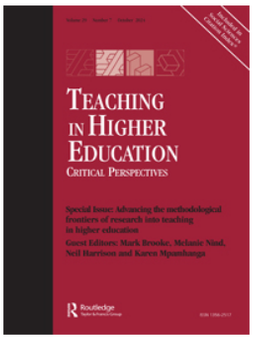

Taylor & Francis
Taylor & Francis Group

Teaching in Higher Education
Critical Perspectives

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cthe20

Measuring and mapping authentic assessment with a novel quantitative typology

Stephanie Baines, Pauldy Otermans, David Tree & Nicholas Worsfold

To cite this article: Stephanie Baines, Pauldy Otermans, David Tree & Nicholas Worsfold (06 Nov 2024): Measuring and mapping authentic assessment with a novel quantitative typology, Teaching in Higher Education, DOI: 10.1080/13562517.2024.2424823

To link to this article: https://doi.org/10.1080/13562517.2024.2424823

Published online: 06 Nov 2024.

TEACHING IN HIGHER EDUCATION
https://doi.org/10.1080/13562517.2024.2424823

Measuring and mapping authentic assessment with a novel quantitative typology

Stephanie Baines [a]*, Pauldy Otermans [a], David Tree [b] and Nicholas Worsfold[c]

[a]Psychology, Department of Life Sciences, Brunel University of London, London, UK; [b]Biosciences, Department of Life Sciences, Brunel University of London, London, UK; [c]Enviromental Sciences, Department of Life Sciences, Brunel University of London, London, UK

ABSTRACT

Authentic assessments are seen as a promising response to many of the challenges currently facing Higher Education. Studies have identified shared characteristics of authentic assessments, but it is also argued that the term is vague and subjective. Drawing on existing frameworks we have established a standardised measure to evaluate authenticity in the axes of product and process allowing graphical representation and mapping across programmes. We have mapped the assessments on two programmes in our institution. We show an appropriate increase of authenticity of assessment year-on-year in one programme and have identified assessments which have been altered to increase their authenticity in the other. Our tool will allow educators to estimate authenticity on their academic programmes, guide discussions and identify areas for improvement. A balanced mix of assessments in a programme ensures both skills development and knowledge requirements are fulfilled, with sustainability of assessment enabling increasingly authentic assessments over time.

DUTCH

Authentieke opdrachten worden gezien als een veelbelovende oplossing voor veel van de uitdagingen waar het hoger onderwijs momenteel voor staat. Studies hebben gedeelde kenmerken van authentieke opdrachten geïdentificeerd, maar het is ook vastgesteld dat de term vaag en subjectief is. Op basis van bestaande theorien hebben we een gestandaardiseerde maatstaf opgesteld om authenticiteit te evalueren in de gebieden van product en proces, waardoor we een grafische weergave kunnen geven op studie programme niveau. We hebben de beoordelingen in twee programma's binnen onze instelling in kaart gebracht. We zien een toename van de authenticiteit van opdrachten van jaar tot jaar in de programma's en hebben opdrachten geïdentificeerd die zijn aangepast om hun authenticiteit te vergroten. Onze tool stelt docenten in staat om de authenticiteit van hun academische

ARTICLE HISTORY
Received 14 March 2024
Accepted 20 September 2024

KEYWORDS
Programme-level approach; authentic assessments; quantitative mapping; sustainable assessments; graphical representations

Trefwoorden
Programmaniveau benadering; authentieke beoordelingen; kwantitatieve mapping; grafische weergave

CONTACT David Tree david.tree@brunel.ac.uk Biosciences, Department of Life Sciences, Brunel University of London, London, UK, UB8 3PH
*All authors contributed equally and are listed in alphabetical order.

programma's in te schatten, discussies te sturen en verbeterpunten te identificeren. Een gebalanceerde mix van opdrachten in een programma zorgt ervoor dat zowel de ontwikkeling van vaardigheden als de kennisvereisten worden vervuld.

Introduction

At the present time, demands on Higher Education (HE) are evolving at pace. Traditional methods of assessment face the challenges of generative artificial intelligence (AI), with subsequent threats to academic integrity and a need for greater student equity (Overono and Ditta 2023). In the UK and worldwide, HE is being asked to justify its existence in a marketplace of ideas (Atherton et al. 2023). The UK Government and the Office for Students (OfS) are placing demands on UK HE to emphasise and prioritise employability qualities. Their key demands include aligning education with the job market, measuring graduate outcomes, ensuring 'quality' education, and fostering entrepreneurial skills (Office for Students 2023). Additionally, a commitment to widening participation and promoting lifelong learning demands a shift towards a holistic, outcomes-focused approach in HE (Office for Students 2023). The context of the development of this work is a widening participation HE institution in the UK.

In this millieu, authentic assessments are a promising route to meet these evolving demands. As assessments shape learning and educational outcomes, the integration of authentic assessments into programmes of study can equip graduates not only with academic qualifications, but also practical skills, adaptability, and qualities essential for success in the job market. Authentic assessments thus address the needs of regulators, students, and employers (Sokhanvar et al. 2021). However, within HE scholarship, the concept of authenticity is far from universal, has a contested nature, and is said to reside in the eye of the beholder, subject to interpretation and context (McArthur 2023, Nieminen et al. 2023).

The term 'authentic assessment' was coined in the 1990s in direct distinction to traditional, standardised testing in the US school system (Wiggins 1990). Standardised tests are characterised by their focus on evaluating students' knowledge through methods that emphasise memorisation, rote learning, and the recall of information. They often involve multiple-choice questions, short-answer questions or traditional essays in which knowledge is largely disconnected from its context or application. In contrast, the authentic assessments described in embryonic form by Wiggins (1990) are more complex, meaningful tasks which emphasise application and integration of knowledge and critical thinking. According to Wiggins, 'authentic tasks involve "ill-structured" challenges and roles that help students rehearse for the complex ambiguities of the "game" of adult and professional life' (Wiggins 1990, 1).

Over the last 30 years the concept of authenticity of assessments has developed. In the context of HE, this means a shift towards assessing students through 'authentic' tasks which mimic those undertaken by professionals in the workplace often referred to as 'real world tasks'. The 1990s and 2000s saw a growing emphasis on constructivist educational theories, which suggest that learners actively construct their understanding of the world (Fosnot and Perry 1996). This form of active learning promotes more effective

retention of knowledge; thus tasks that promote this have become something of a gold standard in HE (Brame 2016, Felder and Brent 2009). Authentic assessments align with a constructivist perspective by encouraging students to engage in active learning and to demonstrate their understanding in contextually relevant ways. Another key shift was the role of assessment in promoting learning rather than just measuring it, encouraging ongoing feedback and reflection, allowing students to learn from their mistakes and improve their performance (Dann 2014). This all reflected a change in educational philosophy during this time toward more student-centred, contextually relevant, and outcomes-focused assessment practices. Besides this, there was a growing recognition that assessing what students can do in authentic contexts provides a more accurate and valuable measure of their readiness for the challenges of the post-graduation world of employment.

An influential concept in the assessment literature, which was developed in parallel with authentic assessments, is that of sustainable assessments. The concept of sustainable assessment (Boud 2000, Boud and Soler 2016) emphasises assessments that not only assess learning outcomes but also contribute to the ongoing development of students and the educational environment. Sustainable assessments are designed to be meaningful, relevant, and beneficial to students beyond merely assigning grades. In this way, they share many characteristics with authentic assessments in aiming to foster deep learning, critical thinking, and self-regulation skills, ultimately preparing students for success in the changing post-graduation world. Sustainable assessments serve multiple purposes, described as double duty, beyond evaluating student performance: they should provide opportunities for feedback that allow students to reflect on their learning experiences and enhance their long-term learning journey. If authentic assessments are to prepare students for their post-graduation lives, they should be structured within a programme of learning that develops the foundational skills learned in high school into the reflective and critical skills that will be crucial in the post-graduation lives.

Building on earlier foundations, recent research has conceptualised a number of characteristics that authentic assessments share (Ashford-Rowe et al. 2014, Gulikers et al. 2004, Ajjawi et al. 2024). A key focus has been on the nature of authenticity as vague (Ajjawi et al. 2020), with attempts to describe it being subjective. Critics argue that the term 'authentic assessment' has become a buzzword, losing its original meaning and with little meaningful operationalisation or application (McArthur 2023).

To address the ambiguities around authentic assessments, we wish to increase the transparency and consistency of the definition of levels of authenticity in assessments by quantifying key aspects of it. We wish to transition from abstract conceptualisation to the practical implementation of a measure of the authenticity of assessments at individual and programme level. To achieve this aim, we developed a standardised measure of authenticity based on previous studies. Gulikers et al. (2004) qualitatively described a five-dimensional framework of authenticity of assessment and Ashford-Rowe et al. (2014) identified eight critical elements of authentic assessments. To a very large extent these frameworks correspond, with the exception of metacognition included in the latter but not the former. Building on these qualitative works (Gulikers et al. 2004, Ashford-Rowe et al. 2014), we have designed a simple questionnaire to measure authenticity of assessment that is not restricted to any type of degree or assessment (Table 1).

The survey contains four questions which address the extent to which different aspects of the task resemble a task students might undertake in post-graduation life; four

Table 1. The 18 statements of the questionnaire, showing how these map onto the six subscales and the two dimensions.

2D	Subscale	Statement
Product	Task	Resembles authentic learning task but in a new situation
		Integrates knowledge skills and attitudes
		Resembles complexity of the real-life situation
		Resembles ownership of the task in real life
	Assessment result or form	Assessment is a quality product or performance that students can be asked to produce in real life
		Assessment should be a demonstration that permits making valid inferences about the underlying competencies
		Assessment is a full array of tasks and multiple indicators of learning in order to come to full conclusions
		Assessment requires students to present work to other people (oral or written form)
	Assessment criteria	Criteria are set and made explicit to learners beforehand
Process	Physical context	Assessment resembles reality
		Assessment resembles number and kind of resources available in real-life situation
		Time for assessment resembles that in the real-world situation (does not rely on unrealistic or arbitrary time constraints)
	Social context	Assessment resembles the social context in which the task takes place in real life
		Collaborative assessment includes social interaction
		Collaborative assessment includes positive interdependency
		Collaborative assessment includes individual accountability
		Individual assessment stimulates competition
Product/ Process	Metacognition	To what extent does the product/process involve metacognition? Examples of metacognitive activities include setting goals for learning, monitoring one's understanding of a topic, selecting appropriate strategies to tackle a problem, and evaluating the effectiveness of those strategies.

questions which address the extent to which different aspects of the resulting output of the assessment is resemble a task students might undertake in post-graduation lifeand one question about the assessment criteria. These are all aspects of the **product** of the assessment. The survey contains four questions about the social context, five about the social context and one question about the extent to which the assessment involves metacognition. These are all aspects of the **process** by which the students do the assessment.

The results of the survey allows us to condense the multiple conceptual aspects of authenticity down to two axes: The **product** of the assessment and the **process** through which the student produces the product. This produces a simple measure of how authentic an assessment is on each of the two most fundamental dimensions of **product** and **process**, which can be graphically displayed and compared (Figures 2 and 3). Crucially, this allows visual comparison of multiple years of assessment across a programme of study. It also allows a straightforward evaluation of the effects of changes to assessments on the pattern of assessment across a programme. The tool will enable other educators to assess the authenticity of the assessments on their programmes in a simple way, and to facilitate an open and informed dialogue about what authenticity means in different contexts.

In the development of this questionnaire, we started by reviewing at the level of individual assessments as is commonly done in the literature but, crucially extended this to examine the pattern of assessments throughout whole programme of study to also determine the sustainability of assessments. We do not make value judgements about individual assessments with different characteristics of authenticity or argue that traditional

Table 2. Intraclass correlation and description of assessment for the BSc Psychology.

Assessment	Intraclass correlation	Type of assessment
PY1A exam	.900	MCQ on campus exam 2 h
PY1Bi interview	.962	Mock job interview to panel of two (10 min + 5 min Q&A) + upload CV
PY1Bii written reflection	.862	Written reflection 1000 words + critique a ChatGPT output using track changes and comments
PY1Ci lab report 1	.914	Lab report on data collected 1200 words
PY1Cii exam 1	.832	MCQ on campus exam (with access to resources) 1 h
PY1Ciii lab report 2	.917	Lab report on data collected with reflection on use of feedback from lab report 1 1200 words
PY1Civ exam 2	.832	MCQ on campus exam (with access to resources) 1 h
PY1Cv experiment participation reflection	.777	Written reflection 200 words
PY1D exam	.860	MCQ on campus exam 2 h
PY1Ei formative essay	.914	Theory-based essay 1000 words
PY1Eii summative essay	.910	Theory-based essay 1500 words
PY2Ai poster	.259	Poster creation based on research article
PY2Aii lab report	.090	Lab report on data collected 1400 words
PY2Aiii exam	.743	MCQ and short-answer on-campus data analysis exam 3 h
PY2Aiv experiment participation reflection	.500	Written reflection 200 words
PY2Bi essay	.607	Theory-based essay 1500 words
PY2Bii reflective piece	.720	Written reflection and rubric construction 150 words + rubric
PY2C report	.266	Qualitative research report based on data collected 2000 words
PY2D essay	.297	Synoptic essay 2 × 500 words
PY2E essay	.354	Synoptic essay 1500 words
PY2F exam	.604	48-hour at-home exam
PY2Gi reflection*	.778	Written reflection on placement 1500 words
PY2Gii essay*	.790	Theory-based essay 1500 words
PY2Hi reflection*	.783	Written reflection on placement 1500 words
PY2Hii essay*	.790	Theory-based essay 1500 words
PY2Ii evidence file**	.839	Placement evidence file 3000 words
PY2Iii oral presentation**	.909	10–15-minutes presentation to panel of two + Q&A
PY2Iiii reflection**	.802	Written reflection on placement 1500 words
PY2Iiv essay**	.790	Theory-based essay 1500 words
PY3A dissertation	.352	Dissertation (empirical research project) 8000 words
PY3Bi portfolio	.802	Portfolio of answers to psychophysics practical questions
PY3Bii essay	.536	Theory-based essay 1500 words
PY3C exam	.713	MCQ and essay on campus exam 3 h
PY3Di essay	.508	Theory-based essay 1500 words
PY3Dii exam	.715	MCQ and short answer exam 2 h
PY3E essay	.620	Applied essay 2000 words
PY3F essay	.502	Theory-based essay 2000 words
PY3Gi essay	.791	Applied risk assessment based on case file 1000 words
PY3Gii exam	.625	MCQ on campus exam 1 h
PY3H essay	.466	Theory-based essay 2 × 1000 words
PY3I written coursework	.841	Applied written assignment (e.g. funding application bid) 2500 words
PY3J essay	.374	Theory-based essay 2500 words
PY3 K exam	.332	Short- and long-answer on campus exam 3 h
PY3L exam	.570	MCQ on campus exam 3 h
PY3M essay	.229	Applied essay 2000 words
PY3Ni essay	.710	Reflective essay 1500 words
PY3Nii reflective journal	.752	Reflective journal of mindfulness practise
PY3O essay	.484	Synoptic essay 1500 words
PY3P essay	.564	Synoptic essay 2 × 500 words
PY3Q exam	.538	48-hour at-home exam

*taken by 'thin' placement students only (2 × 6-month placements).
**taken by 'thick' placement students only (1 × 1-year placement).

assessments should be eliminated. Rather, we emphasise the importance of looking at assessments at a programme level and that a mixed pattern of assessments, with different characteristics of authenticity within a programme, is likely to benefit students. This enables them to develop a broad range of skills to benefit them in their later lives, makes them more competitive in the job market, and captures a more comprehensive view of students' abilities. But it is also important that students have their learning scaffolded and that assessments are sustainable so that students are supported in their journey through different aspects of authenticity through their university experience. Achieving this diversity and sustainability in assessment can only be achieved by taking a programme-level view of the totality of the assessment regime undertaken by a student.

Programme B

Assessments on Programme B are designed to progressively develop necessary knowledge and skills, building on the baseline with which students enrol (Table 3). The earlier education our incoming students receive focuses on traditional essays and exams, thus in their first year in Programme B, there is a preponderance of more familiar traditional assessments. These introduce students to university-level assessment in a less daunting way and complement the more authentic assessments students are not yet familiar with. To encourage critical, integrative thinking and reduce assessment load, there are a number of synoptic assessments in the first two years. These require students to use and integrate content from two modules to effectively complete the work: research methods and statistics, biological and cognitive psychology, social

Table 3. Intraclass correlation and description of assessment for the BSc Biomedical Sciences.

Assessment	Intraclass correlation	Type of assessment
BB1Ai presentation	.783	Group presentation
BB1Aii group essay	.926	Group-written literature review
BB1Aiii portfolio	.965	Portfolio of career development materials
BB1B micro report	.937	Written coursework
BB1C data MCQ	.939	Data Analysis MCQ
BB1D molecular report	.942	Written Coursework
BB1E exam 1	.943	Closed book exam: MCQ
BB1F synoptic 1	.987	Seen exam, at home, 2 h
BB2A career portfolio	.977	Portfolio of career development materials (CV, cover letter, PDP, critical reflection on career plans)
BB2B lit review	.963	2000-word report of literature review
BB2C paper CW	.960	2000-word report of data analysis of case study
BB2D poster CW	.972	Group (3) poster presentation (15 minutes) of data analysis of case study in front of 2 markers and group
BB2E exam 2	.980	Closed book exam: short essays and MCQ
BB2F synoptic 2	.994	Exam, subject known, 2 hours under exam conditions
BB3Ai presentation	.984	10-minute individual presentation in front of 15 others, 5 min questions
BB3Aii scientific report	.987	1500 word written data analysis report
BB3B reports	.995	1000 word written data analysis reports
BB3C exam 3	.995	Closed book exam, 1-hour essays
BB3D synoptic 3	.989	Exam, subject known, 2 h under exam conditions
BB3E dissertation	.974	Dissertation (empirical research project) 5000 words

psychology and individual differences. Furthermore, the degree is accredited, with the relevant accrediting body which stipulate a number of 'required' subjects that must be taught and assessed during the first and second academic year. In the final year, where content is more specialised, students choose four modules from a selection of 12 research-driven options determined by academic staff specialisms. The content and assessment of each of these modules is at the discretion of the academic lead. This allows students to choose the direction of specialism in what is a particularly broad academic field. The assessments across the programme are regularly reviewed by staff with an education focus, who work with module leads to refresh them according to changing needs of the students and developments in the field. This review process is now used to ensure authenticity is appropriately maintained in a manner that builds from less to more authentic as a student progresses through the years of their programme.

Materials and methods

The context of the study

The site of this study is a research-intensive university founded in 1966 and is home to about 18,000 students and 2,500 staff. It is a widening participation institution due to its commitment to increasing access to HE for individuals from diverse backgrounds, including those traditionally underrepresented in academia: low-income families, ethnic minority groups, mature learners, and those with disabilities. Over 90% of the home undergraduate students at the institution are from underrepresented groups. The institution offers a wide range of undergraduate degree courses, two of the most popular of which are the case studies to which the tool measuring authenticity of assessments has been applied and are referred to as Programme A and Programme B.

Programme A

The assessment strategy on Programme A was designed using Integrated Programme Assessment (IPA), which separates study and assessment allowing a holistic and overarching teaching and assessment strategy (Harvey et al. 2021, Table 2). Programme-level assessment has a number of highly positive aspects but is difficult to deliver (Charlton and Newsham-West 2024). This is a departure from a structure where teaching and assessment is contained within single modules, to one where assessment is separated from the teaching and staff are part of the programme team, not isolated in their modules (Charlton and Newsham-West 2024). To ensure assessment was fit for purpose (Brown 2005), a clear assessment rationale was developed to ensure students met the programme level learning outcomes. These were iterated down from the final year to the second year and the first year to ensure students entering university would be supported during every academic year to achieve the outcomes they would need by graduation. Assessments were developed by a team for each academic year and a vertical check between years verified coherence and progression. This ensured students were scaffolded to have the skills to meet learning outcomes

during each year while avoiding unnecessary repetition of tasks. Explicit feedback opportunities are incorporated into assessments to provide students with useable information on their performance to improve on subsequent related tasks. In this way, each assessment was designed to have a 'sustainable' element (Boud 2000), helping students to meet their learning needs for subsequent assessments. Because assessments are integrated, coursework assignments and exam questions are collectively designed, and marking is shared by teams creating a coherent programme assessment strategy and collegiate environment.

Materials

A 17-item questionnaire was devised based on the five-dimensional framework of Gulikers et al. (2004): Task, Assessment result/form, Assessment criteria, Physical context, and Social context. Given the increasing recognition of the importance of meta-cognition in authentic assessment (e.g. Ashford-Rowe et al. 2014), an additional item was included to measure this construct, producing an 18-item final questionnaire. Each item consisted of a statement, such as 'Resembles authentic learning task but in a new situation'. For the wording of each question see Table 1. Participants, educators involved in the teaching and assessment were required to rate their assessment in accordance with each statement using the scale 1: 'Not at all', 2: 'A little', 3: 'Somewhat', 4: 'A lot', 5: 'Extremely', with the option of using 'Not applicable' (given a value of 0, so as not to contribute to the overall score). Participants were instructed to use the questionnaire separately for each assessment they wanted to rate. The item ratings were used to calculate the score for each subscale by calculating the average of the items contributing to the subscale (Task items 1–4, Assessment result/form items 5–8, Assessment criteria item 9, Physical context items 10–12, Social context items 13–17, and Metacognition item 18). To enable easy and comprehensible visualisation of the authenticity of assessments across a whole programme of study the results were condensed into two dimensions: Product (Task, Assessment result/form and Assessment criteria) and Process (Physical context, Social context and Metacognition).

One crucial methodological decision was whether the final score always includes every element (all statements are included), or whether the absence of an element is recorded as 'not a number', thus removed from the calculation of the mean. Since our intention was to assess authenticity based on the five-dimensional framework plus metacognition, and all items contribute to the framework, we chose to use the global score (i.e. all elements contributed). The maximum score for authenticity does include every element and therefore the absence of one of those elements should affect the mean and result in a lower mean score.

Procedure

We conducted proof-of-concept data collection using a two-step process. First, members of the research team completed ratings of each assessment in their programme of study. These were then completed by an independent member of each programme team to test reliability. Data were collected via online survey using JISC (https://www.onlinesurveys.ac.uk/). In addition, unstructured interviews were conducted with colleagues from the Department who were not a part of the programme design team. Interviewees were asked what they thought each item on the questionnaire was asking, whether they

thought there were any alternative meanings, and if any items or words were unclear or confusing. This qualitative data was used to clarify and refine the items, to ensure each item was clear and appropriate to measure the framework element it was designed to test.

Design and analysis

The data were analysed using IBM SPSS Statistics version 28 (IBM Corp 2021). Each assessment was rated for each of the 18 items independently rated by two assessors. Intraclass reliability was calculated for each pair of assessors to determine how close their ratings were, for each assessment. This was calculated as the correlation coefficient. In accordance with prior literature (Koo and Li 2016), lower than 0.5 was deemed poor, between 0.5 and 0.75 was defined as moderate, between 0.75 and 0.9 as good and 0.9 and above as excellent intraclass reliability.

Ethical considerations

All study procedures received ethical approval from the Brunel Research Ethics Online system (Approval reference: 41794-LR-Jan/2023-43484-2). All participants gave consent to take part in the project.

Calculation and graphical representation of results

In order to assess how each programme scored in the different elements of authentic assessment generated by the questionnaire, boxplots were produced for the dimensions Task, Physical Context, Social Context, Result/Form and Metacognition. This was achieved by first calculating the mean of the raw scores (from 0 to 5) generated for each of the criteria within each dimension (except Metacogntion for which there was a single question), giving every assessment a single score for each of the dimensions that could be plotted. For each programme we included all assessments for all optional modules in the analysis.

To collapse the five dimensions of authentic assessment (Task; Result/Form; Social Context; Physical Context; Metacogntion) into two dimensions for plotting and visualisation at programme level, we aggregated Task and Result/Form into 'Product' and aggregated Metacogntion, Social and Physical context into 'Process'. This was achieved by calculating the mean of Product and Process for every assessment using each of the contributory scores (eight for Product and nine for Process) generated by the questionnaire (not the mean of the two means already generated for visualisation of each dimension separately in Figure 1).

Results

Numerical measures and representation of authenticity

Comparing the five factors contributing to authentic assessment separately, there are notable differences between the two programmes (see Figure 1). In Programme A authenticity increased for all areas apart from metacognition as students progressed through

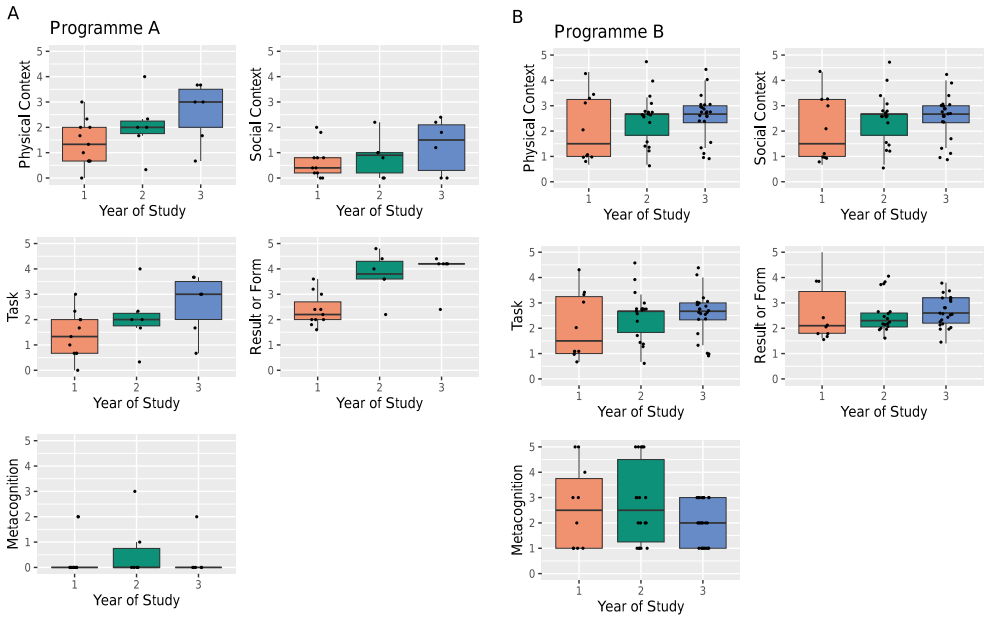

Figure 1. Boxplots showing four components of authentic assessment at each year of study for (A) Programme A and (b) Programme B. Boxes show the median and upper and lower interquartile range and whiskers show the largest and smallest value within 1.5 times the interquartile range above and below the 75th and 25th percentile. Dots represent the raw values for each assessment. Some dots have been moved a small arbitrary distance along the x axis to allow them all to be seen individually. Colour is used here only to facilitate comparison with Figures 2–4.

academic years (see Figure 1A). However, in Programme B the increase in authenticity scores was less pronounced between the academic years (see Figure 1B), although median scores did slightly rise with each year for all variables. Additionally, there was variance in the spread of authenticity scores between the programs, with the first year assessments on Programme B showing a particularly wide range of values for each variable. Looking at the plots of product versus process of the separate years of each programme (Figure 2). Several key observations emerge. The pattern of authenticity of assessment in Programme A increases year-on-year whereas in Programme B the pattern of authenticity increases from the first to the second year but then largely regresses in the third year. This will be discussed further under 'Assessments in Programme B' below. Additionally, in both programmes, assessments vary more widely along the x-axis (product) than along the y-axis (process). It's noteworthy that neither programme includes assessments significantly more authentic in process than in product. The considerable empty space in the upper part of the plot suggests that both programmes have faced challenges in developing assessments that are highly authentic in both social and physical contexts.

Assessments in Programme A

The IPA assessment strategy in program A and the iterative design process used in its development means there are comparatively few assessments on the programme. The

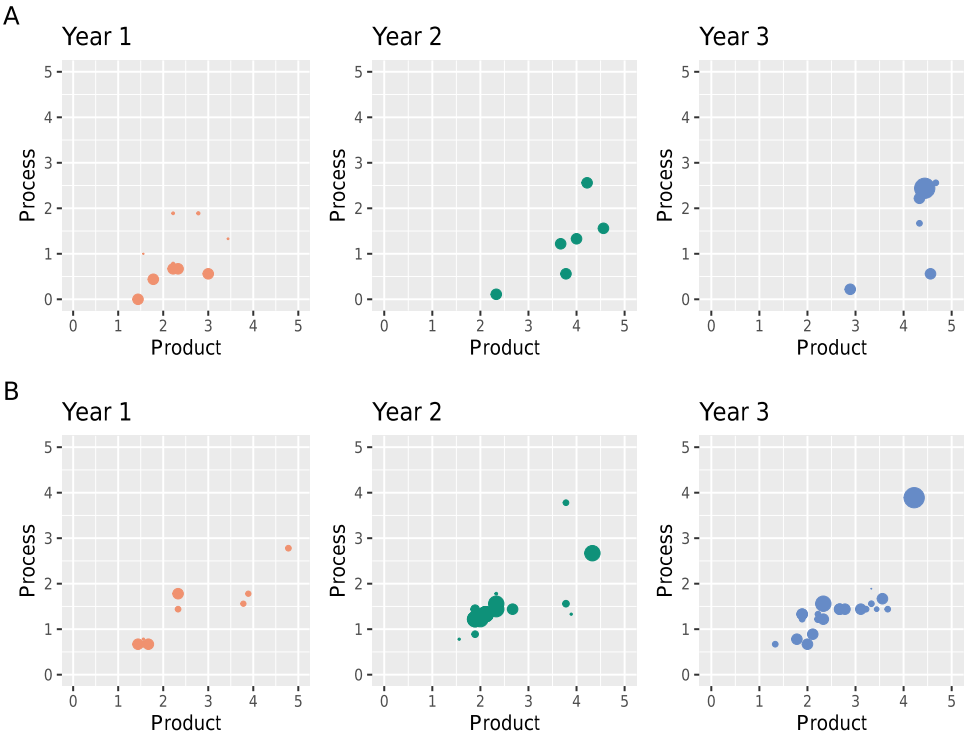

Figure 2. Each assessment for (A) Programme A and (B) plotted as a function of the authenticity of the work produced ('product') and the way in which the work is produced ('process') separately for years 1, 2 and 3 of the programmes. Values for both product and process represent the mean value of each of their constituent criteria. Colours represent the year of study. The area of each point is proportional to the amount of credits it is worth (the largest point on each plot is the 40-credit final honours project (i.e. dissertation)).

analysis presented in Figure 3A is expanded upon in Figure 3B and the underlying elements of assessment strategy are explained below. Notably, there are numerous smaller credit-bearing assessments in the first year which become fewer but higher credit-bearing as students progress through the degree. In each year, subject-specific disciplinary knowledge is assessed in traditional, unseen examinations, seen joined by arrows at the bottom right-hand corner of Figure 3A. The questions on these exams range from knowledge-specific single-best-answer MCQs in the first year to applied questions requiring a higher level of synthesis and critical evaluation in the final year. Synoptic exams are also held each year. These require students to draw together knowledge from within and beyond their programme of study in semi-unseen exams, also joined by arrows and labelled in Figure 3B. Although the latter scores more highly than the former under our scheme, these are not rated as very authentic assessments by our criteria. However, we argue that in the context of a diverse set of assessments across the year of study and programme, they are a highly valid means of determining the extent of knowledge and understanding of students. Therefore, they are an entirely appropriate form of assessment within the curriculum of the degree programme.

It is within the coursework assessments of the programme that a wider range of authentic assessments are deployed. Many of these explicitly prepare students for their final

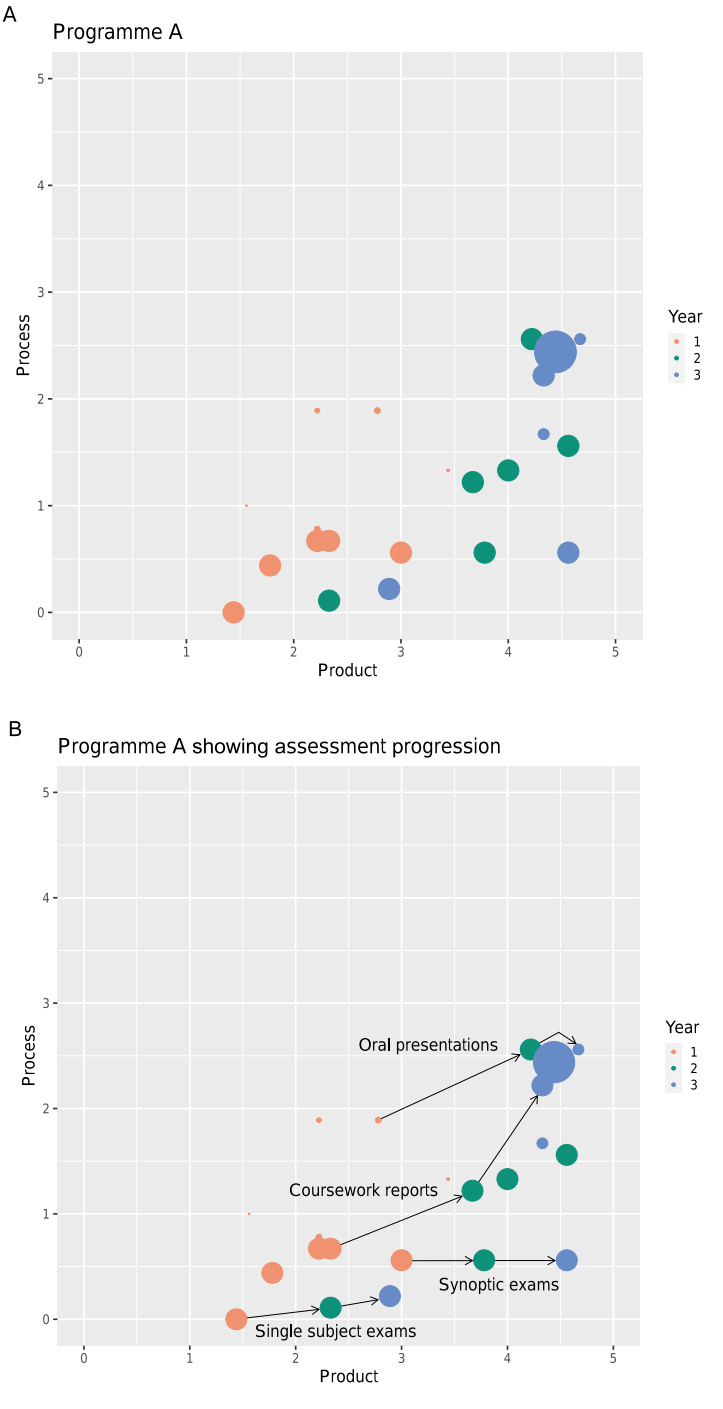

Figure 3. Each assessment for (A) Programme A plotted in a single graph (B) the same data with labels showing how different types of assessments become more authentic through the levels of the degree programme. We suggest these assessments exemplify Boud's concept of sustainability of assessments (2000). Values for both product and process represent the mean value of each of their constituent criteria. Colours represent the year of study. The area of each point is proportional to the amount of credits it is worth (the largest point on each plot is the 40-credit final honours project (i.e. dissertation)).

year research project which is assessed in the form of a dissertation. Students undertake literature review written assessments in the first two academic years. These can be in the form of a traditional essay or formatted as a review article for a journal to enhance their authenticity. These written assessments have been altered recently in response to the advent of Generative AI to be assessed less on the quality of written prose and more on the evidence of the students' engagement with the very recent research literature. In all years students generate data and analyse it and write it up for assessment. These can take the form of a traditional report or be packaged in a more authentic form, enabling students to write them up and present them as the figures for a manuscript for publication in a research journal. These can be seen labelled and connected by arrows in the middle of the graph, again moving from the middle of the bottom left towards the top right of Figure 3B. Data analysis and its presentation in written form reaches its zenith in the final year research project, represented by the largest circle at the top left of the graph. This is the culmination of the studies for students on most programmes, the capstone experience of their degree. As such, it is one of the most authentic pieces of work a student will undertake, for which they employ a wide range of high-level skills including critical thinking, problem-solving, data analysis, time and project management, collaboration, and self-directed learning. Irrespective of their career of choice these are the skills which are authentic to any graduate-level employment. Students are also asked to orally present knowledge and the results of data analysis in every year. This is either in the form of a slide or poster presentation, joined by arrows at the very top right of Figure 3B. These can also be presented in a highly authentic manner, as a student might do in a conference presentation as a PhD student or employee of a company. Again, irrespective of their career path the ability to communicate ideas clearly and persuasively in oral form is an invaluable graduate-level skill. Strong presentation skills facilitate collaboration, enable effective knowledge sharing, and enhance professional relationships across many industries and roles.

Assessments on the Programme B

One of the greatest challenges in designing the assessment strategy for Programme B is the broad range of careers the graduates pursue which necessitates equipping graduates with a broad range of transferable skills. The skills developed through the programme range from critical thinking, academic writing, group work and presenting orally, to use of digital tools and collection and analysis of data. The cohort is also large (200+), so some compulsory exam assessments, taken by all students in the early years of the degree, use automatically marked single best answer questions. Students join the programme from a wide variety of backgrounds and prior education so care has been taken at in the first year to assess using a range of different approaches, from more traditional MCQ exams and essays that students will be familiar with (seen in the bottom left quadrant of Figure 4A) to more authentic assessments such as a mock job interview and reports based on research questions and data developed by the student. In the second year, traditional essays have been made more authentic by introducing synoptic elements. In the final year, students choose four optional modules from a selection of approximately 12. In addition, they must undertake the dissertation, a substantial piece of work that contributes heavily to the final degree classification. As in all

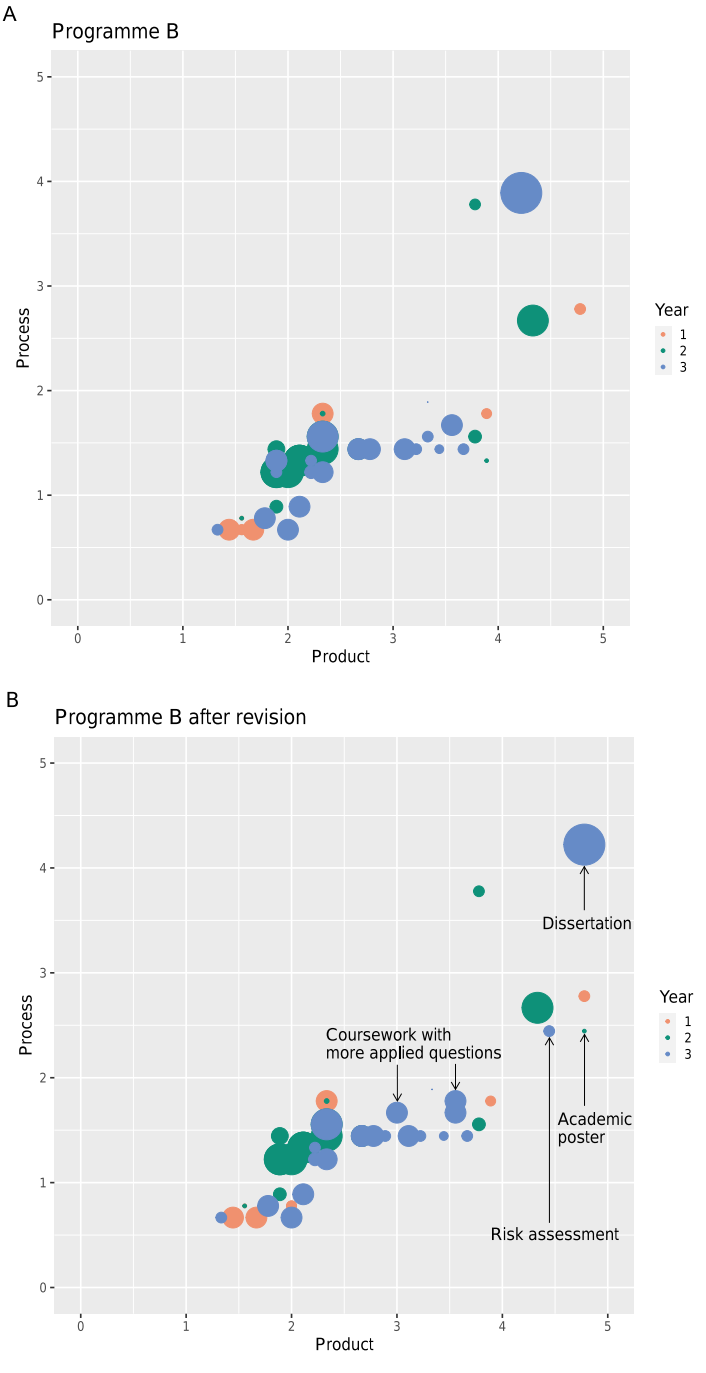

Figure 4. Each assessment for (A) Programme B plotted as a function of the authenticity of the work produced ('product') and the way in which the work is produced ('process') before changes were made to make assessments more authentic and (B) Programme B showing changed assessments with arrows indicating those made more authentic. Values for both product and process represent the mean value of each of their constituent criteria. Colours represent the year of study. The area of each point is proportional to the amount of credits it is worth (the largest point on each plot is the 40-credit final honours project (i.e. dissertation)).

programmes this requires students to incorporate many of the skills they have developed during their degree, such as engaging in diverse social interactions, performing tasks such as design, data analysis, and academic writing, culminating in a report akin to a journal article. It necessitates long-term planning and self-directed work akin to professional settings. The dissertation is therefore highly authentic in both product and process, as evident in Figure 4A, displaying many authentic graduate-level attributes. Students will need whatever their career path may be.

We used a programme-level approach to identify assessments that could be altered in product or process to increase their authenticity. Changed assessments can be see labelled with arrows in Figure 4B. For assessments identified for change, we used the granular data of the subscales to identify element(s) which would benefit from enhanced authenticity. In the first year, the MCQ statistics exams were tweaked to include research methods questions and students were provided with data to analyse rather than simply answering MCQs based on a predetermined output. The exam was converted to be open book, increasing the resources available and these tweaks shifted the assessment to be more authentic (Figure 4B). In the second year, a poster assessment was altered to include a presentation element where students give a presentation to a group of peers and the marker, followed by an interactive question and answer session in the style of a conference presentation. This tweak has increased the authenticity on the process axis (Figure 4B). In the final year the plots showed that there was very little authenticity of assessments across the optional modules which had been independently designed by the module leaders. They were predominantly traditional theory-based essays and exams. Using the data of the six subscales to determine where tweaks would be most effective in boosting authenticity modifications to these assessments have included making the questions more applied or based on clinical scenarios. For the one module, a traditional written coursework was replaced with a risk assessment report where students were provided with a case file of a clinical scenario and were required to use a validated risk assessment questionnaire standard in professional practise to produce a risk assessment and ongoing plan. The alterations in the final year assessments have led to a greater variety of assessment type and authenticity (Figure 4B).

Discussion

The literature contains many instances of conceptual frameworks and descriptions of the commonly recognised aspects of authentic assessment (Ajjawi et al. 2024, Ashford-Rowe et al. 2014, Gulikers et al. 2004). Our first aim in this work was to develop a quantifiable estimate of the level of authenticity of assessments in a reproducible way based on published conceptualisations of authenticity. We have then simplified the measures of authenticity to two axes: Product and Process. This simplification allows the graphical representation of the pattern of authenticity of assessments across a year and/or programme of study. This tool will provide a straightforward way for educators to make estimates of the authenticity of both individual assessments and the assessment patterns on their degree programmes. Our tool has been tested for ease of use across different individuals with different degrees of familiarity with the assessments being scored, and has been found to be reliable, clear and appropriate for use independent of the subject of study.

This programme-wide view of the authenticity of assessments is a novel view of authenticity of assessment in the literature. We believe this is a valuable and important lens through which to view the experience of the educational and assessment journeys students undertake in HE in ways we expand on below. In UK universities, undergraduate degrees are typically organised into programmes of study that are highly structured by discipline to provide students with a comprehensive education in that field. The programme specification outlines the aims and structure of a programme, serving as a blueprint to the skills a student is expected to develop. The programme specifications can be used in concert with the tool we have generated to map the assessment journey of students through their study from entry to exit.

In Figure 4 we have demonstrated that our tool can be used to effectively identify areas for improvement in a programme of study. We show how minor alterations to the assessments in terms of specific aspects of the product and process can be made and what the outcome of those changes can be. Using the tool in this way has allowed us to effectively enhance the authenticity and diversity of assessments across the Programme B. We found that the tool was particularly helpful in guiding conversations with colleagues. It enabled us to show them where their assessment sat in terms of authenticity, but then to show them specific areas that they could address to enhance their authenticity, a starting point from which to support colleagues with practical solutions. It is notable also that in the pattern of assessment through both degrees the authenticity of assessments in the product axis is much greater than that in the process axis. In our experience, it is easier to change authenticity of the product students submit than the assessment process itself, as this usually only requires modifying the assessment question. For example, instead of a traditional essay, a case study could be provided that students need to analyse. The process on the other hand is more difficult as then one must think about changing the physical context and social context (i.e. longer timescales, including an element of group work, or live presentations which are logistically more difficult). The tweaks where we saw the largest improvement in process usually involved inclusion of an oral presentation. We have found social context is the subscale on which assessments are most frequently low on authenticity and the one educators often find the most difficult to boost. Future work could direct efforts towards developing practicable means to implement more authenticity into assessments in the area of social context.

On the other hand, we would argue that not all assessments on a programme need to be or should be authentic, and certainly not authentic in the same dimensions. It is notable that in the assessment pattern on Programme A there are a range of assessments each year which are not particularly authentic in either their product or process. In particular, the subject-specific examinations are not very authentic in terms of either axis and the synoptic exams are not very authentic with regards to process. However, these still build on their authenticity in each successive year of study. Despite their lack of authenticity, subject-specific examinations are important in degree programmes for assessing knowledge, skills, and understanding, offering standardised evaluation, ensuring learning outcomes, and often necessary for professional certification. They underscore the essential role of knowledge in effective thinking and problem-solving, and the importance of cultivating a broad and rich knowledge base to facilitate deeper understanding (Willingham 2006). They are also a form of assessments students are familiar with when arriving at university, so provide the useful scaffolding we advocate. We believe that a degree

programme should ideally contain a balance of assessment opportunities (Wiliam 2011) to ensure assessment for learning for students to develop skills through authentic assessments while also performing assessment of learning for external bodies to ensure students have the required knowledge for accreditation of their degrees.

Unsurprisingly, in general the level of authenticity of assessments tends to increase during a degree programme (Figure 2) as students mature in their studies and assessments become a measure of higher-order skills. We argue that a mix of authenticity of assessments is important for students to have a balanced experience and develop a wide range of skills. However, the case of the assessments on the third year of Programme B shows that a programme team cannot assume that this will be the case in a programme with many options for which individual academics have responsibility for individual assessments and where the programme is not designed with a strong assessment philosophy. This would also apply to HE institutions where degree programmes follow the structure in which students choose their major and select their modules from sets of options at the faculty level, rather than the more prescribed UK design. In such cases, there is a tendency for a academics to regress to relatively inauthentic or standard assessments like traditional essays and exams. We argue that in such a situation a faculty member deciding to introduce an authentic assessment in a programme that has not prepared a student for such an unfamiliar assessment is not ideal for the psychological safety of students (Johnson et al. 2020). It is vital to take a programme-level and quantitative approach to both that ensure limitations in authenticity are identified and that programmes appropriately scaffold learning to ensure the skills to tackle different types of assessment are developed as students progress through their programme.

In the literature, authentic assessments are often viewed as discrete events in the life of a student, which are disconnected from all the other assessments they undertake through their studies. An important element of the concept of sustainable assessments (Boud 2000, Boud and Soler 2016) is that each assessment builds on learning from previous assessments. We therefore need to ensure that we develop the students' skills/knowledge to undertake ever more authentic assessments as they go through their programme, providing useable feed forward feedback to build skills and develop psychological safety (Johnson et al. 2020). Whilst providing excellent preparation for students' eventual post-graduation employment, these skills and knowledge are largely unknown to students when they arrive at university and need to be developed at an achievable velocity. We argue that for students, appropriate scaffolding is of equal importance to varying levels of authenticity to ensure wellbeing and psychological safety (Johnson et al. 2020). Appropriate scaffolding begins in more inauthentic, knowledge-based assessments on entering university, and progresses to the highly authentic assessments they undertake when near to graduation. This increasing gradient of authenticity of assessment must be shallow enough for students not to be overly challenged by any individual assessment for which they have not been prepared.

This is critically important because students should not be expected to be assessed for something for which they have not been explicitly prepared. The anticipation of assessments is a significant contributor to students' stress and poor well-being at university (Oaten and Cheng 2005; Koudela-Hamila et al. 2022). It is thus important that we mitigate the impact of assessments that students do not feel prepared for, especially during significant life changes like the transition from further to HE (Macaskill 2013). When educators introduce authentic assessments to a programme of study, we should always

consider how previous assessments have prepared students and should aim to provide support and consideration for students' well-being. Indeed, Wake et al. (2024) show that students feel less safe or secure when undertaking an authentic rather than traditional assessment and argue that to optimise the impact of authentic assessments, it is crucial for students to be engaged with the process and confident in their ability to complete them. In the programme-level approach we advocate promoting authentic assessments that build sustainably throughout a programme. Teams can strike a balance between implementing new assessment methods with more authentic aspects to them while ensuring a supportive and secure learning and assessment environment for students (Johnson et al. 2020). This incremental approach to building authenticity through a student's learning journey emphasises the importance of fostering growth in assessment practices using successively less-traditional, more authentic assessments, while also prioritising students' emotional well-being, comfort, and sense of security (Arjanggi and Kusumaningsih 2016). This balance recognises that effective teaching and assessment involves both pushing the boundaries of what students are comfortable with and taking risks with more complex and less defined assessments. This will enhance learning experiences while continuing to provide a nurturing and inclusive environment where students feel valued, respected, and empowered to take academic risks without fear of judgment or failure.

Conclusion

In this work, we have demonstrated the value that can be gained from taking a programme-wide quantitative approach to authentic assessment. Quantifying the authenticity of each assessment across a programme of study in terms of product and process allows a simple visualisation. Through this visualisation, one can easily identify which assessments would benefit from tweaks, as well as which dimension(s) to target (i.e. product or process). The granularity of the subscales can then be used to guide where to make these tweaks. As Race, Brown, and Smith (2005) note, '*The results of our assessment influence our students for the rest of their lives and careers – fine if we get it right, but unthinkable if we get it wrong*'. Done right, a thoughtfully designed range of assessments integrated into a sustainable framework with varying and progressively increasing levels of authenticity has the potential to equip students for success in an uncertain future. Done wrongly, authentic assessments inserted into a programme of study with no thought for the readiness of students to undertake them, and no support for the psychological safety of students, may do more harm than good, confusing and alienating students. Attempts to reform assessments should consider their impact on students and their wider, programme-wide experience. Interactive dialogue between students and educators is important, as exemplified in sustainable assessments, where there is an emphasis on formative feedback for students to support them on their learning journeys.

Acknowledgements

The authors would like to thank Dr Martina Bocchetta, Dr Matteo de Marco, Dr Ignazio Puzzo, Dr Sabrina Tosi and Dr Nicholas Morris for providing the second set of ratings, Prof Amanda Harvey,

Dr Nicholas Pound, Dr Matteo de Marco and Dr Ignazio Puzzo for providing qualitative feedback on the tool items and Dr Beverley Pickard-Jones, Prof Ronan McCarthy and Prof Amanda Harvey for providing helpful feedback on earlier version of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

David Tree http://orcid.org/0000-0002-8427-3334
Stephanie Baines http://orcid.org/0000-0001-7293-9517
Pauldy Otermans http://orcid.org/0000-0001-8495-348X

References

Ajjawi, R., J. Tai, M. Dollinger, P. Dawson, D. Boud, and M. Bearman. 2024. From authentic assessment to authenticity in assessment: Broadening perspectives. *Assessment & Evaluation in Higher Education* 49, no. 4: 499–510. doi:10.1080/02602938.2023.2271193.

Ajjawi, R., J. Tai, T.L. Huu Nghia, D. Boud, L. Johnson, and C.J. Patrick. 2020. Aligning assessment with the needs of work-integrated learning: The challenges of authentic assessment in a complex context. *Assessment & Evaluation in Higher Education* 45, no. 2: 304–16. doi:10.1080/02602938.2019.1639613.

Arjanggi, R., and L.P.S. Kusumaningsih. 2016. The correlation between social anxiety and academic adjustment among freshmen. *Procedia – Social and Behavioral Sciences* 219: 104–7. doi:10.1016/j.sbspro.2016.04.049.

Ashford-Rowe, K., J. Herrington, and C. Brown. 2014. Establishing the critical elements that determine authentic assessment. *Assessment & Evaluation in Higher Education* 39, no. 2: 205–22. doi:10.1080/02602938.2013.819566.

Atherton, G., J. Lewis, and B. Bolton. 2023. Higher education in the UK: Systems, policy approaches, and challenges. Research briefings, UK Parliament, https://commonslibrary.parliament.uk/research-briefings/cbp-9640/.

Boud, D. 2000. Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education* 22, no. 2: 151–67. doi:10.1080/713695728.

Boud, D., and R. Soler. 2016. Sustainable assessment revisited. *Assessment & Evaluation in Higher Education* 41, no. 3: 400–13. doi:10.1080/02602938.2015.1018133.

Brame, C. 2016. *Active learning*. Nashville, USA: Vanderbilt University Center for Teaching.

Brown, S. 2005. Assessment for learning. *Learning and Teaching in Higher Education* 1: 81–89. https://eprints.glos.ac.uk/id/eprint/3607.

Charlton, N., and R. Newsham-West. 2024. A conceptual model for program-level assessment. *Higher Education Research & Development* 43, no. 5: 1074–1088. doi:10.1080/07294360.2024.2364094.

Dann, R. 2014. Assessment as learning: Blurring the boundaries of assessment and learning for theory, policy and practice. *Assessment in Education: Principles, Policy & Practice* 21, no. 2: 149–66. doi:10.1080/0969594X.2014.898128.

Felder, R.M., and R. Brent. 2009. Active learning: An introduction. *ASQ Higher Education Brief* 2, no. 4: 1–5.

Fosnot, C.T., and R.S. Perry. 1996. Constructivism: A psychological theory of learning. *Constructivism: Theory, Perspectives, and Practice* 2, no. 1: 8–33. New York: Teacher's College Press.

Gulikers, J.T., T.J. Bastiaens, and P.A. Kirschner. 2004. A five-dimensional framework for authentic assessment. *Educational Technology Research and Development* 52, no. 3: 67–86. doi:10.1007/BF02504676.

Harvey, A., A. La Rooy, S. Kent, M. Rand-Weaver, and D. Tree. 2021. A tale of two CATES – improving assessment through collaboration. In *The CATE collection: together in collaborative educational leadership*, edited by Debbie Tolson, Nicola Watchman Smith, Dawne Irving-Bell, and Miles Weaver, 48–61. London: Advance HE.

IBM Corp. 2021. *Released 2021. IBM SPSS Statistics for Windows, Version 28.0*. Armonk, NY: IBM Corp.

Johnson, C.E., J.L. Keating, and E.K. Molloy. 2020. Psychological safety in feedback: what does it look like and how can educators work with learners to foster it? *Medical Education* 54, no. 6: 559–70. doi:10.1111/medu.14154.

Koo, T.K., and M.Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15 (2): 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Koudela-Hamila, S., J. Smyth, P. Santangelo, and U. Ebner-Priemer. 2022. Examination stress in academic students: A multimodal, real-time, real-life investigation of reported stress, social contact, blood pressure, and cortisol. *Journal of American College Health* 70 (4): 1047–1058. https://doi.org/10.1080/07448481.2020.1784906.

Macaskill, A. 2013. The mental health of university students in the United Kingdom. *British Journal of Guidance and Counselling* 41 (4): 426–441. https://doi.org/10.1080/03069885.2012.743110.

McArthur, J. 2023. Rethinking authentic assessment: work, well-being, and society. *Higher Education* 85, no. 1: 85–101. doi:10.1007/s10734-022-00822-y.

Nieminen, J.H., M. Bearman, and R. Ajjawi. 2023. Designing the digital in authentic assessment: Is it fit for purpose? *Assessment & Evaluation in Higher Education* 48, no. 4: 529–43. doi:10.1080/02602938.2022.2089627.

Oaten, M., and K. Cheng. 2005. Academic examination stress impairs self-control. *Journal of Social and Clinical Psychology* 24 (2): 254–279. https://doi.org/10.1521/jscp.24.2.254.62276.

Office for Students. 2023. About the Teaching Excellence Framework (TEF). https://www.officeforstudents.org.uk/advice-and-guidance/the-tef/about-the-tef/.

Overono, A.L., and A.S. Ditta. 2023. The rise of artificial intelligence: A clarion call for higher education to redefine learning and reimagine assessment. *College Teaching*, 1–4. doi:10.1080/87567555.2023.2233653.

Race, P., S. Brown, and B. Smith. 2005. *500 tips on assessment second edition*. London: Falmer.

Sokhanvar, Z., K. Salehi, and F. Sokhanvar. 2021. Advantages of authentic assessment for improving the learning experience and employability skills of higher education students: A systematic literature review. *Studies in Educational Evaluation* 70: 101030. doi:10.1016/j.stueduc.2021.101030.

Wake, S., M. Pownall, R. Harris, and P. Birtill. 2024. Balancing pedagogical innovation with psychological safety? Student perceptions of authentic assessment. *Assessment and Evaluation in Higher Education* 49 (4): 511–522. https://doi.org/10.1080/02602938.2023.2275519.

Wiggins, G. 1990. The case for authentic assessment. *Practical Assessment, Research, and Evaluation* 2, no. 1: 1–3.

Wiliam, D. 2011. What is assessment for learning? *Studies in Educational Evaluation* 37, no. 1: 3–14. doi:10.1016/j.stueduc.2011.03.001.

Willingham, D.T. 2006. How knowledge helps. *American Educator* 30, no. 1: 30–37.