Examples of low-burden inclusive assessment strategies

This appendix sets out the outcome of a literature review to identify successful inclusive assessment strategies. The majority are from North America rather than the UK. The core reason for this outcome seems to be due to different regulatory environments; the regulatory burden is different in the US with module leaders having considerable autonomy in module delivery and design. Indeed, regulatory burden was mentioned as a barrier in the staff survey.

Across the STEM disciplines, research points to a set of assessment practices that are both low-burden for staff and demonstrably inclusive for students, even when deployed in very large classes. Although almost all of the peer-reviewed evidence originates in North America, the findings are readily transferable to the UK because the interventions rely on routine digital platforms (for example, institutional virtual-learning environments or optical-mark reading of paper quizzes) rather than on any country-specific infrastructure. Nevertheless, some Warwick-specific platforms lack the capability to implement the suggested assessments.

In biology, large-enrolment introductory courses at the University of North Carolina adopted a “moderate-structure” model in which weekly, auto-graded quizzes and in-class clicker questions replaced a single mid-term (Eddy and Hogan, 2014). This shift raised average marks for all students and, crucially, halved the Black–white grade gap by providing frequent feedback without increasing staff marking time. A complementary technique is the two-stage exam, trialled with hundreds of life-science undergraduates in the United States: students sit an exam individually and then immediately retake the same questions in self-selected groups, an arrangement that improves final scores and lowers test anxiety while adding only minutes of extra invigilation effort (Gilley and Clarkston, 2014). A third inclusive biology strategy uses structured “exam-wrapper” assignments; after each test, students diagnose the source of their errors and can regain partial credit by submitting corrections. A recent controlled study showed that test-wrapper tasks especially benefit lower-ACT entrants, thereby narrowing preparation-based gaps without adding extra marking beyond a brief rubric check (Angell et. al.., 2024).

Evidence from chemistry confirms that small, frequent assessments are equally powerful. In a 1300-student General Chemistry sequence, mastery-grading software gave students multiple attempts to meet each learning objective and produced a seven-percentage-point gain on the common final for under-represented minority students while leaving overall staff workload unchanged (Hartman et. al.., 2024). A second inclusive option, again scalable because it uses automated quiz banks, is simply to replace one or two high-stakes tests with weekly online quizzes; a meta-analysis of 52 experimental studies across quantitative disciplines found an average improvement of 0.42 standard deviations in final marks and a halving of D/F rates when weekly quizzes were present (Sotola and Credé, 2021). Where laboratory time allows, chemistry instructors have also borrowed the two-stage exam format; students' anxiety falls and conceptual scores rise because the peer stage immediately repairs misconceptions with no extra grading beyond recording a second answer sheet (Rieger and Heiner, 2014). All three approaches have been implemented in cohorts of several hundred students, so their viability at scale is well established.

Turning to physics, Webb and Paul (2023) analysed four parallel sections of introductory mechanics and showed that allowing voluntary, low-stakes retake examinations entirely removed the gender performance gap that had persisted for years under a one-and-done regime; the administrative burden amounted to releasing an additional practice paper each fortnight and re-using the same marking rubric. Ives (2014) demonstrated that two-stage collaborative testing could be run for classes of over 200 physics majors with no loss of rigour and with significant gains in individual understanding. A third physics innovation is the concept-first assessment sequence reported by Webb and Paul (2023), in which early tests emphasise qualitative reasoning before moving to algebraic manipulation; the authors found no ethnic attainment gap on the common final once the conceptual emphasis was in place.

In mathematics, a multi-institutional study of inquiry-based learning (IBL) courses revealed that replacing routine lectures and algorithmic exams with student-presented proofs and open-problem write-ups improved outcomes most for those entering with the weakest prior preparation and closed the long-standing gender attainment gap (Laursen et. al.., 2014). Mastery testing offers a lighter-weight alternative: DeGoede (2022) enabled first-year calculus students to reattempt five-point quizzes until they reached mastery, cutting self-reported test anxiety while requiring no extra instructor grading thanks to an online engine

that regenerated isomorphic questions. Finally, the analysis by Sotola and Credé (2021) confirms that simply inserting short, auto-graded quizzes each week lifts mathematics course pass rates across institutions. It is worth noting that a recent Office for Student investigation[1] under registration condition B4 means caution must be taken when designing such mastery assessments.

For computer science, early work by Werner, Hanks and McDowell (2004) showed that pair-programming assignments in CS1 sections of 100 students raised pass rates for women from 68 per cent to 88 per cent while having no detrimental effect on men, and the marking load remained unchanged because only one submission per pair was graded. Guzdial's (2003) "media computation" projects, used with cohorts of several hundred non-majors, almost eradicated the gendered withdrawal/failure gap by letting students express programming concepts through creative image and sound manipulation; staff workload was balanced by clear, automated test suites for the code. More recently, Shah (2024) piloted a competency-based grading contract across three programming courses and reported higher overall pass rates together with narrower ethnicity gaps, the administrative cost being limited to one brief progress-tracking meeting per student mid-semester.

Finally, within engineering education, Koretsky et. al.. (2022) embedded an authentic design problem into the group phase of a two-stage exam administered to a 150-student thermodynamics class; students from demographic groups that usually under-perform on timed analytical tests showed disproportionately large score gains, and the intervention required no extra marking because the group sheet was graded with the same rubric as the individual sheet. Women's self-efficacy in engineering design improved markedly when capstone courses introduced structured role rotation and peer-reviewed teamwork rubrics, as documented in a 145-student aerospace programme (Camarillo and Camfield, 2020). A much lighter but equally scalable practice is the reflective learning portfolio: Williams (2002) reported that asking every final-year engineering student to curate evidence of meeting programme outcomes, accompanied by short reflective commentaries, provided a richer picture of competence without adding new assessments, and students who historically fared poorly on closed-book exams responded positively to that latitude.

[1] Office for Students, (2024). Regulatory case report for the University of Wolverhampton – Finding of a breach of condition B4. Published 3 December 2024. Available online: https://www.officeforstudents.org.uk/media/is1cbe3b/uow-case-report.pdf

In summary, the peer-reviewed literature offers at least three large-scale, low-burden and more inclusive assessments for STEM disciplines. Most of the evidence derives from North American institutions, yet the technologies involved – weekly auto-graded quizzes, repeatable online mastery tests, group answer sheets/multi-stage assessment, or portfolio uploads – are readily available in UK higher-education contexts, making local adoption possible. Certain assessments, such as multi-stage assessments or portfolio assessments do represent challenges from quality assurance and also reasonable adjustment perspectives. For example, how might a two-stage examination function? Nevertheless, a problem-solving mindset that seeks to address these UK-based challenges should be able to overcome these issues.