

Assessment & Evaluation in Higher Education

ISSN: 0260-2938 (Print) 1469-297X (Online) Journal homepage: www.tandfonline.com/journals/caeh20

Integrated Programme Assessment: quantitative evidence that curriculum transformation delivers

Mathieu Di Miceli & David Tree

To cite this article: Mathieu Di Miceli & David Tree (05 Oct 2026): Integrated Programme Assessment: quantitative evidence that curriculum transformation delivers, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2026.2739378

To link to this article: https://doi.org/10.1080/02602938.2026.2739378

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

View supplementary material

Published online: 05 Oct 2026.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=caeh20

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2026.2739378

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Integrated Programme Assessment: quantitative evidence that curriculum transformation delivers

Mathieu Di Miceli[a] and David Tree[b]

[a]School of Biomedical Sciences, Faculty of Health and Life Sciences, University of Leeds, Leeds, UK; [b]College of Health, Medicine and Life Sciences, Brunel University of London, Uxbridge, UK

ABSTRACT

Integrated Programme Assessment (IPA) is an approach to curriculum design in which assessment is organised at programme rather than module level, enabling synoptic assessments requiring students to integrate knowledge and skills across a programme rather than within isolated modules. IPA is advocated on the claims that it reduces assessment volume and increases authenticity, neither of which has been subjected to empirical scrutiny using quantitative before-and-after data from real programme transformations. This study applies an established analytical tool for measuring and mapping assessment authenticity to STEM degree programmes at two universities before and after transition to IPA. The two institutions implemented IPA to substantially different degrees: one adopted IPA across the full programme while the other transitioned approximately a third of its curriculum. At one institution assessment volume fell significantly alongside significant gains in authenticity. At the other, no significant change in assessment volume was observed yet authenticity gains were similarly significant. Post-IPA, assessments at both institutions showed a more coherent year-on-year progression in authenticity, consistent with principles of sustainable assessment. These findings provide quantitative evidence of changes following IPA transition, with volume reduction observed following full implementation and authenticity gains following both full and partial redesign.

KEYWORDS

Integrated programme assessment; authentic assessment; programme-level assessment; assessment volume

Introduction

Assessment occupies a central role in higher education, shaping not only how students learn but also how learning is delivered. Within undergraduate programmes, assessment practices signal to students what the curriculum values, guide student effort and influence the development of disciplinary knowledge, attributes, skills and professional competencies. Assessment has long been recognised as a primary driver of student behaviour and learning (Boud 1990; Rust 2002), yet traditional assessment

CONTACT David Tree david.tree@brunel.ac.uk College of Health, Medicine and Life Sciences, Brunel University of London, Uxbridge UB8 3PH, UK.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/02602938.2026.2739378.

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

practices remain largely modular and fragmented. The dominance of modular degree structures, introduced in the 1990s to facilitate flexible credit accumulation, has produced assessment regimes in which students experience a large number of disconnected, relatively low-stakes tasks rather than a coherent developmental journey (Charlton, Weir, and Newsham-West 2022; McCafferty, Barney, and Walker 2025). Traditionally, these assessments have relied heavily on decontextualised formats such as timed examinations and standardised essays. Although traditional tests can support learning and long-term knowledge retention (Murphy, Little, and Bjork 2023), decontextualised assessment formats have been increasingly critiqued for their limited capacity to capture the complexity of real-world application and higher-order learning, especially under time-restricted conditions (Gernsbacher, Soicher, and Becker-Blease 2020). When assessment is compartmentalised into individual academic blocks rather than planned across a programme, students tend not to see the interconnectedness of their learning and instead view each assessment as a unit of learning currency to be completed and discarded (Charlton, Weir, and Newsham-West 2022). In this environment, students are incentivised towards surface learning, struggling to feed-forward insights from one assessment to the next and rarely developing the integrative, higher-order competencies that professional life demands (Jessop and Tomas 2017).

In response, there has been growing interest in authentic assessment (Ajjawi et al. 2024; Wiggins 1990): an approach that seeks to align assessment tasks with the kinds of challenges, practices and contexts students are likely to encounter beyond university, and which has been shown to foster transferable skills and enhance student motivation and engagement (Quinlan, Sellei, and Fiorucci 2025). However, the concept of authenticity has been interpreted in different ways across the literature (Ajjawi et al. 2020; McArthur 2023). Recent work has distinguished between authentic assessment as a design approach and authenticity in assessment as the extent to which features of an assessment reflect real-world practices and contexts (Ajjawi et al. 2024). In the present study, authenticity is treated as multi-dimensional and measurable by degree, rather than 'authentic' or 'inauthentic'. These dimensions include the nature of the task, the form of the assessment product, the physical context in which it is undertaken, and the social context in which it is embedded, and prior work has shown that these dimensions do not respond equally to curriculum intervention, with the social dimension in particular proving resistant to change (Baines et al. 2025). Critically, structural integration of assessment does not automatically produce authentic learning: authenticity requires deliberate and intentional design that embeds real-world relevance, developmental scaffolding and student agency (Tree and Worsfold 2026). The implementation of authentic assessments also presents practical challenges, including issues of scalability, reliability and staff workload (Nkealah 2023; Pallant, Pallant, and Jopp 2024), which partly explains why traditional assessment formats have persisted despite widespread recognition of their limitations.

Alongside growing interest in authentic assessment, the past decade has seen increasing attention paid to programme-level assessment (PLA): the strategic design of assessment across a whole programme of study to support the sequential development of knowledge, skills and professional capabilities (Charlton and Newsham-West 2024a;

Walker 2025). PLA encompasses a spectrum of approaches, from basic coordination of assessments across modules to full integration of assessment at programme-level (McCafferty, Barney, and Walker 2025). Integrated Programme Assessment (IPA) sits towards the integration end of this spectrum. IPA separates teaching and assessment, enabling synoptic assessments that require students to integrate and apply knowledge from across their taught programme rather than within isolated modules (Tree, Harvey, and Rand-Weaver 2025). As illustrated in Figure 1, this represents a fundamental shift in the relationship between teaching and assessment, from subject-specific tasks contained within individual modules (Figure 1A), through partial integration where some assessments draw across subjects (Figure 1B), to full integration where all assessment is synoptic and programme-wide (Figure 1C).

It is important to distinguish IPA clearly from programmatic assessment, a related but distinct approach originating in medical education (van der Vleuten et al. 2012; van der Vleuten and Schuwirth 2005). Programmatic assessment is built on a philosophy of multiple low-stakes assessment datapoints, aggregated towards high-stakes progression decisions made by committee, and has been codified through the Ottawa Consensus principles which set out an international consensus framework for the design and implementation of programmatic assessment (Heeneman et al. 2021). IPA, by contrast, uses fewer, higher-stakes integrative assessments designed to address programme-level learning outcomes and has been adopted across a broader range of disciplines beyond healthcare (Tree, Harvey, and Rand-Weaver 2025). Although both approaches seek greater coherence and longitudinal development, their assessment architectures are substantially different. They also have different histories, governance requirements and approaches to implementation (Francis et al. 2026). Conflating them makes it difficult to compare how programme-level approaches are implemented, what they require of institutions and the outcomes they produce. The distinction is particularly relevant in a field where terminology is inconsistently applied (McCafferty, Barney, and Walker 2025; Tree, Harvey, and Rand-Weaver 2025).

IPA is advocated on the basis of several claimed benefits. These include reducing assessment volume and the associated burden on students and staff; enhancing the authenticity and relevance of assessments; and better preparing graduates for

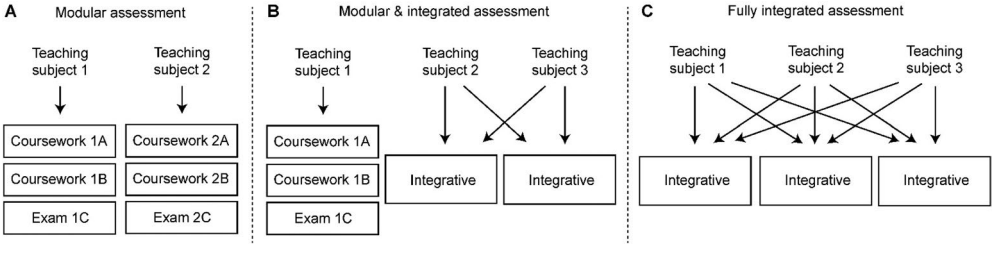


Figure 1. Schematic illustration of three approaches to assessment organisation. In modular assessment (a), each teaching subject has its own discrete assessments with no integration across subjects. In partial IPA (B), subject-specific assessments are retained alongside integrative assessments that draw across subjects. In full IPA (C), all assessment is synoptic and programme-wide, with no subject-specific assessment components.

professional life, equipping them with the discipline-specific knowledge and well-developed generic skills that employability demands (Charlton, Weir, and Newsham-West 2022; McCafferty, Barney, and Walker 2025; Tree, Harvey, and Rand-Weaver 2025). Crucially, while both authenticity and reduced assessment volume are repeatedly cited as outcomes of IPA, neither claim has been subjected to systematic empirical scrutiny using quantitative before-and-after data from real programme transitions. The evidence base remains largely descriptive and institution-specific (Tree, Harvey, and Rand-Weaver 2025), with a recent review finding limited empirical evidence of outcomes in practice (McCafferty, Barney, and Walker 2025). As institutions invest substantially in programme-focussed assessment redesign (Walker 2025) robust evidence of its outcomes is needed.

To bring consistency and measurability to the concept of authenticity of assessment - widely regarded as contextually variable and resistant to precise definition - an analytical tool was developed by Baines et al. (2025) to measure, map and visualise authenticity of individual assessments and of assessment patterns across whole programmes of study. This tool is grounded in published frameworks of authentic assessment (Ashford-Rowe, Herrington, and Brown 2014; Gulikers, Bastiaens, and Kirschner 2004) and breaks down authenticity into constituent dimensions which are then condensed into two overarching axes, product and process, enabling quantitative comparison and graphical mapping of authenticity across years of study and between programmes. Each assessment is rated on 18 items using a five-point scale, with items organised into subscales: task, result/form, physical context and social context, plus a single metacognition item. The product axis aggregates the subscales relating to what students produce (task and result/form) while the process axis aggregates those relating to the conditions under which they produce it (physical context, social context and metacognition), yielding two numerical scores per assessment that can be plotted, compared across programmes and tracked over time.

In a proof-of-concept study at a single institution, the tool was applied cross-sectionally to two programmes, one designed using IPA and one following a conventional modular structure, demonstrating that the IPA programme showed a consistent year-on-year increase in authenticity while the modular programme showed some regression in authenticity in the final year (Baines et al. 2025). That study demonstrated the application of the measurement approach but did not examine change within programmes following a transition to IPA. The present study addresses this question by applying the same analytical framework to curricula before and after IPA implementation at two universities, referred to here as University A and University B. It also examines assessment volume along with authenticity, allowing two of the principal claims made for IPA to be evaluated within the same programme transformations. The datasets analysed here are distinct from those reported by Baines et al. (2025).

The before-and-after design, replicated across two institutions with different degrees of IPA implementation, allows changes in assessment authenticity and volume associated with the transition to IPA to be examined within programmes over time. Both programmes share the same disciplinary context, reducing subject-specific variation in assessment norms. The two IPA implementations differ substantially in scope: at University A, IPA was implemented across the full programme; at University B, approximately one third of the curriculum transitioned to IPA, broadly

corresponding to the intermediate model illustrated in Figure 1B. Assessment volume before and after IPA was also examined. Two research questions are addressed: first, does assessment burden change following transition to IPA, and does the pattern of change differ between programmes with different degrees of implementation; and second, does assessment authenticity change following transition to IPA and does the pattern of change similarly differ according to the extent of implementation.

Materials and methods

This study involved documentary analysis of programme and assessment materials, did not involve human participants and as such does not fall within the remit of the Research Ethics Committee of University A, so no formal ethical approval was required. At University B, ethical approval was granted by the institutional ethics committee (ELPERF #1920). Both universities are research intensive institutions which have a strong widening participation remit and offer a broad range of undergraduate degree programmes. The study examines STEM degree programmes at each institution before and after their respective transitions to IPA. At University A, IPA was introduced for incoming students from 2011 to 12, replacing a conventional modular programme that had been in place previously. By 2014-15, all graduating students had experienced the full IPA programme throughout their studies. Pre-IPA data used in this study are drawn from the 2010-11 academic year, and post-IPA data from 2024 to 25, by which point IPA had been fully embedded for over a decade. At University B, IPA was introduced on a rolling basis from 2019 to 20, reaching full implementation across all years of study by 2022-23; pre-IPA data are from 2020 to 2021 and post-IPA data from 2024 to 25, following a further period of programme refinement. As described in the introduction, the two implementations differed substantially in scope: at University A, IPA was adopted across the full programme, while at University B approximately one third of the curriculum transitioned to IPA. It should be noted that the post-IPA data for University A used in the present study are drawn from the 2024-25 academic year, distinct from the 2021-22 data used in the cross-sectional study by Baines et al. (2025); the two datasets are independent and the present study does not reanalyse any previously published data.

All compulsory assessments for each programme were identified from programme documentation for both the pre- and post-IPA time points. For University A, where programmes included optional modules, the most selected combination of modules totalling 120 credits per academic year was used, reflecting the typical assessment experience of students on each programme. For University B, curriculum mapping was limited to compulsory modules because optional module choices vary between students, providing a consistent basis for analysis. In addition, some assessments at University B do not bear any credits but are compulsory for progression into the next year, for example the Technical Skills Passport, a structured framework whereby students progressively demonstrate practical laboratory competencies across Years 1 and 2, with selected competencies designated as pass-for-progression requirements. In both institutions, for each assessment, the year of study and credit weighting were recorded alongside the authenticity ratings described below.

Authenticity of each assessment was measured using the analytical tool developed by Baines et al. (2025), comprising 18 items that address the key dimensions of authentic assessment drawn from the frameworks of Gulikers, Bastiaens, and Kirschner (2004) and Ashford-Rowe, Herrington, and Brown (2014). Items are organised into four subscales, task, result/form, physical context and social context, with an additional item measuring metacognition. These subscales are aggregated into two overarching dimensions: product (comprising task, result/form and assessment criteria) and process (comprising physical context, social context and metacognition). Each item was rated on a five-point scale from 1 (not at all) to 5 (extremely), with N/A assigned where an item was not applicable to the assessment in question. Mean scores were calculated for each subscale and for the two overarching dimensions.

Ratings were made by members of the authorship team working directly from programme documentation, including module handbooks, assessment briefs and marking criteria. Before rating, the authorship team discussed and agreed the interpretation and application of the scoring framework and co-developed a structured Authentic Assessment Scoring Guide to support consistent application across the two institutions. Individual assessments were subsequently rated by one member of the authorship team using the agreed scoring guide. As an additional calibration check, independently scored assessments with comparable purposes and formats at the two institutions were compared to examine consistency in the application of the scoring framework. The Authentic Assessment Scoring Guide is available as Supplementary File S1. The internal consistency of scores across items in the authenticity measurement tool was assessed using Cronbach's α (Cronbach 1951), calculated separately for each university and across the combined dataset.

Assessment volume was analysed by recording the number and credit weighting of all assessments at each time point for both institutions. Statistical analyses were conducted in R version 4.0.3. Differences in assessment volume before and after IPA transition were analysed using two-way ANOVAs, with year of study and IPA transition as factors. Differences in authenticity scores were analysed using two-way ANOVAs with year of study and IPA transition as factors, with assessment credits included as weights to reflect the differential contribution of assessments to the overall programme. Statistical significance was set at $p < 0.05$. The two factors were year of study (Years 1, 2 and 3) and IPA transition (pre *versus* post). Where significant interactions were identified, Tukey's post-hoc tests were computed to evaluate pairwise comparisons. Figures were produced using the ggplot2 and patchwork packages in R (https://cran.r-project.org/web/packages/ggplot2/index.html; patchwork: https://cran.r-project.org/web/packages/patchwork/index.html; version numbers: ggplot2 = 3.5.2 patchwork = 1.3.0.9000).

Results

Internal consistency

We assessed the internal consistency of scores across items in the authenticity measurement tool developed by Baines et al. (2025), both within individual institutions

and overall. At University A, Cronbach's α was 0.94. Similarly, Cronbach's α was 0.96 for University B. Across all assessments and institutions, the overall Cronbach's α was 0.95. As an additional calibration check, independently scored assessments with comparable purposes and formats at the two universities were compared descriptively. Product and process authenticity scores were generally similar across these paired assessment types (Supplementary Table S2).

Assessment volume and credit distribution

One of the most widely cited benefits of IPA is its capacity to reduce assessment volume and consolidate credit weighting into fewer, more substantial tasks. Figure 2 shows the number and credit weighting of assessments across each year of study at University A and University B before and after the transition to IPA. At University A there were significant effects of year of study ($F(2,80) = 8.42$, $p = 0.0005$) and IPA transition ($F(1,80) = 24.09$, $p < 0.0001$) on assessment volume, together with a significant year-by-IPA interaction ($F(2,80) = 5.72$, $p = 0.005$), indicating that the change in assessment volume differed across years of study. This is seen in the reduction in the number of assessments, most pronounced in Years 1 and 2 (Figure 2A, Supplementary Table S1A), accompanied by a clear shift towards fewer, more substantially credit-weighted assessments post-IPA in Figure 2A. At University B there were no significant effects of year ($F(2,62) = 1.64$, $p = 0.20$), IPA transition ($F(1,62) = 0.51$, $p = 0.48$), or their interaction ($F(2,62) = 0.08$, $p = 0.93$) on assessment volume, consistent with the more modest changes seen in Figure 2B. The contrasting findings show that while a full programme IPA implementation at University A was associated with a statistically significant consolidation of assessment load, a partial implementation at University B did not significantly alter the overall number or distribution of assessments. As detailed below this difference in assessment volume

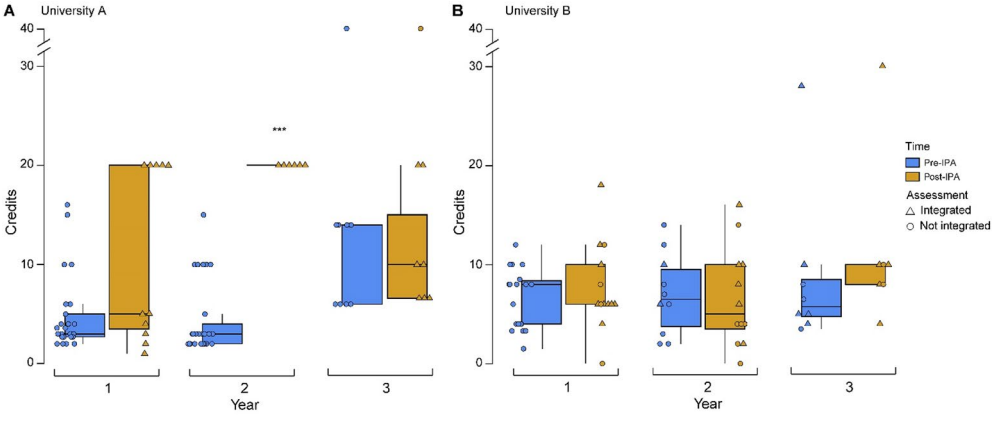


Figure 2. Assessment volume and credit distribution at University A (A) and University B (B) before (pre-IPA) and after (post-IPA) the transition to integrated Programme assessment, shown separately for each year of study (years 1, 2 and 3). Each point represents an individual assessment, with its position on the y-axis indicating its credit weighting. Integrated assessments are presented as triangles, non-integrated assessments as circles.

change does not preclude meaningful change in the authenticity of assessments at University B.

Authenticity of assessment: constituent dimensions/elements

A central aim of this study was to determine whether measurable changes in the authenticity of assessments across the five constituent dimensions of the tool were observed following transition to IPA. Figures 3 and 4 show authenticity scores across the five constituent dimensions of physical context, social context, task, result and metacognition for each year of study at both universities before and after the transition to IPA. These dimensions capture how the assessment task resembles real-world professional challenges (task); the format of the assessed output (result/form); where and how the assessment takes place (physical context); its use of collaboration or professional social contexts (social context); and whether students reflect on their own learning and practice (metacognition).

At University A (Figure 3), IPA transition was significantly associated with physical context ($F(1,80) = 7.17$, $p = 0.009$), task ($F(1,80) = 34.05$, $p < 0.0001$), result ($F(1,80) = 34.05$, $p < 0.0001$) and metacognition ($F(1,80) = 31.60$, $p < 0.0001$). Significant year-by-IPA interactions were also observed for task ($F(2,80) = 3.15$, $p = 0.048$) and result ($F(2,80) = 3.25$, $p = 0.04$), indicating that changes in these dimensions differed across years of study. Social context was the only dimension showing no significant association with IPA transition ($F(1,80) = 0.01$, $p = 0.96$), consistent with evidence that the collaborative and socially contextualised conditions of assessment are among the most resistant to change, reflecting both the logistical demands of designing genuinely collaborative assessments and well-documented student resistance to group-based formats (Baines et al. 2025; Wake et al. 2024).

At University B (Figure 4), despite the more limited scope of IPA implementation, the pattern of change closely mirrors that observed at University A, with IPA transition significantly associated with physical context ($F(1,62) = 27.13$, $p < 0.0001$), task ($F(1,62) = 34.29$, $p < 0.0001$), result ($F(1,62) = 36.42$, $p < 0.0001$) and metacognition ($F(1,62) = 28.16$, $p < 0.0001$), but not social context ($F(1,62) = 2.14$, $p = 0.15$). A highly significant year-by-IPA interaction was observed for the task dimension at University B ($F(2,62) = 9.21$, $p = 0.0003$), indicating that the change in task authenticity differed across years of study. The consistency of findings across two institutions with substantially different degrees of IPA implementation suggests that gains in assessment authenticity across most of the dimensions measured can also be observed following partial implementation of IPA. At University B, comparison of integrated and non-integrated assessments after the IPA transition revealed no significant differences in authenticity across any of the five dimensions (all $p \geq 0.49$; Supplementary Figure S1; full results in Supplementary Table S1C), suggesting that higher authenticity scores post-IPA were observed beyond the formally integrated component of the curriculum.

Authenticity of assessment: product and process dimensions

To examine how changes in authenticity are distributed across the two overarching dimensions of the tool, Figures 5 and 6 show the authenticity of individual

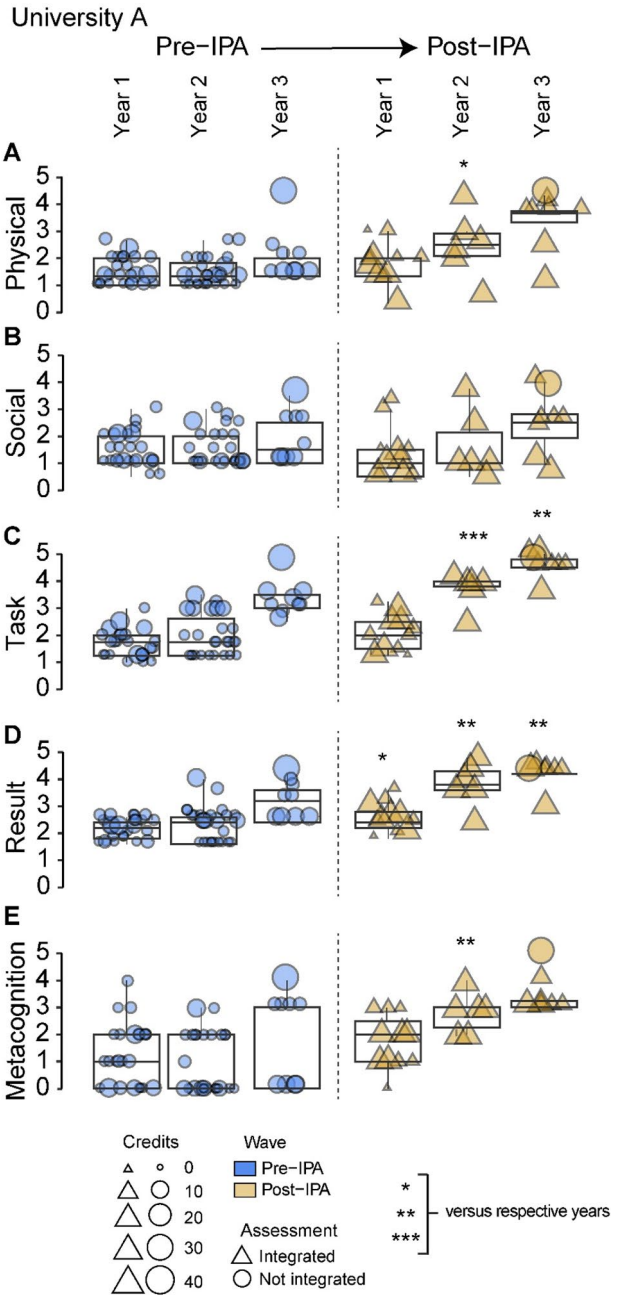


Figure 3. Authenticity scores across the five constituent dimensions of the authenticity measurement tool at University A: physical context (A), social context (B), task (C), result (D) and metacognition (E), shown for each year of study (years 1, 2 and 3) before (pre-IPA) and after (post-IPA) the transition to IPA. Each point represents an individual assessment, with its area proportional to credit weighting. Boxes show the median and interquartile range; whiskers show the largest and smallest values within 1.5 times the interquartile range above and below the 75th and 25th percentiles. Integrated assessments are presented as triangles, non-integrated assessments as circles. Higher scores indicate greater authenticity within the corresponding dimension. Statistical comparisons between pre-IPA and post-IPA within each year of study are indicated with asterisks ($p < 0.05$, $p < 0.01$ and $p < 0.001$).

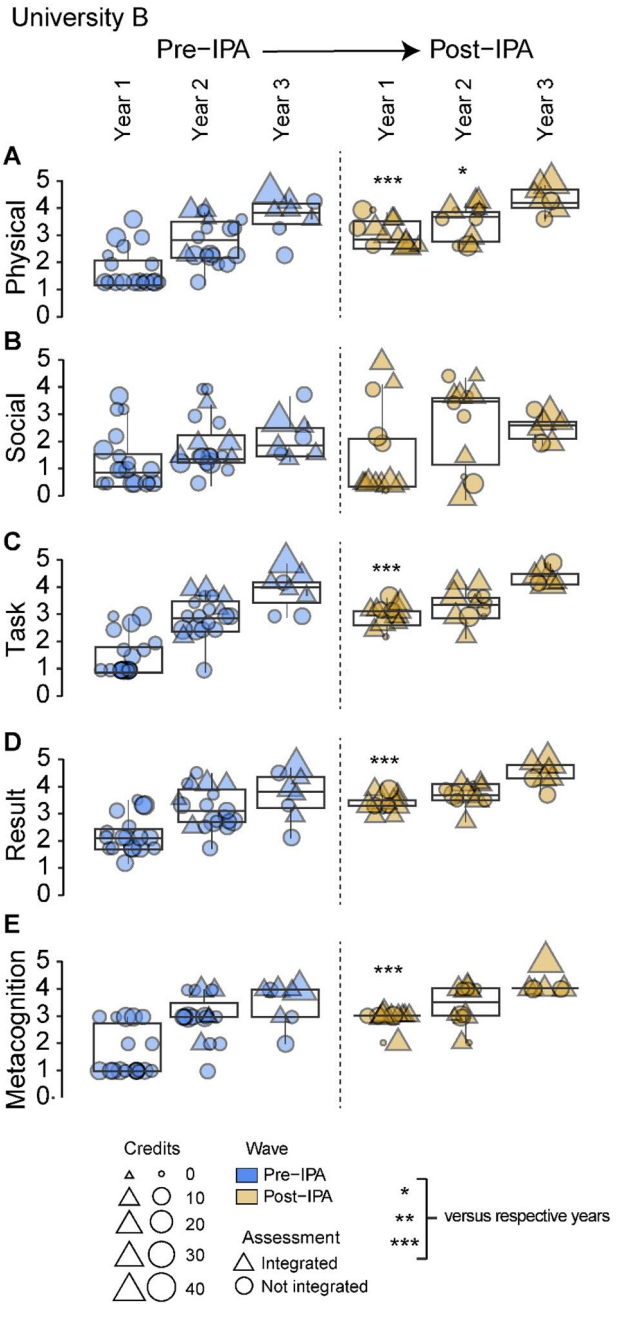


Figure 4. Authenticity scores across the five constituent dimensions of the authenticity measurement tool at University B: physical context (A), social context (B), task (C), result (D) and metacognition (E), shown for each year of study (years 1, 2 and 3) before (pre-IPA) and after (post-IPA) the transition to IPA. Each point represents an individual assessment, with its area proportional to credit weighting. Boxes show the median and interquartile range; whiskers show the largest and smallest values within 1.5 times the interquartile range above and below the 75th and 25th percentiles. Integrated assessments are presented as triangles, with non-integrated assessments as circles. Higher scores indicate greater authenticity within the corresponding dimension. Statistical comparisons between pre-IPA and post-IPA within each year of study are indicated with asterisks ($p < 0.05$, $p < 0.01$ and $p < 0.001$).

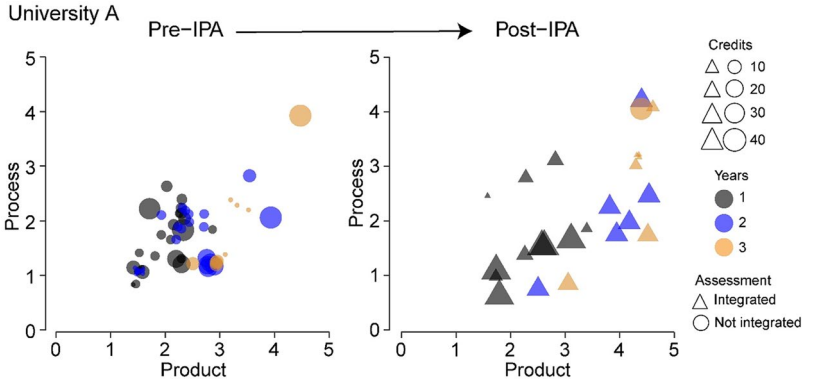


Figure 5. Overall authenticity trajectories at University A before (pre-IPA) and after (post-IPA) the transition to IPA, showing all assessments across years 1, 2 and 3 plotted as a function of product authenticity (x-axis) and process authenticity (y-axis). Values for product and process represent the mean of their constituent dimension scores. Each point represents an individual assessment, with its area proportional to credit weighting. Colours represent year of study (year 1, black; year 2, blue; year 3, orange). Integrated assessments are presented as triangles, non-integrated assessments as circles.

assessments at both universities plotted across the product axis (comprising task and result/form) and the process axis (comprising physical context, social context and metacognition) for each year of study, before and after the transition to IPA.

At University A (Figure 5), pre-IPA all three years are characterised by a large number of small, low credit-weighted assessments clustered in the lower-left region of the plot, indicating consistently low authenticity on both product and process dimensions. Following the transition to IPA, the number of assessments is considerably reduced with those remaining shifting towards higher authenticity scores, particularly on the product axis. The post-IPA plots reveal a smoother and more coherent trajectory in authenticity from Year 1 through to Year 3. In all years and in both pre- and post-IPA conditions at University A, the distribution of assessments show a consistent tendency for product authenticity to exceed process authenticity.

At University B (Figure 6), the pre-IPA assessment landscape showed a more distributed spread of authenticity scores, with several assessments scoring moderately highly on both product and process dimensions prior to IPA, alongside existing assessment practices like the use of at-home rather than on-campus examinations and collaborative assessment formats. Post-IPA, an upward and rightward shift is evident across all years at University B, with some improvement in year-on-year separation, though this is less pronounced than at University A. Notably, Figure 6 shows several assessments at University B for which process authenticity exceeds product authenticity in both pre- and post-IPA conditions, a pattern largely absent at University A.

Overall authenticity trajectories

At University A (Figure 5), pre-IPA assessments from Years 1, 2 and 3 are intermingled across the plot with no clear visual separation between years with the

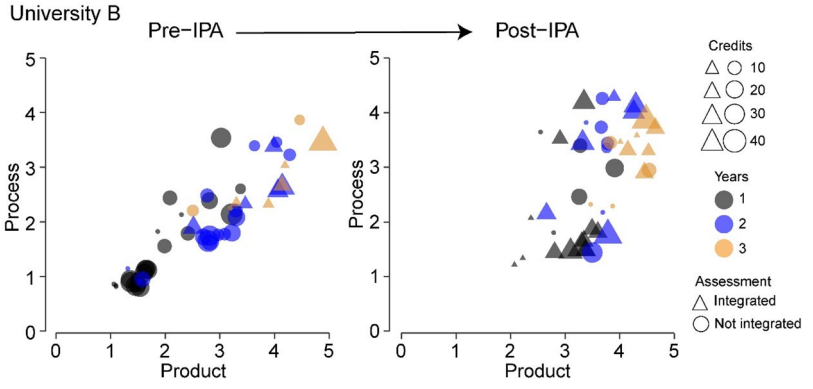


Figure 6. Overall authenticity trajectories at University B before (pre-IPA) and after (post-IPA) the transition to IPA, showing all assessments across years 1, 2 and 3 plotted as a function of product authenticity (x-axis) and process authenticity (y-axis). Values for product and process represent the mean of their constituent dimension scores. Each point represents an individual assessment, with its area proportional to credit weighting. Colours represent year of study (year 1, black; year 2, blue; year 3, orange). Integrated assessments are presented as triangles, non-integrated assessments as circles.

programme lacking a discernible assessment architecture when viewed as a whole. Post-IPA, the three years separate clearly, with Year 1 assessments clustering towards the lower-left, Year 2 in the middle ground, and Year 3 towards the upper-right, accompanied by the consolidation of credits into fewer, more substantially weighted assessments.

At University B (Figure 6), the pre-IPA assessments are already more widely distributed across the plot than the equivalent pre-IPA picture at University A, with a more visible, if incomplete, separation between years of study. Post-IPA, an upward and rightward shift is evident across all years with some improvement in year-on-year colour separation, though this is less pronounced than at University A.

Across both institutions, the post-IPA plots show a more coherent year-on-year progression in authenticity than the corresponding pre-IPA plots. This pattern is consistent with the principles of sustainable assessment in which assessment supports cumulative development across a programme of study (Boud 2000).

Discussion

The findings of this study provide the first empirical before-and-after evidence that measurable changes in both assessment volume and authenticity were observed following transition to IPA, while also revealing differences associated with the extent and context of implementation. Taken together, the results from two institutions with substantially different degrees of IPA implementation advance the empirical evidence base for IPA beyond predominantly descriptive and cross-sectional accounts and raise important questions for institutions considering the adoption of programme-level assessment approaches.

The most striking difference between the two institutions concerns assessment volume. Full implementation at University A was associated with a statistically

significant reduction in assessment number and consolidation of credit weighting, providing quantitative before-and-after evidence consistent with the claim that IPA can reduce assessment burden (McCafferty, Barney, and Walker 2025; Tree, Harvey, and Rand-Weaver 2025). Reducing assessment volume has direct implications for student wellbeing and staff workload, both of which are cited as primary motivations for IPA adoption (Tree, Harvey, and Rand-Weaver 2025). It should also be acknowledged that the transition to IPA, whether full or partial, requires a substantial investment of time and effort from academic and professional staff. Designing integrative assessments that span modules or programme levels is considerably more demanding than designing assessments within individual modules, and institutions should plan accordingly, ensuring that appropriate resource and protected time are available for the redesign process (Tree, Harvey, and Rand-Weaver 2025). The present study does not quantify the amount of staff time or resource required, but the findings reinforce the practical importance of supporting programme-level coordination rather than treating assessment redesign as a series of isolated module-level changes.

The absence of a significant volume change at University B indicates that partial implementation does not necessarily result in a measurable reduction in overall assessment volume. In contrast, the substantial reduction following full implementation in University A suggests that broader structural change across a programme may be needed if reducing assessment volume is a primary objective. With only two implementation profiles these data cannot establish a minimum proportion of the curriculum which must be integrated to achieve assessment reduction. This is an important practical consideration for institutions currently exploring programme-level assessment within the constraints of modular regulatory frameworks and existing quality assurance structures (Charlton and Newsham-West 2024b; McCafferty, Barney, and Walker 2025).

The authenticity findings show that both institutions demonstrated statistically significant gains across the majority of dimensions measured, despite their very different implementation profiles. The magnitude of change was greater at University A, although the present study does not allow this to be attributed solely to the greater extent of IPA implementation. The consistency of the finding across both institutions suggests it is not an artefact of a single institutional context. Since both programmes are in the same STEM discipline, the extent to which these findings generalise across other subject areas and institutional types remains an open question that future work should address. The principle of programme-level coordination is not discipline-specific, but how authenticity and assessment integration are realised is likely to vary between disciplines. Replication across other disciplinary and institutional contexts is therefore needed. For institutions navigating the significant operational and cultural challenges of IPA adoption (Charlton and Newsham-West 2024b; Tree, Harvey, and Rand-Weaver 2025), the key message is that meaningful gains in assessment authenticity were observed even where IPA was implemented only partially, suggesting that institutions unable to undertake full programme transformation may still benefit from targeted programme-level redesign.

The consistency of the findings across two institutions supports the evidence that the observed changes are associated with the transition to IPA. However, the

before-and-after design of this study cannot exclude alternative explanations. Changes in academic staffing, shifts in institutional culture, or broader sector-wide trends in assessment practice occurring over the same period could have contributed to the observed gains in authenticity. This is a particular consideration given that the two implementations are separated by almost a decade with University A transitioning in 2011-12 and University B in 2019-20, during which awareness and practice of authentic assessment across UK higher education developed considerably. This broader temporal context may have also contributed to the already more distributed pre-IPA authenticity profile at University B compared with University A. Nevertheless, significant authenticity gains were observed at both institutions despite their different implementation timelines and substantially different degrees of IPA adoption. The differing pattern for assessment volume, which changed significantly at University A but not University B, also indicated that the observed changes were not uniform across institutions. Taken together, these findings are consistent with an influence of programme redesign, but do not allow that influence to be isolated from other temporal or institutional changes.

The consistent absence of significant change in social context authenticity at both institutions mirrors the findings of Baines et al. (2025) and suggests that this dimension may be less readily changed through programme-level restructuring, reflecting the difficulty of embedding genuinely collaborative and socially authentic assessment formats into degree programmes. This contrasts with the substantial changes observed in task and result/form authenticity, where changes can often be made within the design of an individual assessment. Increasing social authenticity may require wider coordination, for example to enable collaborative or interactive assessment involving multiple students, with implications for staff, timetabling and other programme-level arrangements. At University B, process authenticity exceeds product authenticity for a subset of assessments. These assessments include at-home examinations and collaborative assessment formats, illustrating how specific design choices can support high process authenticity independently of full programme integration.

More broadly, the higher pre-IPA baseline at University B compared to University A may reflect differences in programme history: the University B programme was designed more recently, and growing awareness of authentic assessment principles in curriculum design practice (Baines et al. 2025; Walker 2025) may have influenced its pre-IPA assessment landscape independently of any IPA transition. The absolute level of authenticity at any given time point is therefore likely to reflect both the assessment design philosophy in place and the broader disciplinary and temporal context in which it was developed.

The overall trajectory plots in Figures 5 and 6 show one of the clearest patterns in the dataset. Post-IPA, assessments at both institutions separate more clearly by year of study, revealing a more coherent assessment architecture in which authenticity increases progressively across the years of study. This year-on-year progression is much less apparent in the pre-IPA data. At University A, the smaller change in Year 1 may reflect deliberate scaffolding for students transitioning into higher education, with greater authenticity introduced as students progress through the programme. This structured trajectory is consistent with the principles of sustainable assessment,

in which each assessment is designed to build cumulatively on students' developing skills and knowledge (Boud 2000).

The contrast between University A and University B does not suggest a simple relationship between the extent of IPA implementation and the amount of change. At University A, full implementation was followed by reduced assessment volume, increased authenticity and a more coherent overall assessment trajectory. At University B, partial implementation was followed by increased authenticity and a more coherent trajectory but without significantly reducing assessment volume. The findings suggest that the scope of structural integration may matter differently for different outcomes: substantial consolidation may be more important for reducing assessment volume, whereas gains in authenticity can accompany more targeted programme-level redesign. Integration alone does not ensure authentic assessment; this emerges through deliberate design embedding professional relevance, appropriate scaffolding and programme-level coherence (McCafferty, Barney, and Walker 2025; Tree, Harvey, and Rand-Weaver 2025; Tree and Worsfold 2026).

Several limitations should be considered when interpreting these findings. The study examines two programmes within the same STEM discipline and uses a before-and-after design in which temporal, staffing and institutional changes cannot be separated fully from the transition to IPA. Individual assessments were rated by a single rater following a standardisation exercise and use of a common scoring guide. Although the measurement framework has previously undergone inter-rater reliability testing (Baines et al. 2025), inter-rater agreement was not independently assessed in the present dataset. Comparison of two different implementation profiles cannot establish a minimum level of integration required to achieve particular outcomes. Future research should therefore test these findings across different disciplines and institutional contexts, and across a wider range of implementation models.

This study comes at a timely moment: the move towards programme-level assessment has been identified as one of the defining trends in higher education assessment practice (Walker 2025), and the evidence presented here provides further empirical evidence for institutions navigating this transition. Finally, this study demonstrates the value of the authenticity measurement tool (Baines et al. 2025) as an instrument for evaluating change following curriculum interventions at programme level: sensitive to change, applicable across different implementation contexts, and capable of revealing nuanced patterns that aggregate measures alone would obscure. We hope its application here encourages other institutions to adopt it as both a diagnostic and evaluative tool to guide and evidence their own assessment redesign.

Disclosure statement

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

ORCID

Mathieu Di Miceli http://orcid.org/0000-0003-3713-0370
David Tree http://orcid.org/0000-0002-8427-3334

References

Ajjawi, R., J. Tai, T. L. H. 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 (2): 304–316. https://doi.org/10.1080/02602938.2019.1639613.

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 (4): 499–510. https://doi.org/10.1080/02602938.2023.2271193.

Ashford-Rowe, K., J. Herrington, and C. Brown. 2014. "Establishing the Critical Elements That Determine Authentic Assessment." *Assessment & Evaluation in Higher Education* 39 (2): 205–222. https://doi.org/10.1080/02602938.2013.819566.

Baines, S., P. Otermans, D. Tree, and N. Worsfold. 2025. "Measuring and Mapping Authentic Assessment with a Novel Quantitative Typology." *Teaching in Higher Education* 30 (3): 663–682. https://doi.org/10.1080/13562517.2024.2424823.

Boud, D. 1990. "Assessment and the Promotion of Academic Values." *Studies in Higher Education* 15 (1): 101–111. https://doi.org/10.1080/03075079012331377621.

Boud, D. 2000. "Sustainable Assessment: Rethinking Assessment for the Learning Society." *Studies in Continuing Education* 22 (2): 151–167. https://doi.org/10.1080/713695728.

Charlton, N., and R. Newsham-West. 2024a. "A Conceptual Model for Program-Level Assessment." *Higher Education Research & Development* 43 (8): 1721–1736. https://doi.org/10.1080/07294360.2024.2364094.

Charlton, N., and R. Newsham-West. 2024b. "Enablers and Barriers to Program-Level Assessment Planning." *Higher Education Research & Development* 43 (5): 1074–1088. https://doi.org/10.1080/07294360.2024.2307933.

Charlton, N., K. Weir, and R. Newsham-West. 2022. "Assessment Planning at the Program-Level: A Higher Education Policy Review in Australia." *Assessment & Evaluation in Higher Education* 47 (8): 1475–1488. https://doi.org/10.1080/02602938.2022.2061911.

Cronbach, L. J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16 (3): 297–334. https://doi.org/10.1007/BF02310555.

Francis, N., P. Birtill, M. Davies, L. Fall, P. Hartley, R. James, S. Rutherford, and D. Tree. 2026. "Managing the Shift to Programme-Level Assessment: Structural Barriers, Stakeholder Experiences and Mechanisms of Sustained Transition." *Assessment & Evaluation in Higher Education:* 1–19. https://doi.org/10.1080/02602938.2026.2724543.

Gernsbacher, M. A., R. N. Soicher, and K. A. Becker-Blease. 2020. "Four Empirically Based Reasons Not to Administer Time-Limited Tests." *Translational Issues in Psychological Science* 6 (2): 175–190. https://doi.org/10.1037/tps0000232.

Gulikers, J. T. M., T. J. Bastiaens, and P. A. Kirschner. 2004. "A Five-Dimensional Framework for Authentic Assessment." *Educational Technology Research and Development* 52 (3): 67–86. https://doi.org/10.1007/BF02504676.

Heeneman, S., L. H. de Jong, L. J. Dawson, T. J. Wilkinson, A. Ryan, G. R. Tait, N. Rice, D. Torre, A. Freeman, and C. P. M. van der Vleuten. 2021. "Ottawa 2020 Consensus Statement for Programmatic Assessment – 1. Agreement on the Principles." *Medical Teacher* 43 (10): 1139–1148. https://doi.org/10.1080/0142159X.2021.1957088.

Jessop, T., and C. Tomas. 2017. "The Implications of Programme Assessment Patterns for Student Learning." *Assessment & Evaluation in Higher Education* 42 (6): 990–999. https://doi.org/10.1080/02602938.2016.1217501.

McArthur, J. 2023. "Rethinking Authentic Assessment: Work, Well-Being, and Society." *Higher Education* 85 (1): 85–101. https://doi.org/10.1007/s10734-022-00822-y.

McCafferty, H., A. Barney, and S. Walker. 2025. *Programme-Level Assessment: A Review of the Current Landscape*. QAA Collaborative Enhancement Project.

Murphy, D. H., J. L. Little, and E. L. Bjork. 2023. "The Value of Using Tests in Education as Tools for Learning—Not Just for Assessment." *Educational Psychology Review* 35 (3): 1-29. https://doi.org/10.1007/s10648-023-09808-3.

Nkealah, N. 2023. "Unravelling an Assessment Puzzle: Reflections on Reconciling Effective Assessment and Workload Management in South African Higher Education." *Assessment & Evaluation in Higher Education* 48 (8): 1092–1102. https://doi.org/10.1080/02602938.2023.2173139.

Pallant, J. I., J. L. Pallant, and R. Jopp. 2024. "The Case for Scaling Authentic Learning across Undergraduate and Postgraduate Research Skills Courses." *Teaching in Higher Education* 29 (6): 1442–1459. https://doi.org/10.1080/13562517.2022.2066468.

Quinlan, K. M., G. Sellei, and W. Fiorucci. 2025. "Educationally Authentic Assessment: Reframing Authentic Assessment in Relation to Students' Meaningful Engagement." *Teaching in Higher Education* 30 (3): 717–734. https://doi.org/10.1080/13562517.2024.2394042.

Rust, C. 2002. "The Impact of Assessment on Student Learning: How Can the Research Literature Practically Help to Inform the Development of Departmental Assessment Strategies and Learner-Centred Assessment Practices?" *Active Learning in Higher Education* 3 (2): 145–158. https://doi.org/10.1177/1469787402003002004.

Tree, D., A. Harvey, and M. Rand-Weaver. 2025. "Integrated Programme Assessment: 10 Years On." *Open Scholarship of Teaching and Learning* 4 (1): 166–180. https://doi.org/10.56230/osotl.128.

Tree, D., and N. Worsfold. 2026. "Do Synoptic Assessments Lead to Authentic Learning? A Critical Perspective on Integration and Intentionality in Higher Education Assessment Design." *Education Sciences* 16 (2): 187. https://doi.org/10.3390/educsci16020187.

van der Vleuten, C. P. M., and L. W. T. Schuwirth. 2005. "Assessing Professional Competence: From Methods to Programmes." *Medical Education* 39 (3): 309–317. https://doi.org/10.1111/j.1365-2929.2005.02094.x.

van der Vleuten, C. P. M., L. W. T. Schuwirth, E. W. Driessen, J. Dijkstra, D. Tigelaar, L. K. J. Baartman, and J. van Tartwijk. 2012. "A Model for Programmatic Assessment Fit for Purpose." *Medical Teacher* 34 (3): 205–214. https://doi.org/10.3109/0142159X.2012.652239.

Wake, S., M. Pownall, R. Harris, and P. Birtill. 2024. "Balancing Pedagogical Innovation with Psychological Safety? Student Perceptions of Authentic Assessment." *Assessment & Evaluation in Higher Education* 49 (4): 511–522. https://doi.org/10.1080/02602938.2023.2275519.

Walker, S. 2025. *Trends in Assessment in Higher Education: Considerations for Policy and Practice*. Bristol: Jisc.

Wiggins, G. 1990. "The Case for Authentic Assessment." *Practical Assessment, Research & Evaluation* 2: 2.