & Garcia, 2001), more resilient and autonomous learners (McMillan & Hearn, 2008; Zimmerman, 2002), and have better employability prospects (Yorke & Knight, 2007). ASE evolves over time, and can be influenced in the classroom through a number of teaching techniques aimed to elicit student self-reflection and self-assessment skills.

An HEA-funded pilot study (2013–2014), allowed a pedagogy aimed at increasing ASE in large class Year 1 core module in Introductory Macroeconomics to be devised and piloted. In particular, the pedagogy enabled students to: (i) develop self-assessment skills; and (ii) gain mastery of experience, which both act as core contributors in the development of ASE beliefs. The pedagogy is implemented through formative assessment sessions, rolling over the course of the entire academic year. In each session, students engage in a learning algorithm based on multiple choice questions (MCQs). The learning algorithm is structured as follows: (i) students provide an individual and independent response to an MCQ; (ii) students self-assess their skills and their ability to address the question asked to them; (iii) peer-instruction takes place, as students compare their answers and discuss methods to solve the question asked; (iv) students provide a second individual response to the same MCQ; the correct answer is revealed and the session facilitator provides a final explanation to the problem which students have just attempted to solve. Finally (vi) students re-iterate self-assessment, ranking their confidence at being able to tackle similar questions in the future. The algorithm is repeated for each MCQ in the problem set, which comprises 8–10 questions per session. Intense use of learning technologies, such as student response systems, facilitates the collection of data-sets reporting on student performance and student self-assessment statements. In addition, in accordance with the ASE methodology, students receive individualised student reports on their performance via email, as well as immediate results via the student response system in class.

The pilot study identified four learning gain indicators, which are also used as indicators in the Learning Gain project through newer data-sets collected in 2015–2016 and 2016–2017. These were:

(1) self-efficacy, measured as student self-reported confidence in formative assessment performance;
(2) self-assessment skills, measured as the statistical association of student confidence and student performance;
(3) peer-instruction learning gains, measured as the difference in the proportion of correct responses to questions before and after peer-instruction; and
(4) peer-instruction confidence gains, measured as the difference in the proportion of confidence statements expressed by students before and after peer-instruction.

These indicators are computed at class-level as well as at student-level, and can be matched with past and future sessions to track individual student performance across the entire academic year. The large number of MCQs administered to students ($N > 100$ across all sessions, for class-level analysis) and the large number of students in the cohorts analysed ($N > 200$, for student-level analysis) guarantee that empirical results are statistically significant.

To conduct investigations at class-level, regression analysis is used to identify the correlation between the proportions of correct and confident responses for each MCQ question administered to students over the course of an academic year. Control variables are added to identify the effects of specific sessions. In order to develop student-level analysis, attainment scores and confidence scores are averaged and cross-tabulated to detect the association between the two dimensions of learning through simple Exact Fisher's tests.

Preliminary results of the self-efficacy data obtained from the Introductory Macroeconomics module indicate that the pedagogy introduced allows students to develop good self-assessment skills. In other words, students who display higher ASE are also those who score higher in formative assessment sessions, while students who are less confident about their performance are also those who score lower. This is particularly significant as the dominant empirical self-assessment literature claims that low-performers also display poor self-assessment skills: the Dunning–Kruger effect (Dunning, Johnson, Ehrlinger, & Kruger, 2003; Kruger & Dunning, 1999). The Dunning–Kruger effect is particularly detrimental in an academic context, because it predicts that those students who struggle with learning material are also those who display more difficulties in forming a reliable judgement about their competences. Therefore, by overcoming the Dunning–Kruger effect, our results are consistent with the hypothesis that the pilot pedagogy enables low-performers to identify their difficulties and seek for help when needed. (Qualitative data obtained from focus group interviews for Economics students in 2013–2014 shows that low-performing students developed strategies to react to poor performance, studying more or seeking for academic support by the teaching team). At the same time, the pedagogy also enables high-performers to develop awareness of their achievements, fostering positive self-efficacy beliefs.

The previous result is re-enforced by considering the change in the proportion of correct and confident responses given by students along multiple iterations of the pedagogy's learning algorithm. Positive learning gain is associated with confidence gain. When students learn from each other in the classroom, their confidence at tackling similar problems in the future also increases. These first two results are confirmed when considering gains both at class-level and at student-level.

When considering the association at class level between peer-instruction learning gain and the proportion of correct responses recorded at the first stage of the learning algorithm, an important insight emerges. As we described above, the pedagogy prescribes that students respond to formative questions twice, before and after peer-instruction. Analysing these responses, we find that when the class displays an initially low proportion of correct responses before class discussion, peer-instruction generates a high learning gain. This finding is consistent with the hypothesis that the pedagogy generates a 'catch-up' effect in the classroom. In particular, by exploiting the power of class interaction, students who do not respond correctly to a question before peer-instruction do benefit from discussion with more knowledgeable peers and respond correctly to the questions once they are reassessed at the end of the peer-instruction cycle (Mazur, 1997).

The pedagogy, developed within the School of Economics, has been used with cohorts of students attending modules in the Schools of Chemistry and Pharmacy, both in 2015–2016 and 2016–2017, and in a group of students attending a Foundation Year in Humanities, in 2016–2017.

Concept inventories

The third strand of the project concerns concept inventories. A concept inventory is a research instrument constructed to rigorously determine conceptual understanding in a given discipline. The definitive example is the Force Concept Inventory in physics (Hestenes, Wells, & Swackhamer, 1992). Concept inventory developers will convene student focus groups, and explore their understanding of the underlying concepts, seeking to surface misconceptions. A series of MCQs investigating conceptual understanding are subsequently prepared. These questions are validated by a panel of discipline experts to ensure they do indeed address the relevant concepts. The statistical reliability of all the individual components of the instrument is extensively trialled before application in education research. In order to reach satisfactory levels of validity and reliability multiple iterations and refinements are typically required. This process can take several years and is therefore a considerable investment in time and money. As a consequence, concept inventories only exist for a limited number of topics. We regard concept inventories, where available, as the best measures of student conceptual understanding available.

Concept inventories have been developed not to rank students or institutions but to provide a dependable insight into the conceptual understanding of an individual or cohort at a moment in time. By administering a concept inventory at the beginning and the end of a course we can seek to determine the difference in conceptual understanding during that course.

In order to preserve the utility and integrity of a concept inventory, the questions and their answers should not be published because this would be likely to lead to applications of less rigorous nature via diagnostic and assessment approaches. This would, in turn, risk students being repeatedly exposed to the questions and encourage them to adopt a rote learning approach thus negating the inventory's ability to assess conceptual understanding (Galloway & Lancaster, 2016). However, after the second application, with the learning gain determined, the educator might choose to provide general feedback to the students.

The impact of teaching interventions is often reported as the absolute increase in percentage score in routine assessment versus a control group or the previous year's cohort (Freeman et al., 2014). For a 'before-and-after' concept inventory investigation, the best metric is the *mean normalised gain*. This measure of learning gain directly expresses the fraction of questions the students were getting wrong before the course and are now getting right after the class.

The normalised gain is calculated according to the equation below (where Pre and Post are an individual student's percentage score before and after the course, respectively):

$$g = \frac{\text{Post} - \text{Pre}}{100\% \ - \text{Pre}}$$

While the numerator is a measure of the absolute improvement of the student's mark, the denominator is effectively a correction factor which takes into account the remaining scope for improvement. The normalised gain therefore recognises that it is easier for an initially low-scoring student to have a larger absolute improvement than an initially high-scoring student. The mean normalised gain $<g>$is the mean of the individual gains of each of the students in the cohort.

The maximum possible normalised gain is 1.0. This would occur if the student answered all the questions incorrectly in the pre-test and all the questions correctly in the post-test. If the student made no improvement the normalised gain would be 0. It is possible for a student's mark to go down as well as up. If this were to happen the normalised gain would be negative. A criticism sometimes levelled at the metric is that it is asymmetric. A student who does very well in the pre-test and whose mark deteriorates dramatically will record a large negative gain. The maximum value of g is 1, but the minimum value is $-\infty$.

In his seminal study, Hake (1998) used mean normalised gains on the Force Concept Inventory to compare the Physics instruction of US college students. One group were taught using traditional passive instruction and the second group using engaging interactive teaching approaches. For traditional lectures $<g>$ was in the range of 0–.3, whereas for the interactive classes $<g>$ varied from .2 to .7, and typically .5. Hake (1998) concluded that interactive teaching was generally significantly more effective than purely transmissive delivery. However, there was an overlap and that some interactive courses did less well than the most successful lectures.

Currently, there are not many concept inventories in chemistry nor are they as well known as the Force Concept Inventory in physics discussed above. Of the small number of chemistry concept inventories that have been published, the *Bonding Concepts Inventory* developed by the Lowery-Bretz group in the US had by far the best match with our target module and was therefore chosen to be used in this project. The *Bonding Concepts Inventory* was used with the kind permission of the authors (Luxford & Bretz, 2014) on condition that the questions were not released into the public domain. Although bonding concepts are a less significant learning outcome on the Pharmacy and Environmental Science degrees, the Bonding Concepts Inventory was also used with students at FHEQ Level 4. Another chemistry concept inventory published by Mulford and Robinson (2002) aims to determine the conceptual understanding of students at the FHEQ Level 3. As a result it was unsuitable to the evaluation of modular undergraduate teaching at FHEQ Level 4.

FHEQ Level 4 Biology students took a sub-set of the Molecular Life Sciences concept inventory developed by Howitt and colleagues (Howitt, Anderson, Costa, Hamilton, & Wright, 2008). The questions were selected by the lead academic on the basis of their relevance to the module. In contrast, the Bonding Concept Inventory was administered as a single instrument without editing to Chemistry, Pharmacy and Environmental Science Students. In this paper, we will only focus on discussing the findings from the Chemistry cohort.

In order to determine the learning gain, the concept inventories were administered as closely as practically possible to the beginning and end of the respective modules. The first sitting of the inventory was locked away until the second sitting had taken place and before any marking was conducted. Despite the objective nature of the MCQs, the scripts were marked by an independent party and by doing so we sought to minimise the risk of investigator bias.

For the chemistry cohort, 73 students took both the first and second sit of the bonding concept inventory. The highest individual learning gain, *g*, found was .82, the lowest result was −.67. The mean normalised gain, <*g*>, was .19, indicating that the students did better on the inventory at the end of the module. Students were not informed in advance when the second run was to take place and the marks obtained did not contribute to the module grading. Based on these results, it appears that a conceptual learning gain consequential to the module took place. The regression analysis undertaken indicates that those students who performed worse in the first run exhibited larger absolute improvements in their conceptual understanding than their better performing peers.

Discussion

In this concluding section, we summarise the main findings of the project and consider the potential strengths and weaknesses of the three approaches, and possible implications of introducing these types of learning gain approaches. Finally, we outline some possible future scenarios for the use of learning gain metrics in English higher education.

The first research question focused on assessing what each of the approaches (student marks and GPA/self-efficacy/concept inventories) can offer as an attempt to measure learning gain at UEA and in higher education more broadly. In developing and implementing learning gain initiatives generally, the fact that there was already strong staff buy-in at UEA helped to facilitate the implementation of the project. Existing capacity as well as a strong teaching-focused culture alongside the University's research culture acted as enablers and helped to facilitate staff interest. The fact that the UEA project built upon previous work also helped. This should be a key consideration for institutions wishing to implement experimental learning gain initiatives and projects. The concept of standardised learning remains contested in higher education generally, as it does at UEA, and attempts to construct measures can be seen by some staff to be misconceived. For some, 'measurement' in the university is unwelcome and uncomfortable and the notion of the 'measured university' is often discussed in dystopian terms. Peseta, Barrie, and McLean (2017, pp.453–4) make a strong argument, which is worth quoting in full:

> The demand to count, measure, rank, quantify, evaluate and judge the work of universities (along with those who labour and study in them) haunts virtually all aspects of our work: from the quality of research, to targets for income generation, counts of patents, citations of articles and public testimonies of policy impact made visible and likeable online; from the quality of curriculum, to teaching with technology, responding to student feedback, watching the employment destinations and salaries of graduates as a comment on the value of their education; to whether a university is healthy, sustainable, sufficiently globalized or doing enough to position itself as the world leader in this or that discipline. Every day, our conduct is being shaped to procure a commitment to institutional indicators, targets, standards and benchmarks that help us to diagnose ourselves (and others) as worthy and successful academics.

Any initiative to develop quantitative measures around the efficacy of learning and teaching can potentially be interpreted as part of wider surveillance and performance management cultures – a suspicion that is only further fuelled by national initiatives like the TEF. Another issue to consider is research ethics, which can add an additional burden to pursuing these types of research initiatives. This raises the general question of where evaluating teaching ends and doing research begins. Institutions need to consider this issue carefully.[4]

When it comes to analysing and using student marks as a way to measure learning gain, existing systems and units in place that deal with student marks are important. At UEA, the role of the University's Business Intelligence Unit (BIU) and their existing systems for data analysis regarding student performance facilitated the use of student marks and GPA as measures of learning gain. The BIU's capabilities include the ability to analyse, question and internally publish these data. In addition, UEA was already interested in the GPA concept and were associated with the HEA GPA project. UEA has been actively working on understanding 'Good Honours' issues and introduced revised Bachelor and Integrated Masters award regulations in order to deepen student academic engagement. Standardised university-wide guidance on how marks are arrived at, the Senate Scale, has helped heighten interest in issues around spread of marks and standardising approaches. There is also an interest in understanding differences in marking cultures and assumptions across different discipline areas. However, policy and regulations can change over time and therefore can distort the picture that the data provides. Good Honours forms part of league tables and so risks becoming competitive, contentious and politicised through, for example, arguments about gaming and 'dumbing down' in the higher education sector more widely. Indeed, the recent introduction of a so-called 'grade inflation' metric to the TEF requires that any improvement in the proportion of students being awarded Good Honours be considered by TEF assessors as potentially signalling a loss of 'rigour and stretch' rather than improved pedagogy, student effort and attainment. The logic of such a TEF assessment may, perversely, discourage institutions from exploring how better to support improved student attainment. Conservatism within distinct disciplinary marking cultures can also hamper the use and analysis of student marks in standardised ways across an institution as a means of performance improvement.

A distinct benefit of assessing ASE is that it is embedded in the pedagogy and so does not require additional student buy-in or engagement. At UEA, assessing ASE has been actively championed within the School of Economics for a number of years and there has subsequently been some wider interests in other schools to implement this type of formative assessment. Findings from our project have indicated that assessing ASE helps to crystallise interest and catalyse action among a set of similarly interested academic staff across the University who are interested in raising student self-efficacy. In some programmes at UEA, an attempt to raise student confidence is a core objective because some students can lack confidence. Work in ASE will continue at UEA beyond the life of the HEFCE project because the approach is well aligned to personalised learning and reflection. However, in order to avoid ASE assessment 'fatigue' and over-burdening staff, there is a need to consider an overarching strategy at course or department level to determine in which modules assessing ASE should be pursued.

With concept inventories, having an active and interested champion leading this strand of work was a key factor at UEA and this is likely to be the case at other universities interested in trialling concept inventories. We were also able to hire a student intern to improve the data analysis capacity, which could have otherwise presented challenges due to the time-consuming nature of such work. Unlike with ASE, because concept inventories can be seen by students as being additional rather than being embedded in the pedagogy, this impacted on student engagement. Furthermore, cross-disciplinary relevance of the tool limits the range of the applicability of concept inventories, which can limit staff interest and buy-in.

The second research question focused on assessing the benefits and drawbacks of each approach to measuring learning gain. The advantages of using student marks to measure learning gain are, firstly, that data are collected routinely already, so no additional data collection is required and the data measures are already embedded within current assessment pedagogy. A further strength of GPA is that it has a higher resolution than standard degree classification. These data are also more conducive to robust quantitative analysis, which can have benefits such as the ability to make generalisations. There are some disadvantages, however. For example, marks are a function of examination or assessment technique as well as student learning. It cannot be clear whether differences in distance travelled measured by marks are the result of learning gain or different assessment norms and cultures. In other words, does material get progressively 'harder'? (Of course, it is a significant area of contention in educational research whether assessment measures learning at all, or whether it simply indicates aptitude for passing assessments). Our study showed that different marking and assessment cultures are present across disciplines and that using a consistent marking scale does not necessarily equate with consistent award of marks. Simply using a comparable marking scale (1–100%) does not necessarily imply comparability between marks awarded in different disciples because of underlying influences on the award of marks. Further research needs to be carried out in what is a complex and multifaceted area.

An advantage of ASE is that the assessment is embedded in the pedagogy. The approach also helpfully reveals which kinds of questions are most effective in enhancing students' confidence in their learning. Being open and explicit about the question of student confidence appears to be good for student well-being and the approach produces a 'co-determined' measure – the result of both the teacher teaching and the student as an active partner in learning. As with student marks, data are also more conducive to quantitative analysis.

With concept inventories, a key strength of the approach is the validity and reliability of the measure which means high level of confidence among experts. It is also strongly focused on conceptual understanding. Furthermore, the universality of the measure enables comparisons across cohorts, institutions and countries. However, concept inventories are very time-consuming and labour-intensive (and so expensive) to develop. They are not universally available, only covering particular concepts in some disciplinary areas. Their international origins can also mean problems of translation (e.g. in units of measure such as pounds and ounces). While they have a strong focus on conceptual understanding they are less able to capture creativity.

None of the three approaches offer a single 'magic bullet' solution to the problem of identifying and measuring learning gain. They each have their own value and can produce useful insights within and across disciplines. However, even the approach that might at first sight be assumed to be most generalisable – using student marks – is revealed to have some shortcomings when applied across disciplines within a single institution. These difficulties are only likely to be compounded if the approach were attempted to be applied across institutions.

The project presented in this paper is situated in the context of a larger programme run by HEFCE investigating learning gain metrics in higher education as well as the new policy context of the TEF. The TEF assessment criteria for Year 2 (Department for Education, 2016) set out three aspects of quality: teaching quality, learning environment and student outcomes and learning gain. In this iteration of the TEF, the quantitative data used to make judgements about student outcomes and learning gain was limited to the Destination of Leavers from Higher Education Survey (DLHE), soon to be replaced by the new Graduate Outcomes Record. The HEFCE Learning Gain programme sets out to widen the available pool of available data on student learning gain that could potentially be used for TEF purposes. However, as our project shows, there are significant barriers to establishing comparable measures of learning gain even within an institution. This means that it will be challenging, to say the least, to establish nationally comparable learning gain measures.

Notes

1. For further details of the different projects, see: http://www.hefce.ac.uk/lt/lg/projects/.
2. Further details of the work undertaken by the Centre of Inquiry at Wabash College in the US can be found at: www.liberalarts.wabash.edu/study-overview/
3. http://www.theguardian.com/education/universityguide
4. For this project, the research was considered and approved by the Research Ethics Committees in the School of Economics and the School of Education and Lifelong Learning.

Acknowledgements

We would like to thank the Higher Education Founding Council for England (HEFCE) for funding this project under its Learning Gain initiative (Grant Code Number: G0177/IEAS01/LGP1517). We are also grateful for the constructive feedback from two sets of referees.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Higher Education Funding Council for England [G0177/IEAS01/LGP1517].

ORCID

A. Ylonen http://orcid.org/0000-0001-6692-7528

References

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215.

Chemers, M.M., Hu, L., & Garcia, B.F. (2001). Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational Psychology, 93*(1), 55–64.

Department of Business, Innovation and Skills (2015). *Fulfilling our potential: Teaching excellence, social mobility and student choice*. London: Author.

Department for Education (2016). *Teaching Excellence Framework: Year two specification*. London: Author.

Douglass, J.A., Thomson, G., & Zhao, C.-M. (2012). The learning outcomes race: The value of self-reported gains in large research universities. *Higher Education, 64*, 317–335.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3), 83–87.

Freeman, S., Eddy, S., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H., & Wenderoth, M.P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410–8415.

Galloway, R., & Lancaster, S. (2016, May). Learning gains. *Education in Chemistry*, 26–29.

Gibbs, G. (2010). *Using assessment to support student learning*. Leeds: Leeds Metropolitan University.

Hake, R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*, 1.

Higher Education Academy (2015). *Grade point average: Report of the GPA pilot project 2013–14*. York: HEA.

Higher Education Funding Council for England. (2015). *Learning gain*. Retrieved from http://www.hefce.ac.uk/lt/lg/

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141–158.

Howitt, S., Anderson, T., Costa, M., Hamilton, S., & Wright, T. (2008). A concept inventory for the life sciences: How will it help your teaching practice? *Australian Biochemist, 39*, 14–17.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Luxford, C.J., & Bretz, S. (2014). Development of the bonding representations inventory to identify student misconceptions about covalent and ionic bonding representations. *Journal of Chemical Education, 91*(3), 312–320.

Mazur, E. (1997). *Peer instruction: A user's manual*. Englewood Cliffs, NJ: Prentice Hall.

McGrath, C., Guerin, B., Harte, E., Frearson, M., & Manville, C. (2015). *Learning gain in higher education*. Cambridge: RAND Corporation.

McMillan, J., & Hearn, J. (2008). Student self-assessment: The key to stronger student motivation and higher achievement. *Educational Horizons, 87*(1), 40–49.

Mulford, D.R., & Robinson, W.R. (2002). An inventory for alternate conceptions among first-semester general chemistry students. *Journal of Chemical Education, 79*(6), 739–744.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*(4), 543–578.

Peseta, T., Barrie, S., & McLean, J. (2017). Academic life in the measured university: Pleasures, paradoxes and politics. *Higher Education Research and Development, 36*(3), 453–457.

Pokorny, H. (2016). Assessment for learning. In: H. Pokorny & D. Warren (Eds.), *Enhancing teaching practice in Higher Education* (pp. 69–90). London: Sage.

Sambell, K., McDowell, L., & Montgomery, C. (2013). *Assessment for learning in higher education*. Abingdon: Routledge.

Schleicher, A. (2015). *Value-added: How do you measure whether universities are delivering for their students? Higher Education Policy Institute Annual Lecture 2015*. London: HEPI.

Thomson, G., & Douglass, J.A. (2009). *Decoding learning gains, Research & Occasional Paper Series, Centre for Studies in Higher Education*. Berkley: University of California.

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). *Assessment of higher education learning outcomes (AHEHO): Feasibility Study Report Volume 1 – Design and Implementation*. Paris: OECD.

US Department of Education (2006). *A test of leadership: Charting the future of US higher education*. Washington, DC: Author.

Yorke, M., & Knight, P. (2007). Evidence-informed pedagogy and the enhancement of student employability. *Teaching in Higher Education, 12*(2), 157–170.

Zimmerman, B.J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*, 64–70.