Rogaten, J., Rienties, B. (2020). A critical review of learning gains on its methods and approaches. In C. Hughes and M. Tight. *International Perspectives on Higher Education Research*. Emerald.

A critical review of learning gains on its methods and approaches

Jekaterina Rogaten & Bart Rienties

Abstract

In the last five years there is an increased interest across the globe and the UK in particular to define, conceptualise, and measure learning gains. The concept of learning gains, briefly summarised as the improvement in knowledge, skills, work-readiness and personal development made by students during their time spent in higher education, has been hailed by some as an opportunity to measure "excellence" in teaching. This chapter will review some of the common definitions and the methods employed in research on learning gains. Secondly, we will provide a critical evaluation of the computational aspects of learning gains (e.g., raw gain, normalised gain). Finally, we will critically reflect upon the lessons learned and what is not yet known in terms of learning gains.

Introduction

Since the Burgess Report (Universities UK, 2004) substantial efforts have been made by a significant number of researchers, teachers, and higher education institutions (HEIs) across the globe to define, conceptualise, and measure learning gains (Cahill et al., 2014; Evans, Kandiko Howson, & Forsythe, 2018; Rogaten, Rienties, et al., 2019; Roohr, Liu, & Liu, 2017). The concept of learning gains, briefly summarised as the improvement in knowledge, skills, work-readiness and personal development made by students during their time spent in higher education, has been hailed by some as an opportunity to measure "excellence" in teaching. In the last four years in the UK, the concept of learning gains has obtained substantial traction in research, media, and policy making (Hughes, 2018; McGrath, Guerin, Harte, Frearson, & Manville, 2015; McKie, 2018).

In particular most recently interest has been raised in linking learning gains within initiatives such as the Teaching Excellence Framework (TEF). The yet unresolved question of how to measure learning gains at a national level has led to an interesting paradox within TEF as HEIs were asked to identify how they measure learning gain as if this a normal and common practice (Hughes, 2018; Kandiko Howson, 2019; McKie, 2018). This of course now leaves open scope for the development of multiple definitions of learning gain along with an associated plethora of measurement methods (Evans et al., 2018; Kandiko Howson, 2019).

As also highlighted elsewhere in this book (XXX), a range of definitions have been proposed to describe learning gains. The simplest perhaps most elegant form suggests that learning gain can be defined as what is learned between two (or more) time points (Pampaka et al., 2018). Baume (2018, p. 51) provided a slightly broader conceptualisation of learning gains, namely "the academic, professional and/or personal value added, that higher education provides". Rogaten, Rienties, et al. (2019, p. 321) specifically linked the notion of learning gains to intended and desired learning outcomes, whereby learning gains are defined as "as growth or change in knowledge, skills, and abilities over time that can be linked to the desired learning outcomes or learning goals of the course".

The different emphasis of each definition suggests a complexity. However, in reality many policy makers and researchers anticipate a link between good to excellent teaching and students subsequently achieving higher learning gains. Thus, effectively defining and then establish a measure for learning gains may help policy makers to determine which institutions provide the best value for money (Everson, 2017; Hake, 1998; McKie, 2018), and perhaps more controversially should be rewarded accordingly. It is perhaps not a surprise that in a recent special issue on learning gains Evans et al. (2018) indicated that "measuring learning gain is considered a policy panacea, a holy grail". However, the jury is still out whether we can actually define and measure learning gains, and whether or not governments should use the concept of learning gains to inform policy.

In this Chapter we aim to provide a brief methodological overview of the learning gains field and hope to inform teachers, learners, policy makers and researchers about potential future research directions and applications of learning gains for assessing HE excellence. Therefore, in this chapter we will provide an overview of our own experiences testing, evaluating, and implementing learning gains and then concentrate on the three main aspects of measuring learning gains. Thus, this chapter firstly will cover some of the definitions and the methods commonly employed in research on learning gains. Secondly, we will provide a critical evaluation of the computational aspects of learning gains. Finally we will critically reflect upon the lessons learned and what is not yet known in terms of learning gains.

What is known about measuring learning gains

The introduction of accountability process in HE is one of the key trends and in the last ten years it attracted substantial attention from the researchers, educators and policy makers

(Everson, 2017; Varsavsky, Matthews, & Hodgson, 2014). Australia (Boud, 2000, 2018; Varsavsky et al., 2014), UK (Department of Business, 2015; Forsythe, Evans, Kandiko Howson, & Edwards, 2018), and US (Arum & Roksa, 2014; Douglass, Thomson, & Zhao, 2012) are front-runners who have specialised agencies monitoring institutional quality and teaching standards, such as the Office for Students in the UK. For example, the UK government introduced the TEF process to help to reshape the UK HE landscape by encouraging universities to put students at the centre of their activity (Ashwin, 2017; Kandiko Howson, 2019; Turner et al., 2018). However, obvious questions can be raised in terms of the standards of excellence against which universities can and should be assessed (and who sets these standards), the degree of scalability and generalisation of proposed measures, the absolute and relative costs of introducing such measures, and obviously the validity and reliability of the methods used in evaluation.

UK universities traditionally relied on students' academic performance for measuring quality of its degree provision (Richardson, 2008). Later surveys like the Graduate Employment Survey, which was developed to assess employability of graduate students graduating from different universities (Shah, Pell, & Brooke, 2004), and National Students Survey (NSS, Richardson, 2013), which assesses opinions of students about their experience of a HE degree courses, were developed to go beyond simple learning outcomes and degree classifications. Survey instruments like the NSS are commonly used to rank universities (Richardson, 2013), whereby there is a recent push towards implementing NSS as a longitudinal instrument across the various years of study (Havergal, 2019). However, questions remain whether such measures are adequate representations of actual excellence of teaching, students' learning, skills and/or personal development. Indeed, a wide range of academics have challenged the validity and suitability of those surveys in assessing quality of degree courses at different universities (e.g., Langan & Harris, 2019; Richardson, 2013; Rienties & Toetenel, 2016). For example, in a recent review of 1.8 million NSS returns by Langan and Harris (2019) showed an increasing similarity between institutions in terms of overall satisfaction, leading to questions about validity and usefulness to distinguish good from excellent teaching.

One alternative way of assessing the 'value' of HE is to look at learning gains. Learning gain is a commonly used term that refers to change in students' knowledge and skills in relation to desired learning outcomes (Boud, 2018; Evans et al., 2018; Hake, 1998; McGrath et al., 2015). In the last 10 years there has been an emergence of learning gains

literature, mainly from the US (Arum & Roksa, 2011; Cahill et al., 2014; Hake, 1998; Pascarella & Blaich, 2013) but increasingly also from the UK (Evans et al., 2018; Rogaten & Rienties, 2018; Rogaten, Rienties, & Whitelock, 2017). Although the construct of learning gains has been widely used in educational research over the years, to the best of our knowledge only one systematic review has critically analysed and reflection on learning gains research findings (Rogaten, Rienties, et al., 2019). This systematic literature review of 51 studies with 41K students indicated a rich but diverse variety of adopted methodologies and approaches were used by researchers attempting to "measure" learning gains. In particular, Rogaten, Rienties, et al. (2019) found a lack of consistency in the ways in which learning gains were measured and reported. These inconsistencies and limitations might hamper any attempt to make effective comparisons between teaching excellence and learning gains. These inconsistencies also confirm previous findings from an influential US study where large differences between credentials and changes in critical thinking could be observed (Arum & Roksa, 2011, 2014).

Measuring learning gains

Our review of the existing research on learning gains first and foremost outlined a vast range of methods used for capturing learning gains (Rogaten, Rienties, et al., 2019). Learning gains research ranges from the use of the standardised tests in a pre-post repeated measures design to the use of the self-reported measures of learning gains in the cross-sectional research. This section we will mainly focus on pre-post test design and summarise and examine their strengths and shortcomings.

As identified by Rogaten, Rienties, et al. (2019), pre-tests and post-tests are a standard method and arguably one of the most appropriate ways of measuring change or gain. In educational research pre-test and post-tests are commonly used to evaluate effectiveness of any particular course (Cahill et al., 2014; Hake, 1998; Roohr et al., 2017), or change in the design of a course (teaching method). Rogaten, Rienties, et al. (2019) showed that most of the learning gains studies using pre-test and post-tests for measuring learning gains could be separated into those that use pre-test and post-test on one group of students (e.g., class, cohort), and those that compare two or more groups of students (group taught through traditional lectures and group taught in a non-traditional way e.g., active learning, blended learning, problem based learning).

According to the review conducted by Rogaten, Rienties, et al. (2019) pre-post tests were the most common method used. Out of 51 empirical studies measuring learning gains,

36 studies totalling 79 student samples (70% of all student samples; (e.g., Andrews, Leonard, Colgrove, & Kalinowski, 2011) used a pre-post test design to assess learning gains. Furthermore, the systematic review identified 23 studies where a comparison between two or more samples of students were made (e.g., Hill, Sharma, & Johnston, 2015; Roohr et al., 2017) totalling 64% of all student samples. Thus, by far most of the quantitative research on learning gains favours this quasi-experimental design to demonstrate the benefits of any one particular way of learning.

In terms of the actual tests used in this quasi-experimental learning gains research one of the most common approaches are the use of multiple choice tests with correct and incorrect answers (e.g., Hill et al., 2015; Pentecost & Barbera, 2013), such as the Chemical Concepts Inventory (CCI) or the Force and Motion Concept Evaluation test (FMCE). For example, Pentecost and Barbera (2013) used the Chemical Concepts Inventory consisting of 22 multiple choice questions amongst 2,392 undergraduate chemistry students and found relatively low learning gains across four universities, ranging from 0.04 to 0.14. Although standardised test such as CCI or FMCE provide a quite reliable way of assessing knowledge, the validity of these types of test should be challenged. Learning should be considered as a transformational experience, and not a mere acquisition of facts and as such, more authentic and encompassing ways of assessment should be considered (Evans et al., 2018; Hughes, 2018).

While there is now a wide established body of literature using standardised tests, there is some emergent literature using other forms of tests in a pre-post test format. Indeed Rogaten, Rienties, et al. (2019) found 18 studies who aimed to do just that and used self-reported measures of learning gains. These self-reported surveys of learning gains commonly use a combination of Likert-response scale items and open questions. The self-reported measures of learning are used either as stand-alone measures or in the combination with the multiple choice tests when knowledge alone may not be sufficient to judge what learning progress students are making.

For example, in a biology laboratory class of 38 students Beck and Blumer (2012) measured their students' confidence in addition to knowledge in designing an experiment using a 12-item self-reported survey in a pre-post manner. Using a similar approach, (Mathabathe & Potgieter, 2014) measured 91 STEM students' knowledge and confidence in

stoichiometry[1] alongside with their confidence in their knowledge and found that overall students' confidence in their knowledge of stoichiometry improved. However, a substantial number of students were overconfident in their self-reported scores, and perhaps surprisingly their overconfidence increased over time, whereas realistic confidence (i.e., confidence corresponding with the level of knowledge) decreased as a result of the course (Mathabathe & Potgieter, 2014).

A less commonly used method for assessing learning gains is the use of self-reported surveys in the cross-sectional design studies. Rogaten, Rienties, et al. (2019) identified 18 studies that used self-reported measures of learning gains in cross-sectional design to assess learning gains. There are number of instruments and approaches that were developed specifically for this type of research, which furthermore may be discipline specific or non-discipline specific. For example, Learning Attitudes about Science Survey (CLASS) (e.g., Cahill et al., 2014; Gok, 2012) is primarily used for students studying science courses like physics. The non-discipline specific survey is Student Assessment of Learning Gains survey (SALG) (e.g., Gill & Mullarkey, 2015; Ojennus, 2016) that is commonly used with students enrolled in science courses like chemistry and biology as well as psychology courses, human nutrition, computer science and information technology. However, it is not uncommon for the researchers to develop their own measures to capture context specific skills and abilities (e.g., Liu, Liu, & Chi, 2014; Matthews, Adams, & Goos, 2015).

Obviously, there are several limitations associated with both pre-post tests and self-reported measures. In relation to the pre-post testing, the first limitation that should be taken into account is whether the tests used are the same tests at the pre- and the post-test stages. If the same test or similar questions was administered twice, by default students will always perform better at the post-test than at the pre-test just as a result of the mere exposure to the testing environment. This is not just the case for the knowledge tests, as these findings are also quite common with the other skills assessments that require practice and attention. As such, when interpreting the findings of the studies that used identical assessments for the pre-test and the post-test we should consider how much of the improvement can be attributed to the actual learning and how much of the improvement is just due to the exposure. To avoid

[1] Stoichiometry is a section of chemistry that is focusing on the ways of determining exact proportions of elements to make pure chemical compounds, alloys or ceramic crystals

the negative effects of completing the same test twice, one can choose two different tests, but the issue of comparability of the test difficulty should be addressed as well as attempts of removing the order effect. Furthermore, the direct comparison of the results of pre-test and post-test may produce less reliable learning gains, but the computational aspect of the pre-post test research will be further discussed later on in the chapter.

In addition to the practical design and selected measures limitations there are practical limitations to such research. Mainly learning gain studies can be conducted on a relatively limited sample of students, and if they are not forming part of the assessment practice students are at risk of being over assessed. In addition, the time span between the pre- and post-test should be considered, as most of the studies reviewed by Rogaten, Rienties, et al. (2019) only looked at learning gains that occurred within couple or weeks or within a semester. All these practical limitations pose questions about the scalability of such research.

Furthermore, a limitation of self-reported measures of learning gains in comparison to standardised tests is the reliability of self-reported measures. It is possibly the single most detrimental shortcoming of learning gains research that uses surveys in a pre-post test settings as well as in cross-sectional studies. The rating of students' learning gains is always benchmarked against their 'feeling' of learning or 'feeling of knowing'. This may be a rather problematic issue to address. One interesting study that tried to address the issue of benchmarking and incorporate pre-post self-reported responses in a cross-sectional manner was done Douglass et al. (2012). The authors asked students to self-report their level of knowledge and skills for when they started the course, as well as when they finished the course. The benchmark for judging the learning gain in this case was the same state of knowledge, and as such the difference between the two points was a measure of learning gains. Their results demonstrated much higher self-reported learning gain than was observed in any other research that used self-reported measures or pre-post tests. The administration of such surveys is a feasible, although costly, way of assessing learning gains on a larger scale, but it still does not elevate the concern over the objectivity of self-reported measures.

One alternative approach to measuring self-reported or objective learning gains is to use pre-existing grades as a pre-test, and subsequent grades as potential proxies for a post-test metric. Although this measure heavily relies on the validity of the assessments used, this data is readily available in most of the institutions and can be used to better understand students' academic progress as a proxy for learning gain. For example, Rogaten and Rienties (2018) used

longitudinal academic performance data of 4,222 first year STEM students across 10 modules and analysed using both pre-post testing as well as multilevel growth-curve modelling. As illustrated in Table 1, it would be relatively straightforward to calculate the pre-post test scores, or take into consideration the scores on various intermediary tests. However, as illustrated in Figure 1 even though the learning gains of these 10 modules would seem easy to calculate, there is substantial diversity in the actual grading trajectories.

➔ Insert Table 1 and Figure 1 about here

Computation of learning gains

Probably one of the challenges in the learning gain research that uses pre-test and post-test is the computation of learning gain (Baume, 2018; Linn & Slinde, 1977; Lord, 1956; Pike & Killian, 2001; Zimmerman & Williams, 1982). The computation of learning gain or change was actively debated in areas like psychology and education since the 1950s (e.g., Lord, 1956). On the one hand, if one wants to examine the level of knowledge students developed on a course, one would assume that subtracting beginning of a semester knowledge test score from end of a semester knowledge test score will produce an accurate level of change/gain in academic achievement. This way of computing gain is referred in a literature as raw gain (Baume, 2018; Pampaka et al., 2018). Although this computation of a gain makes sense, it has number of limitations and was criticised repeatedly for not being an accurate representation of gain/change.

There are three main considerations that have to be taken into account when computing learning gains and interpreting research findings. Firstly, looking at the raw gain as a value of gain may be inaccurate due to the difference between scores being less reliable than scores themselves, i.e., raw gain represents compound error of pre-test and post-test. For example Lord (1956) argued that the representativeness of learning gain as a delta score is only valid in case of perfectly reliable tests, which as argued by Boud (2017); (2018) rarely happens in HE contexts. Furthermore, in psychological and educational testing measurement errors are bound and therefore should be taken into account during the computation (Cronbach & Furby, 1970; Dimitrov & Rumrill Jr, 2003; Pike, 1992; Pike & Killian, 2001).

Secondly, looking at the raw change is misleading because that change depends on the starting point of the performance continuum “Useful comparisons of the gains of students

who start out at quite different parts of the score scale require either an arbitrary assumption or an empirical demonstration that numerically equal intervals in different parts of the score scale are actually “equal" in some meaningful and useful sense” (Lord, 1956, pp. 19-20).

Thirdly, raw change in scores between pre-test and post-test assumes a linear relationship between test scores and ability when actually most of the time this relationship is non-linear. Thus, the raw learning gain on a simple test will be higher for a “low-ability” student and lower for a “high-ability" student, whereas the raw learning gain on a difficult test will be higher for a high-ability student and lower for a low-ability student. In other words, the magnitude of change depends on the difficulty of the test (Fischer, 1976).

Given these potential inaccuracies, over the years multiple mathematical solutions were created aiming to more accurately compute gain. A number of these were specifically developed for and tested in educational settings providing measures of learning gain. The first attempt to compute learning gain taking into account measurement errors was to compute a *true gain*. The true gain is based on a linear regression procedure where the true gain for each individual is the difference between group mean pre-test score and group mean post-test score assuming the reliability of estimates is satisfactory i.e., the pre-test post-test variances and reliability are equal (Lord, 1956). However, when these assumptions are violated the reliability of raw gain scores is actually high and therefore use of raw gain scores produces would be a more accurate representation of a gain that true gain (Zimmerman & Williams, 1982).

Building on linear regression procedure for measuring true gain, a second option is to compute *residual gain* (Cronbach & Furby, 1970). The computation of residual gain is compatible with the raw gain. The advantage of this computation is that it removes the change from the post-test scores that are predicted from the pre-test (Linn & Slinde, 1977). Although, residual gain allows to identify individuals that showed more or less than expected gain (i.e., are superior of inferior learners) residuals do not represent change as such, they only represent what was not predicted linearly (Baird, 1988). As they are residuals (essentially deviations) half of students will by default be above the mean and half bellow which makes judgement on the effectiveness of learning inappropriate (Pike, 1992).

A third option to compute gain was *normalised gain* (Hake, 1998). As indicated by Rogaten, Rienties, et al. (2019) normalised gain is a commonly used measure of learning gain. The main advantage of using normalised gain is that it solves the problem of ceiling effect

which occurs when students reach the possible maximum score. Thus, tests with the ceiling effect have potential bias towards strong students (high pre-testers) by using the difference between the maximum test score and pre-test score as denominator. Thus, normalised gain demonstrates realised gain to the maximum of possible gain.

A fourth option is to use *average normalised gain*, which can be computed using either individual scores or group means for pre-test and post-test (Bao, 2006). Although in both normalised and average normalised gain the same principles are used, the two methods may yield different results for the same sample. This is due to the asymmetrical distribution of differences between low and high scoring pre-testers on their post-tests, or scoring lower on the post-test than in the pre-test (for a review, see Bao, 2006). In addition, the normalised gain cannot be computed for individuals who scored absolute maximum on the pre-test scores, while the average normalised gain cannot be computed using individual scores if any one person scored maximum on the pre-test (Marx & Cummings, 2007).

In cases where post-test scores are lower than pre-test scores computation of *normalised change* is more meaningful (Marx & Cummings, 2007). The normalised change has advantage over normalised gain in cases of negative gain by using analogous computation where observed loss is the ratio of possible maximum loss. However, this method does not apply to those students who scored the possible maximum or possible minimum on both pre-test and post-test (Marx & Cummings, 2007). In addition, normalised change ranges from -1 to +1 and removes low pre-test scores bias. Thus, students who score 90% on a pre-test can obtain a change ranging from -1 to +1, and the same is the case for students who scored 50% or 20%. As such, it is much easier and more intuitive to interpret normalised change rather than normalised gain. Marx and Cummings (2007) further argued that normalised change is suitable for computation if the pre-test and post-test are not the same. However, it is important to understand that when averaging gains or losses to a group, both gains and losses are relative only to the maximum possible gain i.e., the result will show more gain than loss.

For example, Cahill et al. (2014) assessed the learning gains of 1100 students across three years who were studying Physics. The aim of the study was to examine the effectiveness of the Interactive-Engagement teaching technique. In their study Cahill et al. (2014) used both normalised gain and normalised change measures to assess the magnitude of learning gain. The normalised gain was used for students who scored higher at the post-test than at the pre-test, and the normalised change was used for students who showed a decrease in scores from the pre-test to post test. Because the majority of students showed an increase in their scores

the difference between the normalised gain scores and normalised change scores was negligible.

Another problem with computing change is the phenomenon of regression to the mean. Regression to the mean essentially masks true change over time and is largely due to poor reliability of pre-test and post-test materials and random measurement error (Campbell & Cook, 1979; Rocconi & Ethington, 2009). There are a number of ways proposed to address the regression to the mean phenomenon, but none of them eliminate it completely. One needs to examine whether there is a regression to the mean by looking at the relationship between change and pre-test scores. If the correlation is negative, an adjustment to the pre-test score is needed. Once the pre-test scores adjusted they should be used in the further analysis of computing gains (for a review, see Rocconi & Ethington, 2009).

Comparing different groups' learning gains

In cases when the research design involves comparison between two or three groups, the most commonly used methods for computing learning gains from pre-test and post-test scores are: analysis of variance (ANOVA) and analysis of covariance (ANCOVA). The advantage of ANOVA and ANCOVA is in that they both reduce error variance and increase the power of the test (Sörbom, 1976). There are four ways of data analysis commonly used in research that has both within and between participants observations i.e., mixed design research. The preferred method is to use ANCOVA on pre-test and post-test scores or ANOVA on raw gain scores. The two least favourable analyses are ANOVA on residuals and repeated measures ANOVA (Dimitrov & Rumrill Jr, 2003). In ANCOVA pre-test scores are used as covariates of post-test scores because this reduces error variance by adjusting post-test means to the pre-test. ANCOVA will produce reliable results when the assumptions of linear relationship, randomization and homogeneity of variances are met. A same assumption should be met for ANOVA on raw gains analysis. However, ANCOVA is more powerful than ANOVA and more flexible on the assumption of linear relationship. ANCOVA will also produce more accurate computation of gain if the relationship between pre-test and post-test has quadratic or cubic component, or if the regression slopes are not equal the Johnson-Neyman technique can be used for the regions of significance (Cahen & Linn, 1971; Dimitrov & Rumrill Jr, 2003).

The computation of the gain using ANOVA on residuals is considered not a viable option as the results will most certainly show overestimated level of significance when residuals are obtained from the full model (pooled within group regression coefficient) and being too conservative when regression coefficient is based on the restricted model (all observations are

combined into one group) (Maxwell, Delaney, & Manheimer, 1985). The repeated measures ANOVA on two of more groups (mixed ANOVA) is also considered to be inferior to ANCOVA. The main criticism is with the reporting and interpretation of the difference between the groups i.e., main effect of between participants factor. The main effect of a group is usually reported and interpreted incorrectly as the main effect of group is based on the mean of pre-test and post-test for each group. For example, if one group was a control and another group received an intervention, at the base line level (pre-test) two groups will be very similar (assumed randomisation). The effect of the intervention therefore will be the difference between the two groups at the post-test only as intervention possibly could have no effect on pre-test scores. The bases for computing main effect for between-participants factor (group) is to compute an overall mean for both groups. The overall mean would include both pre-test and post-test scores. As such, the difference between groups (effect of intervention) is only partially presented by the main effect for between participants factor. A better representation of the main effect would actually be an interaction (Cronbach & Furby, 1970; Huck & McLean, 1975). In all, all four ways of data analysis have limitations, but results obtained from ANCOVA seems to be most accurate for the comparison of learning gains between two or more groups.

Conclusions and moving forwards

To sum up the core points this chapter of data collection and analysis there are two main constrains associated with the research on learning gains. Firstly, based upon our systematic review (Rogaten, Rienties, et al., 2019) in the overwhelming majority of the studies on learning gains researchers used either tests or self-report surveys for data collection, which requires an additional effort on behalf of the researchers and students. This usually results in a biased data collection with the most engaged students taking part. We proposed to address this limitation associated with the additional data collection and self-selection biases in a sample by using existing data on students' assessments that is readily available in each university (Rogaten & Rienties, 2018; Rogaten et al., 2017).

Secondly, the learning gain is mainly computed looking at the gain between two points in time. This only allows to focus on learning gain that occurs in a short time period and usually does not capture the full extent of learning that occurs as a results of HE. In our own research we have attempted to remove this limitation by conducting an analysis across several points in time looking at the progress students make over the years. Thus, in our own research we

have used was multilevel modelling on longitudinal data (Rogaten & Rienties, 2018; Rogaten et al., 2017).

With regards to the data analysis, multilevel modelling has a number of advantages over other analyses techniques described earlier in the chapter. Firstly, all of the described earlier methods assume that pre-test and post-test observations from one participant are independent from pre-post-test observations of another participant. In a context of HE research this assumption is usually violated, as students who study in the same class, same subject and taught and assessed by the same lecturer/teacher have more in common regardless of individual differences, and as such share similar experience . Therefore their error variance in the performance is correlated/shared (Snijders & Bosker, 2012). Multilevel modelling allows researcher to address this limitation by nesting the error variance at different levels of the hierarchy. Thus, in our research we used the three-level hierarchy that allowed us to estimate individual students' learning trajectories in a context of the average course trajectory. Furthermore, multilevel modelling allows to look at these trajectories taking into account individual students' characteristics such as socio-demographic factors (Nguyen, Rienties, & Richardson, 2019; Rogaten, Clow, Edwards, Gaved, & Rienties, 2019).

References

Andrews, T. M., Leonard, M. J., Colgrove, C. A., & Kalinowski, S. T. (2011). Active Learning Not Associated with Student Learning in a Random Sample of College Biology Courses. *CBE-Life Sciences Education, 10*(4), 394-405. doi:10.1187/cbe.11-07-0061

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

Arum, R., & Roksa, J. (2014). *Aspiring adults adrift: Tentative transitions of college graduates*. Chicago: University of Chicago Press.

Ashwin, P. W. H. (2017). *Making sense of the Teaching Excellence Framework (TEF) results*. Retrieved from Lancaster: http://eprints.lancs.ac.uk/86901/1/making_sense_of_the_tef.pdf

Baird, L. (1988). Value added: Using student gains as yardsticks of learning. In *Performance and Judgement: Essays on Principles and Practice in the Assessment of College Student Learning,* (pp. 205-216). Washington, DC: US Government Printing Office.

Bao, L. (2006). Theoretical comparisons of average normalized gain calculations. *American Journal of Physics, 74*(10), 917-922. doi:10.1119/1.2213632

Baume, D. (2018). Towards a measure of learning gain. A journey. With obstacles. *Higher Education Pedagogies, 3*(1), 51-53. doi:10.1080/23752696.2018.1467213

Beck, C. W., & Blumer, L. S. (2012). Inquiry-based ecology laboratory courses improve student confidence and scientific reasoning skills. *Ecosphere, 3*(12), 1-11. doi:10.1890/ES12-00280.1

Boud, D. (2000). Sustainable Assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151-167. doi:10.1080/713695728

Boud, D. (2017). Standards-Based Assessment for an Era of Increasing Transparency. In D. Carless, S. Bridges, C. Chan, & R. Glofcheski (Eds.), *Scaling up Assessment for Learning in Higher Education. The Enabling Power of Assessment* (Vol. vol 5). Singapore: Springer.

Boud, D. (2018). Assessment could demonstrate learning gains, but what is required for it to do so? *Higher Education Pedagogies, 3*(1), 54-56. doi:10.1080/23752696.2017.1413671

Cahen, L. S., & Linn, R. L. (1971). Regions of Significant Criterion Differences in Aptitude-Treatment-Interaction Research. *American Educational Research Journal, 8*(3), 521-530. doi:10.3102/00028312008003521

Cahill, M. J., Hynes, K. M., Trousil, R., Brooks, L. A., McDaniel, M. A., Repice, M., . . . Frey, R. F. (2014). Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum. *Physical Review Special Topics - Physics Education Research, 10*(2), 020101. doi:10.1103/PhysRevSTPER.10.020101

Campbell, D. T., & Cook, T. D. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago: Rand McNally College Publishing Company

Cronbach, L. J., & Furby, L. (1970). How we should measure" change": Or should we? *Psychological Bulletin, 74*(1), 68. doi:10.1037/h0029382

Department of Business, I., and Science. (2015). *Fulfilling our Potential: Teaching Excellence, Social Mobility and Student Choice* (Cm 914). Retrieved from London:

Dimitrov, D. M., & Rumrill Jr, P. D. (2003). Pretest-posttest designs and measurement of change. *Work, 20*(2), 159-165.

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*(3), 317-335. doi:10.1007/s10734-011-9496-x

Evans, C., Kandiko Howson, C., & Forsythe, A. (2018). Making sense of learning gain in higher education. *Higher Education Pedagogies, 3*(1), 1-45. doi:10.1080/23752696.2018.1508360

Everson, K. C. (2017). Value-Added Modeling and Educational Accountability. *Review of Educational Research, 87*(1), 35-70. doi:10.3102/0034654316637199

Fischer, G. H. (1976). Some probabilistic models for measuring change. In D. N. D. Gruijter & L. J. T. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement* (pp. 97–110). New York: Wiley.

Forsythe, A., Evans, C., Kandiko Howson, C., & Edwards, C. (2018). Learning gain: political expedient or meaningful measurement? Retrieved from https://www.timeshighereducation.com/features/learning-gain-political-expedient-or-meaningful-measurement

Gill, T., & Mullarkey, M. (2015). Taking a Case Method Capstone Course Online: A Comparative Case Study. *Journal of Information Technology Education, 14*.

Gok, T. (2012). The Impact of Peer Instruction on College Students' Beliefs About Physics and Conceptual Understanding of Electricity and Magnetism. *International Journal of Science and Mathematics Education, 10*(2), 417-436. doi:10.1007/s10763-011-9316-x

Hake, R. 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), 64-74. doi:10.1119/1.18809

Havergal, C. (2019, 21 October 2019). All-years National Student Survey 'huge burden' on universities. *Times Higher Education*. Retrieved from

https://www.timeshighereducation.com/news/all-years-national-student-survey-huge-burden-universities
Hill, M., Sharma, M. D., & Johnston, H. (2015). How online learning modules can improve the representational fluency and conceptual understanding of university physics students. *European Journal of Physics, 36*(4), 045019.
Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin, 82*(4), 511-518.
Hughes, C. (2018). How do we measure what students learn at university? 30 September 2019.
Kandiko Howson, C. (2019). To say the Office for Students (OfS) Learning Gain Programme ended with a whimper may be an overstatement. 22 July 2019.
Langan, A. M., & Harris, W. E. (2019). National student survey metrics: where is the room for improvement? *Higher Education*. doi:10.1007/s10734-019-00389-1
Linn, R., & Slinde, J. (1977). The Determination of the Significance of Change Between Pre- and Posttesting Periods. *Review of Educational Research, 47*(1), 121-150. doi:10.3102/00346543047001121
Liu, H., Liu, J., & Chi, X. (2014). Regulatory mechanism of self-determination involvement in higher education: assessing Chinese students' experiences. *Higher Education, 67*(1), 51-70. doi:10.1007/s10734-013-9640-x
Lord, F. M. (1956). The measurement of growth. *ETS Research Bulletin Series, 1956*(1), i-22. doi:10.1002/j.2333-8504.1956.tb00058.x
Marx, J. D., & Cummings, K. (2007). Normalized change. *American Journal of Physics, 75*(1), 87-91. doi:10.1119/1.2372468
Mathabathe, K. C., & Potgieter, M. (2014). Metacognitive monitoring and learning gain in foundation chemistry. *Chemistry Education Research and Practice, 15*(1), 94-104.
Matthews, K. E., Adams, P., & Goos, M. (2015). The Influence of Undergraduate Science Curriculum Reform on Students' Perceptions of their Quantitative Skills. *International Journal of Science Education, 37*(16), 2619-2636. doi:10.1080/09500693.2015.1096427
Maxwell, S. E., Delaney, H. D., & Manheimer, J. M. (1985). Anova of Residuals and Ancova: Correcting an Illusion by Using Model Comparisons and Graphs. *Journal of Educational Statistics, 10*(3), 197-209. doi:10.3102/10769986010003197
McGrath, C. H., Guerin, B., Harte, E., Frearson, M., & Manville, C. (2015). *Learning gain in higher education*. Retrieved from Cambridge: www.rand.org/t/rr996/
McKie, A. (2018). Standardised tests measuring learning gain fail to make the grade. Retrieved from https://www.timeshighereducation.com/news/standardised-tests-measuring-learning-gain-fail-make-grade
Nguyen, Q., Rienties, B., & Richardson, J. T. E. (2019). Learning analytics to uncover inequality in behavioural engagement and academic attainment in a distance learning setting. *Assessment and Evaluation in Higher Education*. doi:10.1080/02602938.2019.1679088
Ojennus, D. D. (2016). Assessment of learning gains in a flipped biochemistry classroom. *Biochemistry and Molecular Biology Education, 44*(1), 20-27. doi:10.1002/bmb.20926
Pampaka, M., Swain, D., Jones, S., Williams, J., Edwards, M., & Wo, L. (2018). Validating constructs of learners' academic self-efficacy for measuring learning gain. *Higher Education Pedagogies, 3*(1), 118-144. doi:10.1080/23752696.2018.1454264

Pascarella, E. T., & Blaich, C. (2013). Lessons from the Wabash National Study of Liberal Arts Educaton. *Change: The Magazine of Higher Learning, 45*(2), 6-15. doi:10.1080/00091383.2013.764257

Pentecost, T. C., & Barbera, J. (2013). Measuring Learning Gains in Chemical Education: A Comparison of Two Methods. *Journal of Chemical Education, 90*(7), 839-845. doi:10.1021/ed400018v

Pike, G. R. (1992). Lies, damn lies, and statistics revisited a comparison of three methods of representing change. *Research in Higher Education, 33*(1), 71-84. doi:10.1007/bf00991972

Pike, G. R., & Killian, T. S. (2001). Reported Gains in Student Learning: Do Academic Disciplines Make a Difference? *Research in Higher Education, 42*(4), 429-454. doi:10.1023/a:1011054825704

Richardson, J. T. E. (2008). *Degree attainment, ethnicity and gender: a literature review*. Retrieved from York, UK: www.heacademy.ac.uk/assets/York/documents/ourwork/research/J_Richardson_literature_review_Jan08.pdf

Richardson, J. T. E. (2013). The National Student Survey and its Impact on UK Higher Education. In M. Shah & C. S. Nair (Eds.), *Enhancing Student Feedback and Improvement Systems in Tertiary Education* (Vol. 5, pp. 76–84). Abu Dhabi, UAE: Commission for Academic Accreditation.

Rienties, B., & Toetenel, L. (2016). The impact of learning design on student behaviour, satisfaction and performance: a cross-institutional comparison across 151 modules. *Computers in Human Behavior, 60*, 333-341. doi:10.1016/j.chb.2016.02.074

Rocconi, L. M., & Ethington, C. A. (2009). Assessing Longitudinal Change: Adjustment for Regression to the Mean Effects. *Research in Higher Education, 50*(4), 368-376. doi:10.1007/s11162-009-9119-x

Rogaten, J., Clow, D., Edwards, C., Gaved, M., & Rienties, B. (2019). Are assessment practices well aligned over time? A big data exploration. In M. Bearman, P. Dawson, R. Ajjawi, J. Tai, & D. Boud (Eds.), *Re-imagining University Assessment in a Digital World*. Dordrecht: Springer.

Rogaten, J., & Rienties, B. (2018). Which first-year students are making most learning gains in STEM subjects? *Higher Education Pedagogies, 3*(1), 161-172. doi:10.1080/23752696.2018.1484671

Rogaten, J., Rienties, B., Sharpe, R., Cross, S., Whitelock, D., Lygo-Baker, S., & Littlejohn, A. (2019). Reviewing affective, behavioural, and cognitive learning gains in higher education. *Assessment & Evaluation in Higher Education, 44*(3), 321-337. doi:10.1080/02602938.2018.1504277

Rogaten, J., Rienties, B., & Whitelock, D. (2017). Assessing Learning Gains. In D. Joosten-ten Brinke & M. Laanpere (Eds.), *Technology Enhanced Assessment. TEA 2016. Communications in Computer and Information Science* (Vol. 653, pp. 117-132). Cham: Springer.

Roohr, K. C., Liu, H., & Liu, O. L. (2017). Investigating student learning gains in college: a longitudinal study. *Studies in Higher Education, 42*(12), 2284-2300. doi:10.1080/03075079.2016.1143925

Shah, A., Pell, K., & Brooke, P. (2004). Beyond First Destinations:Graduate Employability Survey. *Active Learning in Higher Education, 5*(1), 9-26. doi:10.1177/1469787404040457

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2 ed.). London: SAGE.

Sörbom, D. (1976). *A statistical model for the measurement of change in true scores*. New York: John Wiley & Sons.

Turner, R., Sutton, C., Muneer, R., Gray, C., Schaefer, N., & Swain, J. (2018). Exploring the potential of using undergraduates' knowledge, skills and experience in research methods as a proxy for capturing learning gain. *Higher Education Pedagogies, 3*(1), 222-248. doi:10.1080/23752696.2018.1449127

Universities UK. (2004). *Measuring and recording student achievement: Report of the Scoping Group chaired by Professor Robert Burgess*: Universities UK.

Varsavsky, C., Matthews, K. E., & Hodgson, Y. (2014). Perceptions of Science Graduating Students on their Learning Gains. *International Journal of Science Education, 36*(6), 929-951. doi:10.1080/09500693.2013.830795

Zimmerman, D., & Williams, R. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement, 19*(2), 149-154. doi:10.1111/j.1745-3984.1982.tb00124.x

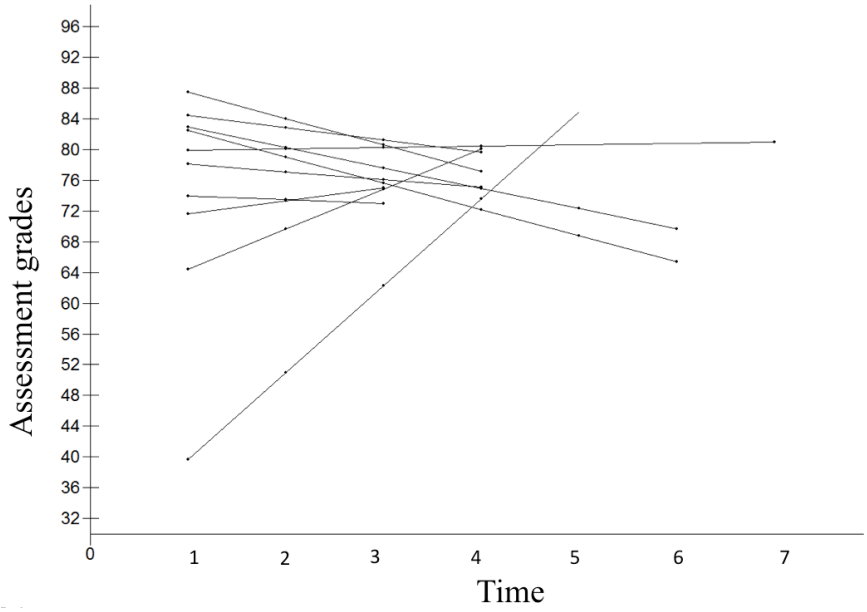
Table 1: Mean and standard deviations across all continuous assessments for each of the 10 level-1 modules

Module	*First Assessment*		*Final Assessment*		*Average Continuous assessment score*	
	M	*SD*	*M*	*SD*	*M*	*SD*
Module 1	88.0	9.7	78.5	15.7	82.4	14.4
Module 2	83.2	11.2	79.5	17.0	72.4	21.5
Module 3	81.2	10.1	71.5	16.4	74.3	18.1
Module 4	78.1	12.5	63.2	22.5	73.4	16.9
Module 5	85.0	11.0	85.6	11.2	76.7	13.4
Module 6	77.2	12.1	73.8	15.5	72.5	12.7
Module 7	74.8	10.8	71.4	14.3	82.5	13.4
Module 8	71.3	17.6	75.0	17.0	62.1	22.0
Module 9	54.9	22.0	73.6	19.5	80.7	14.6
Module 10	41.0	5.6	79.2	16.8	76.4	13.9

Note: the Tutor Marked Assessment (i.e., assignments) are marked on a scale from 0 to 100. The minimum passing mark is 40.

Source: Rogaten and Rienties (2018)

Figure 1. Learning gain trajectories across 10 STEM modules.



Source: Rogaten and Rienties (2018)