The Moderating Role of Assessment Diversity and Load on Student Outcomes

Abstract

Assessment design is a critical, yet underexplored determinant of student academic performance in higher education. Drawing on the testing effect and cognitive load theory, this study advances the proposition that the positive effect of assessment load on performance is moderated by assessment type diversity, which imposes extraneous cognitive load through task-switching demands. Using large-scale administrative data from UK higher education (N = 707,288 student-module observations), cross-classified multilevel models revealed a significant positive main effect of assessment load on module marks and a significant negative moderating effect of assessment type diversity. Subgroup analyses demonstrated that this moderating mechanism was substantially stronger among non-STEM than STEM students, theorised through Bernstein's distinction between horizontal and vertical knowledge structures. Robustness checks across three diversity operationalisations confirmed directional consistency. These findings offer theoretically grounded, empirically robust guidance for assessment design policy across disciplinary contexts in higher education.

Key words: assessment load, assessment type diversity, higher education, multi-level modelling

Introduction

Assessment design is widely recognised as one of the most powerful influences on student learning in higher education, shaping both what students know and how they engage with the curriculum (Gibbs & Simpson, 2004). Two structural features of assessment design have attracted particular attention, these are: assessment load and assessment type diversity. The former refers to the density of assessments within a module, while the latter indicates the range of different assessment formats employed. While existing scholarship has examined each of these dimensions in isolation (Author, 2026; Tomas & Jessop, 2019; Jessop & Maleckar, 2016), the effects of both on student academic performance remain poorly understood.

This study draws on two well-established theories from the educational and cognitive psychology literature to form a theoretical framework and, therefore, situate both the frequency and diversity of assessment within curricula. The testing effects theory suggests that frequent assessment strengthens long-term memory consolidation and promotes learning through repeated retrieval (Roediger & Karpicke, 2006; Roediger & Butler, 2011; Pan & Rickard, 2018), while cognitive load theory cautions that excessive extraneous demands, including the cognitive switching costs imposed by diverse assessment formats, may exhaust cognitive capacity and inhibit deep learning (Sweller, 1988; Sweller et al., 2019). Resolving this tension requires moving beyond the study of assessment load or type diversity in isolation, and examining how their interaction shapes student outcomes across different disciplinary contexts.

Based on large-scale administrative data from UK higher education, this study employs cross-classified multilevel models to examine how assessment load and assessment type diversity jointly predict student module performance, and whether this relationship varies systematically between STEM and non-STEM students. In doing so, it contributes large-scale empirical evidence to an area dominated by small-scale qualitative or survey-based research, and provides theoretically grounded guidance for assessment design policy.

Testing Effect and Assessment Load

Drawing on the conceptualisation of assessment as a sequence of retrieval opportunities embedded within a curriculum (Gibbs & Simpson, 2004), this study draws on the testing effect to theorise the role of assessment load in shaping learning outcomes. The testing effect, also referred to as retrieval practice, demonstrates that actively retrieving information from memory enhances long-term retention more effectively than repeated re-study techniques, such as reading (Roediger & Karpicke, 2006; Roediger & Butler, 2011). Within this paradigm, assessment load can be conceptualised as the frequency with which students are required to retrieve and reconstruct knowledge, acting as a key driver of memory consolidation and schema development (Sweller et al., 2019). The cognitive mechanisms underpinning this effect are well established: effortful retrieval activates and strengthens neuronal networks (schema) within semantic memory by creating more durable retrieval pathways (Endres & Renkl, 2015; Karpicke, 2017). Moreover, repeated and well-timed assessment provides structured feedback opportunities, which enable students to identify their known unknowns, gaps in understanding, adapt learning strategies, and achieve cumulative gains in performance (Hattie & Timperley, 2007; Pickering, 2017).

The empirical evidence for retrieval practice is substantial and is widely regarded as one of the most robust phenomena in educational psychology (Sigayret et al., 2026). A meta-analysis of 19 classroom-based studies reports an overall effect size of $d = 0.56$ (Schwieren et al., 2017), with Phelps (2019) reporting $d = 0.84$, indicating a consistent and meaningful positive impact of retrieval practice on learning outcomes in authentic educational settings. In higher education, undergraduates engaging in retrieval-based strategies consistently outperform those using elaborative encoding approaches (Brame & Biel, 2015), with particular gains among lower-attaining students (Pinto et al., 2020). Beyond consolidation, the forward effect of testing demonstrates that retrieval can enhance the acquisition and transfer of subsequently encountered knowledge (Issa et al., 2013; Pastötter & Bäuml, 2014), suggesting that frequent assessment may play a broader role in structuring ongoing learning processes through schema development and modification (Bjork & Bjork, 2011).

However, despite this foundation, evidence on assessment intensity in real-world higher education settings remains inconsistent. Some studies associate reduced assessment load with deeper learning and satisfaction (Jessop et al., 2014), while others report negative associations between high load and performance in both non-STEM (Yu et al., 2023) and STEM contexts (Griffith & Faulconer, 2024). A more recent large dataset analysis from a single centre,

provided evidence to suggest a positive association with assessment load and module performance (Author, 2026).

This inconsistency in findings highlights a critical limitation in the existing literature, with the tendency to treat assessment load as a uniform construct independent of its broader design context. In practice, cognitive demands are shaped by both assessment frequency and the diversity of formats through which assessment is delivered. Accordingly, this study advances the proposition that the relationship between assessment load and student performance is moderated by the structural diversity of assessment formats. While higher load may enhance learning through more frequent retrieval, these benefits may be attenuated or offset when combined with high assessment diversity that impose additional cognitive demands. This perspective provides a more nuanced account of when and why increased assessment enhances learning, and forms the basis for the analyses that follow.

Cognitive Load Theory and Assessment Type Diversity

Although Cognitive Load Theory (CLT) has predominantly been applied to the design of discrete learning tasks and lessons, its underlying principles can also inform curriculum design. CLT conceptualises learning as dependent upon the interaction between the processing demands imposed by a task, the learner's existing knowledge structures, and the limited capacity of working memory, with effective instructional design seeking to manage these demands in ways that facilitate schema acquisition (Sweller et al., 2019; Paas & van Merriënboer, 2020). At programme level, curriculum structures do not themselves constitute a cumulative cognitive load; rather, they shape the sequence, frequency and configuration of the learning and assessment tasks through which cognitive load is repeatedly tested. This distinction provides a basis for extending CLT to assessment approaches, with assessment load, temporal spacing and diversity all potentially influencing how frequently students encounter high-demand tasks, the extent to which successive assessments draw upon established schemas, and whether learners can progressively automate and integrate knowledge before encountering novel task requirements. Conversely, a coherent assessment sequence may manage intrinsic cognitive demands through progressive increases in complexity, repeated engagement with related knowledge structures and the gradual withdrawal of instructional support, principles already proposed for curriculum-level applications of CLT (van Merriënboer & Kirschner (2024); Leppink & Duvivier, 2016). From this perspective, programme-level assessment design can be understood as shaping the conditions under which task-level cognitive load occurs, rather than generating a continuous cognitive load across time. The distribution and sequencing of assessment may therefore constitute an important, but currently under-examined, mechanism through which curriculum architecture influences students' opportunities for schema construction, consolidation and transfer. Moreover, it is necessary to note that the focus of this manuscript is primarily focused on summative, or assessment of learning, rather than formative, or assessment for learning, forms of assessment. This study addresses this gap by applying the principles of CLT to the curriculum level, and treats the structural diversity of

assessment formats as a source of extraneous cognitive load that operates across the assessment programme as a whole.

Within higher education, students do not engage with assessments in isolation, they typically navigate multiple assessment events within a module, with each assessment imposing its own cognitive demands alongside ongoing consolidation and revision of material. Abolfathi et al. (2022) demonstrate that continuous and varied assessment can benefit students, but risks generating excessive cognitive workload when poorly calibrated. Moreover, Rai et al. (2026), examining diverse assessment formats in engineering programmes, similarly note that while interactive and varied formats can enhance motivation and foster deeper learning in some contexts, they simultaneously introduce additional cognitive demands that, if unmanaged, may impair performance.

For the purposes of this study, assessment type diversity refers to the range of assessment modes employed within a module. Higher diversity reflects the inclusion of multiple distinct formats, such as written examinations, coursework assignments, and oral presentations, whereas lower diversity indicates a more homogeneous assessment structure. Each format involves distinct modes of cognitive processing, strategic preparation for the learners, and temporal organisation and planning. Preparing for a written examination requires concentrated knowledge retrieval under time pressure, whereas producing a coursework assignment demands extended argumentative writing and source integration, whilst an oral presentation demands verbal articulation and real-time response to questioning. The continual shifting between these qualitatively distinct cognitive demands introduces extraneous load associated with task-switching and the reconfiguration of an individual learner's cognitive approach (Sweller, 2010; Paas & van Merriënboer, 2020).

Critically, the cognitive cost of assessment diversity is likely to be contingent on the level of assessment load, rather than being equally distributed. Therefore, it could be argued that when assessment load is low students are more likely to retain sufficient cognitive capacity to accommodate the demands of switching between assessment formats, with the potential benefits of assessment diversity, such as exposure to varied forms of knowledge representation and skill development, may then be experienced at limited cognitive cost. However, as assessment load and diversity increases, the frequency of retrieval events places progressively greater demands on cognitive processing. In this context, the additional extraneous load introduced by high assessment diversity compounds the intrinsic demands of frequent retrieval to generate a cumulatively higher cognitive burden (Sweller, 2010; Sweller et al., 2019). This interactive mechanism forms the central theoretical proposition of the present study.

Testing effect and CLT thus provide a coherent framework for potentially ~~potenitally~~ understanding how assessment load and assessment type diversity may jointly shape student attainment, a relationship that existing literature has not directly examined. Existing studies demonstrate that learners adapt their cognitive strategies in response to assessment format (Scouller, 1998; Struyven et al., 2005), and that assessment demands can negatively (Jessop et al., 2014; Yu et al., 2023) or positively (Author, 2026) affect performance, but these strands have been developed largely in isolation from one another. By integrating insights from testing

effect and cognitive load theory, this study proposes a unified theoretical account of how increased assessment load can enhance, or hinder, learning. To operationalise assessment type diversity at the module level, three complementary measures are employed: Shannon entropy, the Herfindahl-Hirschman Index (HHI), and the count of distinct assessment types (*n_types*), each capturing different aspects of distributional diversity across four assessment categories.

Disciplinary Heterogeneity of Assessment Structure

In addition, and drawing on Bernstein's (1971; 2000) theory of knowledge structures, the relationship between assessment load and assessment diversity is further expected to vary across disciplinary contexts. Specifically, Bernstein distinguishes between two forms of discursive organisation: 1) vertical discourse, in which knowledge is structured hierarchically and cumulatively, with new understanding integrating into and building upon prior knowledge; and 2) horizontal discourse, in which knowledge is organised in contextually embedded ways, with different concepts operating relatively independently to form an integrated hierarchical structure (Bernstein, 2000). STEM disciplines are paradigmatic instances of vertical discourse, while non-STEM disciplines exemplify horizontal discourse (Annala, 2022).

This structural distinction carries direct implications for the cognitive demands of diverse assessment environments. Under vertical knowledge structures, different assessment types activate overlapping and hierarchically related cognitive schemas: because new knowledge is built cumulatively upon existing understanding (Bernstein, 2000), preparing for one assessment type is likely to reactivate and reinforce the same underlying schema drawn upon by another. This shared cognitive logic across assessment types limits the extraneous cognitive load associated with switching between assessment types, which is consistent with cognitive load theory's argument that extraneous load arises primarily when task-switching requires the reconstruction of unrelated schemas rather than the extension of an already-activated one (Sweller, 2010; Sweller et al., 2019). . By contrast, under horizontal knowledge structures, diversified assessment types, including presentations, assignments and posters, draw upon relatively independent cognitive frameworks, with limited shared cognitive logic between them. For instance, the argumentative and analytical competencies required for a written assignment may differ substantially from the real-time verbal and interpersonal competencies required from an oral presentation (Tawfik, 2022). Preparing for each assessment type therefore requires students to construct a largely distinct cognitive schema rather than extend an existing one, which may increase the extraneous load incurred with each switch between preparation modes. Therefore, when assessment load is high, non-STEM students have to simultaneously manage the retrieval demands of frequent assessment and the high switching costs of cognitively distinct preparation modes. This theoretical expectation aligns with empirical evidence on systematic disciplinary differences in assessment type composition (Jessop & Maleckar, 2016) and the relationships between disciplinary knowledge structures and assessment practices (Neumann, 2001).

These considerations suggest that, at equivalent levels of assessment load and diversity, students in non-STEM disciplines may be more susceptible to the constraining effects of

assessment diversity on learning. Accordingly, this study examines whether the interaction between assessment load and assessment diversity varies across disciplinary groupings, to provide further insight into how assessment design operates within different educational contexts. While prior research has examined assessment load and assessment type independently, there is limited research integrating these perspectives to examine how the frequency and structural diversity of assessment jointly shape learning outcomes. Drawing on the theoretical frameworks outlined above and existing research literature, this study addresses the following research questions:

1. To what extent is assessment load, conceptualised as the frequency of retrieval opportunities, associated with student module performance, after accounting for individual background characteristics and institutional factors?
2. To what extent does the structural diversity of assessment types moderate the relationship between assessment load and student module performance?
3. Does the interaction between assessment load and assessment type diversity vary across disciplinary contexts, reflecting differences in cognitive demands between STEM and non-STEM fields?

Methods

Data source

Data for this study were sourced from a comprehensive administrative database at a research-intensive Russell Group university in the UK from the academic year of 2017/18 to 2024/25. This study focused on full-time undergraduate home students; part-time, postgraduate and international students were excluded from the analysis. The original dataset is structured in long format, with repetition in student ID; this has been reshaped into a wide format, where each row represents a unique student-module pair, with separate variables capturing the essential features of that module.

The dataset includes anonymised student characteristics, and the same students have been tracked and recorded as they progress through each year of study. In total, the dataset comprises 707,288 observations (STEM = 336,823; non-STEM = 370,465), including 52,420 students across 3,606 modules.

Dependent variable

The dependent variable was student module marks (ranging 1-100). In higher education, a module can be defined as a relatively self-contained unit of work within a programme of study (Dejene, 2019), characterised by clearly specified learning outcomes, a defined syllabus, and assessment requirements, and a unique identifying code. Outliers in the data, referring to observations with module marks deviating more than three standard deviations from the sample mean, were removed as suspected reporting errors.

Independent variable

Assessment load was operationalised as the number of assessments per 10 module credits to enable comparability across modules of different sizes. This normalisation accounts for systematic variation in credit weighting and avoids conflating assessment volume with module size. Using assessment density rather than raw counts also mitigates possible multicollinearity concerns arising from the strong institutional correlation between the number of assessments and credit value.

Assessment diversity is measured by Shannon Entropy, the Herfindahl-Hirschman Index (HHI) and the numbers of assessment types. In order to make results interpretable, we re-classified over 60 assessment types into four groups, including, written coursework, written examination, non-written coursework and non-written examination, according to the university's classification scheme. Inter-rater reliability (IRR) was conducted using Cohen's κ (Cohen, 1960) for each pair of raters and Light's κ (Light, 1971) as an overall measure of agreement between the three raters (Hallgren, 2012). The Light κ = 0.83, indicating almost perfect agreement (Landis & Koch, 1977). All diversity measures are calculated based on these four categorisations.

Shannon entropy (Shannon, 1948) captures the evenness of the distribution of assessment types, providing a measure of the informational richness of the assessment structure. Entropy is calculated as:

$$H = -\sum_{i} p_i ln p_i$$

where p*i* denotes the proportion of assessment type *i*. Values range from 0, indicating complete concentration in a single assessment type, to ln(4), indicating an equal distribution across the four categorised assessment types within the dataset and therefore maximum diversity (i.e. *Higher* values reflect *greater* diversity).

Secondly, the Herfindahl–Hirschman Index (HHI), originally developed to assess the concentration of foreign trade or steel industry in the US (Hirschman, 1945; Herfindahl, 1950), provides a complementary measure of concentration by summing the squared proportions of each assessment type. The index ranges from 1/4, indicating equal distribution across four types (maximum diversity), to 1, indicating complete concentration in a single type (no diversity). In contrast to entropy, *higher* values of HHI reflect *lower* diversity.

Thirdly, the number of distinct assessment types (n_types) provides a count-based measure of diversity. This is calculated as the number of assessment categories present within a module, ranging from 1 (a single assessment type) to 4 (all assessment types represented in the dataset). While this measure does not account for proportional distribution, it offers a transparent indicator of the breadth of assessment formats employed. Overall, these measures capture both the distributional and structural dimensions of assessment diversity, enabling a more nuanced examination of how diversity interacts with assessment load to shape student outcomes.

Table 1 presents descriptive statistics for the assessment structure variables across disciplinary groups. STEM modules exhibited substantially higher mean assessment density (1.84) than non-STEM modules (0.99). Shannon entropy and n_types values (i.e. number of distinct assessment types) indicate that most modules employed only a few assessment types.

Table 1 Descriptive statistics by disciplinary group

Variable	Overall	STEM	non-STEM
Assessment structure (M(SD))			
Assessment load	1.39(0.85)	1.84(0.95)	0.99(0.49)
Shannon entropy	0.22(0.31)	0.25(0.32)	0.19(0.31)
HHI	0.81(0.27)	0.79(0.27)	0.84(0.26)
n_types	1.36(0.62)	1.44(0.65)	1.29(0.59)
Assessment type composition (%)			
Written coursework	42.7	27.3	56.7
Written examination	42.1	56.1	29.4
Non-written coursework	6.6	8.2	5.2
Non-written examination	4.6	3.7	5.4

Assessment type composition differed markedly between groups. STEM modules were dominated by written examinations, followed by written coursework, reflecting the examination-centred assessment tradition. Non-STEM modules, by contrast, were predominantly assessed through written coursework, with written examinations comprising a smaller share. Non-written formats were minor in both groups. These compositional differences provide descriptive support for the premise that STEM and non-STEM students operate within structurally distinct assessment environments (Jessop & Maleckar, 2016).

We use Shannon entropy as the primary measure of assessment diversity, while employing the HHI and n_types as alternative specifications for robustness checks.

Covariates

The dataset provides student-level demographic characteristics (Appendix A), including sex, ethnicity, POLAR4 quintiles and UCAS tariff points. Sex is categorised into two-way: male and female; ethnicity into seven-way: White, Arab, Asian, Black, Mixed, Other and Missing. POLAR4 quintiles classify local areas into five groups based on young people's participation rates in higher education, from quintile 1 (lowest participation rates) to quintile 5 (highest). In the model, POLAR4 quintiles have been recoded into binary group: Q1-2 versus Q3-5. These categorical variables were incorporated into the multilevel model using standard treatment coding (dummy), with the most advantaged group (i.e. White, Female and Q3-5) as the reference group. Tariff points provide a standardised measure of entry qualifications, with higher scores reflecting stronger pre-university performance. The module-level control variable is module credits and a binary variable of whether this module has formative assessment.

Analytical framework

This study employs a multilevel linear model for analysis (Baayen et al., 2008; Gelman & Hill, 2007). The data structure features a cross-nested relationship between students and modules, where individual students enrol in multiple modules and each module is taken by multiple students. To account for this structure, we incorporate random intercepts for student, modules and institutions (module owner schools) simultaneously.

The analysis proceeds in three stages. First, a null model, which only contains random effects, is estimated to assess the relative variance contribution across level via Intraclass Correlation Coefficients (ICC). It confirms the necessity of a multilevel approach.

$$Y_{ijk} = \gamma_{000} + \mu_{0i} + \nu_{0j} + \varpi_{0k} + \epsilon_{ijk}$$

Second, we estimate the main model, including all fixed effects, the primary interaction term (assessment load*assessment diversity), and a comprehensive set of control variables.

$$Y_{ijk} = \gamma_{000} + \gamma_1 Load_j + \gamma_2 Diversity_j + \gamma_3 (Load * Shannon)_j + X_{ijk}\beta + \mu_{0j} + \nu_{0j} + \omega_{0k} + \epsilon_{ijk}$$

Finally, we conduct subsample analyses grouped by STEM and non-STEM disciplines to examine potential heterogeneity. The statistical significance of differences between these groups is formally tested using a full-sample model with a three-way interaction term (assessment load*assessment diversity*STEM).

$$Y_{ijk} = \gamma_{000} + \gamma_1 Load_j + \gamma_2 Diversity_j + \gamma_3 STEM_j + \gamma_4 (Load * Diversity)_j + \gamma_5 (Load * STEM)_j + \gamma_6 (Diversity * STEM)_j + \gamma_7 (Load * Diversity * STEM)_k + X_{ijk}\beta + \mu_{0j} + \nu_{0j} + \omega_{0k} + \epsilon_{ijk}$$

All analyses are implemented using the lme4 and lmerTest packages in R (Bates et al., 2015).

Results

Null Model and Variance Decomposition

A null model was specified to evaluate the hierarchical structure of the data (Table 2). The results of the null model indicated that the Intraclass Correlation Coefficients (ICC) at student, module and institution level were calculated as 34.7%, 10.2%, and 1.9% respectively, accounting for 46.8% of the total variance. These findings demonstrate significant

heterogeneity in student achievement across individuals, modules, and schools, thereby providing a robust empirical justification for the adoption of a cross-classified multilevel model.

Table 2 Variance Decomposition and Intraclass Correlation Coefficients

	Variance	Std.Dev.	ICC (%)
Random Effects			
Student	43.26	6.58	34.7%
Module	12.68	3.56	10.2%
Institution	2.38	1.54	1.9%
Residual	65.17	8.07	

Main Model Results

After controlling for individual- and module-level characteristics, the main model revealed significant effects of assessment load and diversity on academic performance. Regarding main effects, assessment load exhibited a significant positive association with module marks (β = 0.47) (Model 1 in Table 3). The main effect of Shannon diversity was also significant (β = 0.47), indicating that diversity in assessment types is, to an extent, positively correlated with performance.

While the effects of assessment load and diversity reached statistical significance, their practical magnitude warrants comment. A one-unit increase in assessment load was associated with a 0.47-point increase in module mark (out of 100). Across the observed range of assessment load (up to 7), this corresponds to a maximum predicted difference of approximately 3.3 points. This is smaller than the gap associated with several sociodemographic covariates. For instance, Black students scored 4.46 points lower than their White counterparts controlling for other variables. This gap suggests that although assessment structure is associated with attainment, its magnitude is modest relative to persistent attainment gaps related to student backgrounds, and should be interpreted as one lever, not the primary factor, among several relevant to closing attainment gaps.

Table 3 Fixed effects of main model and interaction model

	Model 1		Model 2	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	52.22 ***	0.56	52.39 ***	0.57
Assessment load	0.47 ***	0.05	0.29 **	0.10
Shannon Entropy	0.47 ***	0.14	1.66 ***	0.22
Tariff Point	0.09 ***	0.00	0.09 ***	0.00
Male	-1.43 ***	0.07	-1.43 ***	0.07
sex missing	-2.05 ***	0.48	-2.05 ***	0.48
Arab	-1.78 ***	0.47	-1.79 ***	0.47
Asian	-2.31 ***	0.11	-2.31 ***	0.11
Black	-4.46 ***	0.20	-4.46 ***	0.20
Mixed	-0.69 ***	0.14	-0.69 ***	0.14
Other	-1.74 ***	0.17	-1.74 ***	0.17

Ethnicity missing	-0.73	***	0.21	-0.73	***	0.21
POLAR Q1-2	-1.12	***	0.07	-1.12	***	0.07
STEM	-0.36		0.58	-0.70		0.59
Module credits	0.04	***	0.01	0.04	***	0.01
Formative assessment inclusion	-1.16	***	0.08	-1.16	***	0.09
Assessment load:shannon	-0.36	***	0.08	-1.31	***	0.17
Assessment load:STEM				0.25	*	0.11
Shannon Entropy:STEM				-1.68	***	0.32
Assessment load:Shannon:STEM				1.26	***	0.20

Note: ***p<0.001, **p<0.01, *p<0.05

The critical finding, however, was the significant negative interaction between assessment load and Shannon entropy (β = -0.37). This result suggests that as the diversity of assessment types increases, the positive effect of assessment load is systematically attenuated. In contexts of high diversity, this effect may even become negative.

In terms of control variables, Tariff points exerted a strong positive effect on marks. Female students significantly outperformed male students, and White students did much better than any other BAME students. Students from more disadvantaged areas (POLAR Q1-2) scored on average 1.12 marks lower than their more advantaged counterparts from POLAR Q3-5. ~~Socio-economic status (POLAR4) quintiles showed a significant positive gradient, reflecting the persistent impact of socioeconomic background on educational equity.~~ Notably, the presence of formative assessment was significantly associated with lower module marks.

Model 2 was conducted to examine the three-way interactive effects of assessment load, Shannon entropy and disciplines, which indicates a significant difference between STEM and non-STEM.

Sub-group heterogeneity

To elucidate the mechanisms of disciplinary heterogeneity, subgroup models were estimated for STEM and non-STEM subjects (Table 4). In the STEM group, assessment load exhibited a significant positive main effect on module marks (β=*0.56*). This suggests that in STEM contexts, higher assessment density facilitates learning by potentially reinforcing the effect of retrieval. The main effect of Shannon diversity was non-significant, which indicates that diversity in assessment types alone does not systematically influence performance for STEM students. The interaction between assessment load and Shannon diversity was not significant as well, which is different from the non-STEM group, suggesting STEM students' performance may be less likely influenced by assessment type diversity .

In the non-STEM group, the main effect of assessment load and Shannon entropy were significantly positive (β=*0.22/1.60*). Critically, the interaction between assessment load and Shannon diversity was substantially significantly negative (β=*-1.27*), with an effect size approximately 11.5 times larger than that of the STEM group. This finding demonstrates that in non-STEM contexts, the positive influences of assessment load on performance may be

impaired as assessment diversity increases, aligning with CLT's expectations regarding the additive effects of extraneous cognitive load.

The relative effect magnitudes within each disciplinary group displayed a similar pattern with the full-sample model. Across the full observed range of assessment load, the maximum predicted difference attributable to load alone was approximately 4 points for STEM students and 2 points for non-STEM students, which were both smaller than, or comparable to, the attainment gap associated with Black students in the same subgroup. Considering this modest magnitude, we therefore interpret our findings as indicating a robust and theoretically consistent pattern, rather than large-magnitude effects with practical consequence for individual students' attainment.

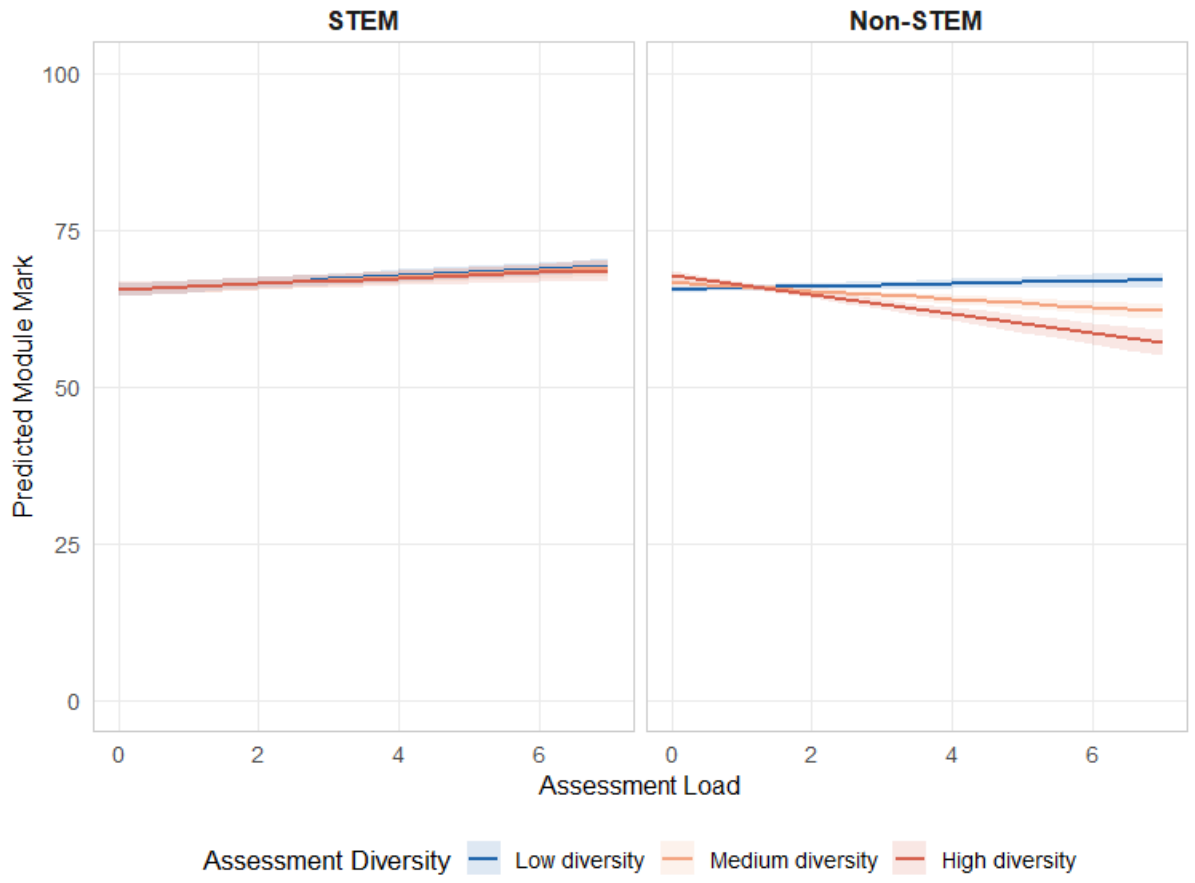
Table 4 Fixed effects of subgroup models

	STEM			Non-STEM		
	Estimate		Std. Error	Estimate		Std. Error
(Intercept)	50.62	***	0.74	53.34	***	0.49
Assessment load	0.56	***	0.06	0.22	*	0.10
Shannon Entropy	0.07		0.26	1.60	***	0.19
Tariff Point	0.10	***	0.00	0.07	***	0.00
Male	-1.70	***	0.11	-1.29	***	0.07
sex missing	-1.60	*	0.80	-2.19	***	0.53
Arab	-1.64	*	0.70	-2.22	***	0.60
Asian	-2.49	***	0.17	-1.90	***	0.13
Black	-5.40	***	0.33	-3.58	***	0.22
Mixed	-0.66	**	0.24	-0.60	***	0.15
Other	-1.76	***	0.28	-1.40	***	0.18
Ethnicity missing	-0.77	*	0.33	-0.64	**	0.23
POLAR Q1-2	-1.15	***	0.11	-1.04	***	0.07
Module credits	0.01		0.01	0.08	***	0.01
Formative assessment inclusion	-1.31	***	0.12	-0.65	***	0.13
Assessment load: Shannon	-0.09		0.11	-1.27	***	0.15

Note: ***p<0.001, **p<0.01, *p<0.05

Figure 12 illustrates the relationship between assessment load and predicted module marks at low, medium and high levels of Shannon diversity (with Shannon entropy values of 0, ln(2), and ln(4), respectively) presented separately for STEM and non-STEM students. In the STEM panel, all three regression lines show positive and parallel slopes, which may reinforce the argument above that academic performance for STEM students is less likely to associate with assessment diversity. .

Figure 12 Moderating effect of assessment type diversity on the assessment load-performance relationship



In the non-STEM panel, the three lines form a clear fan-shaped divergence. Specifically, under low diversity conditions the slope is positive, such that higher assessment load is associated with improved performance, whereas under higher diversity conditions the slope reverses to negative, indicating that higher load is associated with declining performance. This crossover pattern provides compelling visual evidence of the negative moderating role of assessment type diversity on the load-performance relationships, and demonstrates that this mechanism carries substantive practical significance in non-STEM disciplinary contexts.

Robustness check

To assess the sensitivity of the main findings to the choice of diversity measure, two alternative operationalisations of assessment type diversity, the HHI and the n_types, were employed as robustness checks. Both indicators were substituted for Shannon entropy in the full-sample main model, the interaction model, and the disciplinary subgroup models.

Appendix B profiles the results of models using HHI as the diversity measure. In the full-sample model, the positive main effect of assessment load remained significant. The load*HHI interaction term also reached statistical significance ($\beta=0.21$). It is important to note that HHI and Shannon entropy are directionally inverted and a positive load*HHI interaction is equivalent to a negative load*Shannon entropy, both indicating that greater assessment type

diversity attenuates the positive effect of load on performance. The three-way interaction model yielded a highly significant load*HHI*STEM term, confirming disciplinary heterogeneity. In the subgroup models, the load*HHI interaction was significantly negative among non-STEM students, fully consistent in direction with the Shannon sub-group result.

The count-based diversity measure (n_types) also provided consistent support for the main findings (Appendix C). In the full-sample model, the load*n_types interaction was significantly negative (*β=-0.17*), fully aligned with the Shannon model. The three-way interaction model yielded a highly significant load*n_types*STEM term, providing strong evidence for disciplinary heterogeneity. In the sub-group models, the load*n_types interaction was significantly negative among non-STEM students, while the STEM subgroup interaction was non-significant, which is a pattern consistent with the Shannon and HHI subgroup results.

Discussion

Drawing on large-scale administrative data at the student-module level in UK higher education, this study employed multilevel linear models to examine the joint effects of assessment load and assessment type diversity on student academic performance, and further investigated whether these effects varied systematically across STEM and non-STEM disciplinary contexts. We found that assessment load demonstrated a significant positive effect on module marks, but assessment type diversity exerted a significant negative moderating effect on this relationship, attenuating the positive effect of load. Furthermore, this moderating mechanism varied across disciplines in the following ways: it was significant among non-STEM students, whereas no significant moderation was detected among STEM students, indicating that the attenuating effect of assessment diversity is not a uniform feature of the load-performance relationship but is instead concentrated within specific disciplinary contexts. Robustness checks across alternative diversity operationalisations confirmed the directional consistency of these findings.

Assessment load and the testing effect

The robust positive association between assessment load and module performance is consistent with the core proposition of the testing effect: frequent assessment strengthens long-term memory consolidation and may promote deeper encoding through the repeated activation of retrieval processes (Roediger & Karpicke, 2006; Schwieren et al., 2017). Each retrieval event requires effortful reconstruction of knowledge from semantic memory, which may consolidate previously encoded material and also strengthen the neural pathways through which that knowledge can subsequently be accessed (Endres & Renkl, 2015; Karpicke, 2017). Under conditions of relative assessment type consistency, higher assessment density systematically enhances academic performance through the cumulative reinforcement of retrieval practice and structured feedback loops (Hattie & Timperley, 2007). This finding adds large-scale administrative evidence to a literature dominated by experimental and small-scale studies, and is consistent with the proposition that the positive effects of retrieval practice extend to higher education settings, although we do not intend to claim retrieval practice as the underlying mechanism based on current evidence.

Notably, the positive main effect of assessment load was robust across both disciplinary subgroups. This may suggest that the benefits of retrieval practice associated with assessment frequency operate as a relatively general feature of the load-performance relationship in higher education, independent of disciplinary context, whereas the extent to which this benefit is attenuated by assessment diversity is disciplinary-contingent. This distinction indicates assessment load, rather than assessment diversity, as a more generalisable lever identified in this study for associating with students' attainment across disciplines.

The moderating role of assessment type diversity

The central theoretical contribution of this study lies in its demonstration of the significant negative moderating role of assessment type diversity in the load-performance relationship, providing large-scale empirical support for the extension of CLT to the module level. As theorised, diverse assessment types impose extraneous cognitive load by requiring students to shift between distinct cognitive modes and preparation strategies (Sweller, 2010; Paas & van Merriënboer, 2020). Preparing for written examinations, extended assessments (i.e. coursework-based projects or essays), and oral presentations, each demand a fundamentally different cognitive orientation, with the former relying more on intensive knowledge retrieval under time constraint, to extended argumentative construction and real-time verbal articulation for the later. Each of these types requires the learner to configure and then reconfigure their cognitive approach, which introduces switching costs that are not incurred within a homogeneous assessment environment. When assessment density is low, students retain sufficient cognitive capacity to absorb these switching demands, and the potential benefits of diverse knowledge representation may be realised. However, as load increases, the cognitive demands of retrieval practice and the extraneous load of task-switching compound, to approach or exceeding working memory capacity and risk inhibiting the deep processing necessary for schema development (Sweller et al., 2019). The result is a progressive erosion of the conditions under which the testing effect operates, as cognitive resources are increasingly diverted toward managing format transitions rather than consolidating retrieved knowledge. This interactive mechanism, anticipated theoretically but not previously tested at scale, is consistent with the pattern observed empirically here. Note that this observational design cannot definitively isolate cognitive load as the mechanism driving this attenuation. For instance, it would not be tenable to fully rule out the possibility that modules with higher assessment load and greater assessment type diversity also differ systematically in the difficulty of each assessment session, which we were unable to directly test in the current study.

Disciplinary heterogeneity

The moderating effect of assessment diversity on the load-performance relationship was significant among non-STEM students, while no significant moderation was observed among STEM students. Importantly, descriptive statistics reveal only marginal differences between the two groups in mean diversity measures, indicating that the disciplinary heterogeneity

cannot be attributed to differences in the quantity of diversity. Instead, as theorised through Bernstein's (1971; 2000) distinction between vertical and horizontal knowledge structures, the critical difference may lie in the composition of dominant assessment types and the cognitive preparation demands they entail.

STEM modules are dominated by written examinations, reflecting the vertical knowledge structure of these disciplines, in which knowledge accumulates hierarchically and different assessment types activate overlapping and hierarchically related cognitive schemas (Bernstein, 2000; Annala et al., 2022). Even when STEM modules incorporate multiple assessment formats, the examination-centred preparation framework provides a relatively stable cognitive logic that may render the switching costs associated with additional assessment formats negligible, consistent with the absence of a statistically detectable moderating effect in this disciplinary group. Non-STEM modules, by contrast, are predominantly assessed through written coursework, consistent with horizontal knowledge structures characterised by contextually embedded and segmented knowledge organisation. Under horizontal discourse, different assessment types, such as essays, oral and poster presentations, draw upon relatively independent cognitive frameworks with limited shared logic between them (Bernstein, 2000). When non-STEM modules introduce additional assessment types alongside coursework, students must allocate limited cognitive resources across multiple preparation logics that are already cognitively demanding and structurally distinct, to risk substantially amplifying the overall extraneous load. This mechanism offers a theoretically grounded and empirically consistent explanation for the stronger moderating effect observed among non-STEM students, and aligns with evidence on disciplinary differences in assessment composition (Jessop & Maleckar, 2016) and the relationship between disciplinary knowledge structures and assessment practices (Neumann, 2001).

The formative assessment paradox

The negative association between the presence of formative assessment and module marks warrants careful interpretation and does not contradict the broader evidence base supporting formative assessment. Firstly, formative assessment was operationalised as a binary indicator of its presence or absence, without capturing the quality, timeliness, or student uptake that constitute the active ingredients of effective formative feedback (Morris et al., 2021; Hattie & Timperley, 2007). Secondly, modules incorporating formative assessment tend to exhibit higher overall assessment density, and within high-density assessment environments, students may resort to surface learning strategies to manage competing demands, and potentially offsetting the benefits of formative feedback (Tomas & Jessop, 2019). These mechanisms suggest that the net effect of formative assessment on performance is not a fixed property of the practice itself, but is highly contingent on the broader assessment structure within which it operates – a conclusion consistent with Bulut et al.'s (2025) finding that the effectiveness of formative assessment varies markedly depending on how it is operationalised and when students engage with it.

Implications

The findings in this study carry direct implications for assessment design in higher education. For non-STEM curriculum designers, introducing assessment type diversity should be accompanied by careful attention to overall assessment density. When a module requires multiple assessment formats, a corresponding reduction in assessment frequency may be necessary to prevent the compounding of extraneous cognitive loads from undermining performance. For STEM fields, the examination-centred assessment environment confers cognitive efficiency advantages through schema coherences; the absence of a significant moderating effect of diversity in this context suggests that modules in STEM discipline may have comparatively more flexibility to introduce diverse assessment types, although this interpretation should be treated cautiously . More broadly, the disciplinary heterogeneity documented here cautions against uniform institution-wide assessment reform policies, which risk producing divergent and potentially counterproductive effects across disciplinary groups. Assessment design sensitive to disciplinary differences in knowledge structure and cognitive demand is both theoretically warranted and practically necessary.

Limitation and Future Directions

This study has several limitations stated as follows. Firstly, reliance on secondary administrative data precludes capturing the nuanced and subjective experiences of students as they navigate assessment demands. Recoding a large number of distinct assessment activities into four broad categories inevitably masks meaningful within-category variation. However, without this aggregation, retaining a much larger number of detailed assessment categories would not adequately accommodate the variation in how similarly labelled assessment types are implemented in practice across modules. Furthermore, we were unable to directly measure or control for the difficulty of individual assessment components, which means we cannot fully rule out the possibility that modules with a higher assessment load simply comprise easier and more narrowly scoped tasks, rather than the load itself conferring a retrieval-practice benefit.

Secondly, the outcome variable of this study, module mark, is an aggregated composite across assessment components, which could possibly be sensitive to the assessment structure under investigation. Students may compensate for weaker performance on one assessment with higher marks in another, It is therefore possible that part of observed association between assessment structure and module mark reflects the arithmetic properties of score aggregation. Future research using assessment-component-level data will be required to disentangle these possibilities.

Thirdly, our operationalisation of assessment diversity is defined at the within-module level and does not capture potential cross-programme patterns of consistency or diversity in assessment designs. A student's overall cognitive load may depend not only on the diversity of assessment types within a single module, but also on whether those types are consistent or divergent across all modules they take in their undergraduate study. This represents a promising avenue for future research.

Finally, as a single-institution study, the findings may reflect specific institutional cultures. While the large-scale longitudinal data provides significant internal validity, the generalisability of these findings across diverse institution contexts requires further multi-institutional validation.

Conclusion

This study employs a multi-level hierarchical model to explore the joint effects of assessment load and assessment type diversity on student academic performance based on large-scale administrative data from a UK university. The results demonstrate that assessment load exerts a significant positive effect on module marks, consistent with the predictions of the test effect theory. However, this positive effect is not unconditional; assessment type diversity constitutes a significant negative moderating factor, where greater diversity systematically attenuates the positive effect of load, and in non-STEM contexts, renders the overall load-performance relationship negative. This moderating mechanism is theoretically grounded in CLT which describes how diverse assessment types impose extraneous cognitive load through task-switching demands, and risks exhausting students' cognitive resources under conditions of high assessment frequency. This finding enriches the theoretical understanding of assessment design in higher education and provides concrete practical guidance: assessment load and assessment type diversity may constitute a structurally interdependent combination that jointly determines student cognitive load within disciplinary contexts; practitioners should avoid treating them as independent design elements. Optimising assessment design in higher education, therefore, may require attending to the joint effects of these two dimensions against a backdrop of disciplinary difference, rather than pursuing higher assessment frequency or greater assessment variety in isolation.

Data availability statement for Basic Data Sharing Policy

The data that support the findings of this study are held by the host institution and are not publicly available due to their sensitive nature and legal restrictions.

Declaration of interest statement

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Reference

Abolfathi, A., Tahir, N., & Ahmed, K. (2022, September). Different types of assessments and their effect on students' learning and workload in remote learning. ISEE.

Annala, J. (2022). Disciplinary knowledge practices and powerful knowledge: a study on knowledge and curriculum structures in regions. *Teaching in Higher Education*, *27*(8), 1084–1102. https://doi.org/10.1080/13562517.2022.2114340

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390-412.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of statistical software*, *67*, 1-48.

Bernstein, B. (1971). *Class, Codes and Control: Theoretical Studies Towards a Sociology of Language*. Routledge.

Bernstein, B. (2000). *Pedagogy, Symbolic Control and Identity* (revised ed.). Rowman & Littlefield.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Brame, C. J., & Biel, R. (2015). Test-enhanced learning: the potential for testing to promote greater learning in undergraduate science courses. *CBE—Life Sciences Education*, *14*(2), es4.

Bulut, O., Gorgun, G., & Yildirim-Erbasli, S. N. (2025). The impact of frequency and stakes of formative assessment on student achievement in higher education: A learning analytics study. *Journal of Computer Assisted Learning*, *41*(1), e13087.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Dejene, W. (2019). The practice of modularized curriculum in higher education institution: Active learning and continuous assessment in focus. *Cogent Education*, *6*(1), Research-Article.

Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: An empirical investigation of retrieval practice in meaningful learning. *Frontiers in psychology*, *6*, 1054.

Gibbs, G., & Simpson, C. (2004). Does your assessment support your students' learning. *Journal of Teaching and learning in Higher Education*, *1*(1), 1-30.

Griffith, J. C., & Faulconer, E. K. (2024). Do We Need All Those Graded Assessments in This Course?. *Online Journal of Distance Learning Administration*, *27*(2), n2.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, *8*(1), 23.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112.

Herfindahl, O. C. (1997). *Concentration in the steel industry*. Columbia university.

Hirschman, A. O. (1980). *National power and the structure of foreign trade* (Vol. 105). Univ of California Press.

Issa, N., Mayer, R. E., Schuller, M., Wang, E., Shapiro, M. B., Darosa, D.A. (2013) Teaching for understanding in medical classrooms using multimedia design principles. Med Educ 47:388–396.

Jessop, T., El Hakim, Y., & Gibbs, G. (2014). The whole is greater than the sum of its parts: A large-scale study of students' learning in response to different programme assessment patterns. *Assessment & Evaluation in Higher Education*, *39*(1), 73-88.

Jessop, T., & Maleckar, B. (2016). *Studies in Higher Education*, 41(4), 696–711.

Karpicke, J. D. (2017). Retrieval-Based Learning: A Decade of Progress. *Grantee submission*.

Klionsky, D. J. (2008). The quiz factor. *CBE—Life Sciences Education*, *7*(3), 265-266.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.

Light, R. J. (1971). Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological bulletin*, *76*(5), 365.

Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, 9(3), e3292.

Neumann, R. (2001). Disciplinary differences and university teaching. *Studies in Higher Education*, 26(2), 135–146.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational psychologist*, *38*(1), 1-4.

Paas, F., & Van Merrienboer, J. J. (2020). Cognitive-load theory: Methods to manage working memory load in the learning of complex tasks. *Current directions in psychological science*, *29*(4), 394-398.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological bulletin*, *144*(7), 710.

Pastötter, B., & Bäuml, K. H. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in psychology*, *5*, 286.

Phelps, R. P. (2019). Test frequency, stakes, and feedback in student achievement: A meta-analysis. *Evaluation Review*, *43*(3-4), 111-151.

Pickering, J. D., (2017). Cognitive Engagement: a More Reliable Proxy for Learning? *Med.Sci.Educ*. (2017) 27:821–823

Pinto, C. M., Mendonça, J., Babo, L., & Ferreira, M. H. (2020, April). Assessment practices in higher education: A case study. In *2020 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1964-1968). IEEE.

Rai, N., Thakkar, Y., & Salvi, R. (2026). Effect of diverse assessment methods on undergraduate engineering students' performance in mathematics courses. *Journal of Research in Mathematics, Science, and Technology Education*, *3*(1), 24-36.

Roediger Iii, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20-27.

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249-255.

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The Testing Effect in the Psychology Classroom: A Meta-Analytic Perspective. *Psychology Learning & Teaching*, *16*(2), 179-196.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher education*, *35*(4), 453-472.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Sigayret, K., Parmentier, J. F., & Silvestre, F. (2026). Testing the testing effect on prolific: when retrieval practice fails to boost learning. *Frontiers in Psychology*, *17*, 1727423.

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & evaluation in higher education*, *30*(4), 325-341.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, *12*(2), 257-285.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review*, *22*(2), 123-138.

Sweller, J., van Merriënboer, J.J.G. & Paas, F. Cognitive Architecture and Instructional Design: 20 Years Later. *Educ Psychol Rev* **31**, 261–292 (2019). https://doi.org/10.1007/s10648-019-09465-5

Tawfik, M. (2022). Complexity and Interaction across Oral, Written and Online Discourse. *International Journal of TESOL & Education*, *2*(1), 272-295. https://doi.org/10.54855/ijte.222117

Tomas, C., & Jessop, T. (2019). Struggling and juggling. *Assessment & Evaluation in Higher Education*, 44(1), 1–10.

Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational psychology review*, *17*(2), 147-177.

Yu, D., Swanepoel, C., Sims, D., & Anciano, F. (2023). Do assessment loads affect student academic success? An investigation at an Economics Department at the undergraduate level. *South African Journal of Higher Education*, *37*(5), 237-259.

Appendix A Covariates distribution in the sample (%)

	Overall	STEM	Non-STEM
Female	62.2	58.1	63.7
White	79.2	77.2	81.1
Asian	0.4	0.5	0.3
Black	8.3	10.2	6.6
Mixed	2.4	2.5	2.2
Other	4.3	4.2	4.5
Missing ethnicity	3.1	3.1	3
POLAR1	8.2	8.6	7.4
POLAR2	11.9	12.7	10.8
POLAR3	15.1	15.6	14.1
POLAR4	19.7	19.5	19.6
POLAR5	40.6	39.7	43.2

Appendix B Robustness check via Herfindahl-Hirschman Index

	Full Model			Interaction Model			STEM			non-STEM		
	Estimate		Std. Error	Estimate		Std. Error	Estimate		Std. Error	Estimate		Std. Error
(Intercept)	52.33	***	0.58	53.60	***	0.60	50.68	***	0.77	54.44	***	0.52
Assessment load	0.29	***	0.07	-0.93	***	0.18	0.48	***	0.10	-0.96	***	0.16
HHI	-0.10		0.14	-1.18	***	0.25	-0.07		0.25	-1.11	***	0.22
Tariff Point	0.09	***	0.00	0.09	***	0.00	0.10	***	0.00	0.07	***	0.00
Male	-1.43	***	0.07	-1.43	***	0.07	-1.70	***	0.11	-1.29	***	0.07
sex missing	-2.05	***	0.48	-2.05	***	0.48	-1.60	*	0.80	-2.19	***	0.53
Arab	-1.79	***	0.47	-1.79	***	0.47	-1.64	*	0.70	-2.23	***	0.60
Asian	-2.31	***	0.11	-2.31	***	0.11	-2.49	***	0.17	-1.91	***	0.13
Black	-4.46	***	0.20	-4.46	***	0.20	-5.40	***	0.33	-3.58	***	0.22
Mixed	-0.69	***	0.14	-0.69	***	0.14	-0.66	**	0.24	-0.60	***	0.15
Other	-1.74	***	0.17	-1.74	***	0.17	-1.76	***	0.28	-1.40	***	0.18
Ethnicity missing	-0.73	***	0.21	-0.73	***	0.21	-0.77	*	0.33	-0.64	**	0.23
POLAR Q1-2	-1.12	***	0.07	-1.12	***	0.07	-1.15	***	0.11	-1.04	***	0.07
STEM	-0.37		0.58	-1.96	**	0.65						
module credits	0.04	***	0.01	0.04	***	0.01	0.01		0.01	0.08	***	0.01
Formative assessment inclusion	-1.17	***	0.08	-1.17	***	0.08	-1.32	***	0.12	-0.69	***	0.13
Assessment load: HHI	0.13	.	0.08	1.17	***	0.23	0.07		0.10	1.13	***	0.20
Assessment load: STEM				1.44	***	0.20						
HHI: STEM				1.20	***	0.34						
Assessment load:HHI:STEM				-1.14	***	0.25						

	Full Model			Interaction Model			STEM			non-STEM		
	Estimate		Std. Error	Estimate		Std. Error	Estimate		Std. Error	Estimate		Std. Error
(Intercept)	52.02	***	0.56	51.72	***	0.58	50.58	***	0.74	52.72	***	0.50
Assessment load	0.63	***	0.06	0.77	***	0.16	0.64	***	0.08	0.66	***	0.14
n_types	0.22	**	0.07	0.78	***	0.11	0.03		0.12	0.75	***	0.10
TariffPoint	0.09	***	0.00	0.09	***	0.00	0.10	***	0.00	0.07	***	0.00

Male	-1.43	***	0.07	-1.43	***	0.07	-1.69	***	0.11	-1.29	***	0.07
sex missing	-2.05	***	0.48	-2.05	***	0.48	-1.60	*	0.80	-2.19	***	0.53
Arab	-1.78	***	0.47	-1.79	***	0.47	-1.64	*	0.70	-2.23	***	0.60
Asian	-2.31	***	0.11	-2.31	***	0.11	-2.49	***	0.17	-1.91	***	0.13
Black	-4.46	***	0.20	-4.46	***	0.20	-5.40	***	0.33	-3.58	***	0.22
Mixed	-0.69	***	0.14	-0.69	***	0.14	-0.66	**	0.24	-0.60	***	0.15
Other	-1.74	***	0.17	-1.74	***	0.17	-1.76	***	0.28	-1.40	***	0.18
Ethnicity missing	-0.73	***	0.21	-0.73	***	0.21	-0.77	*	0.33	-0.65	**	0.23
POLAR Q1-2	-1.12	***	0.07	-1.12	***	0.07	-1.15	***	0.11	-1.04	***	0.07
STEM	-0.36		0.58	0.04	***	0.01						
module credits	0.04	***	0.01	-1.16	***	0.09	0.01		0.01	0.08	***	0.01
Formative assessment inclusion	-1.16	***	0.08	-0.58	***	0.08	-1.32	***	0.12	-0.65	***	0.13
Assessment load:n_types	-0.18	***	0.03	-0.04		0.62	-0.07		0.05	-0.56	***	0.07
Assessment load:STEM				-0.15		0.17						
n_types: STEM				-0.78	***	0.15						
Assessment load:n_types:STEM				0.53	***	0.09						

Appendix C Robustness check via n_types