DOI: 10.5324/njsteme.v6i1.4873

Maintaining quality assessment practices in Norwegian higher education after the two-evaluator law

Yael Harlap[1], Christian Jørgensen[2] and Sehoya Cotner[2]

[1]Department of Education, University of Bergen, Norway

[2]Department of Biological Sciences, University of Bergen, Norway

Abstract: In May 2021, the Norwegian parliament voted unanimously to again require the use of two evaluators to assess all student work given a grade on the A-F scale in higher education. This revision of the law regulating higher education marks a return to a rule that had been rescinded with the Quality Reform of 2001, and has the potential to lead to a cascade of negative consequences for the quality of practices in higher education. We first provide an overview of the problem, and then offer practical, constructive, and evidence-based suggestions for how instructors can meet these requirements while still offering students opportunities to gain formative feedback, to engage in deep and meaningful learning, and be assessed in ways that are aligned with the intended learning outcomes of the course.

Keywords: Assessment, constructive alignment, formative assessment, grading, Norwegian higher education, learning outcomes

I think it will be twice as safe, although it will require more work.
- Henrik Asheim, then-Minister of Research and Higher Education, 2020

When the how's of assessment preoccupy us, they tend to chase the why's back into the shadows.
-Alfie Kohn, 2006

INTRODUCTION

The aim of this paper is to offer some suggestions on how engaged and motivated instructors can meet the new requirement of two evaluators for all A-F grades in Norwegian higher education while still offering students opportunities to gain formative feedback, to engage in deep and meaningful learning, and be assessed in ways that are aligned with the intended learning outcomes of the course. Our goal is to be practical, concrete, constructive, and even inspirational. Before we offer alternatives, we provide an overview of the current problem in Norway, but the reader should note that our points and subsequent suggestions are relevant across higher education, as instructors grapple with the challenge of providing students with meaningful assessment without an excessive demand on instructor time.

THE LEGAL BASIS

The Quality Reform of 2001 (Regjeringen, 2001) intended to shift Norwegian higher education from being comprised of "exam-driven institutions" (Regjeringen, 2001) to having a greater focus on higher-level learning, constructive alignment between stated learning outcomes, learning activities and assessment (Biggs & Tang, 2010), and greater use of formative assessment. The government thus removed a requirement for two evaluators for every assessment of student learning and was explicit as to why: "new methods for teaching and assessment call for a change in the use of external evaluators (…) The introduction of more frequent assessments and feedback to students make it less appropriate to use external evaluators to evaluate all exams" (Regjeringen, 2001, p. 32).

However, in May 2021, the Norwegian parliament voted unanimously to again *require* the use of two evaluators to assess *all student work given a grade on the A-F scale in higher education*. This revision of the law regulating higher education marks a return to a rule that had been rescinded with the Quality Reform. The two new sentences read (§3-9 (2), our translation): "There shall be at least two evaluators for all assessments where a grade scale of A-F is used. At least one of the evaluators shall be without involvement in that part of the education where the person will be an evaluator." Although implementation was postponed by parliament in June 2022, this rule seems likely to become active law from August 2024.

There is also a shift in how the law about assessment of student performance is described in the government's description and interpretation. Formerly, the description of grade-setting was predicated on the need for "evaluative judgment" (skjønnsmessige vurderinger); now, the focus is on "students' perception of fairness" (Regjeringen, 2021, p. 48): "Such can two evaluators provide a more neutral and independent evaluation of the student's performance. This can therefore increase the students' legal security/right and contribute to students being able to feel more secure that the grade is correct."

THE TWO-EVALUATOR CHALLENGE

The use of two evaluators is a long-standing tradition in Norwegian higher education. Colleagues (often the course instructor and an instructor from another institution) separately assign provisional grades and confer to make a final determination when their initial grades differ. This practice is also commonplace in the United Kingdom and to a degree in Australia, and in English is referred to as second marking, double marking, or (more generally) moderation or external examination (Beutel et al., 2016; Bloxham, 2009; Bloxham et al., 2016; Bloxham & Price, 2015; Smith, 2012). After the Quality Reform, the use of two evaluators remained mandatory for thesis work and oral exams, but became optional for other grading assessments. National student organizations lobbied for the new requirement, though some leaders in student politics have since changed position and argued for the law to be rescinded before it takes effect. And faculty value the opportunity to discuss grading with a peer, both because they feel it helps ensure fair grading, and because it opens up developmental conversations about education in the discipline. But that two evaluators may be a preferred choice does not justify it as a *mandatory* rule for *all* graded assessments imposed on the higher education sector, in the belief that two evaluators are always better than one.

There are a number of reasons why we, and most of the institutional actors in our sector, argue against an absolute requirement for two evaluators. There are principled arguments anchored in autonomy – for institutions, departments and instructors to make pedagogical decisions about how they structure their educational offerings without government interference. There is some research evidence that evaluators regularly disagree and that two evaluators may provide a fairer assessment than one, but the effects are variable and depend on subject discipline (Rye, 2014, Bonsaksen et al., 2018). Then there are the arguments that posit that extensive use of two evaluators will drain resources away from research-based pedagogical practices in teaching and assessment. Because external evaluators in Norway are paid for their work, and because there are never enough of them available, a significant increase in their use will result in greatly increased costs to each department in time and money – without any increase in resources from the government to meet these expenses, and in fact in the face of significant budget cuts to higher education in 2022.

This will lead to a cascade of negative consequences for the quality of educational practices. Teachers will use more hours for finding and guiding external evaluators, and evaluating and grading their students. They will use more time working outside of ordinary work hours as external evaluators for other institutions. They will have less time each semester for planning, teaching and giving ongoing feedback to their own students (Figure 1). More time set aside for grading means less time set aside for teaching and feedback activities, nudging higher education teaching practices towards: more lecture-format teaching with larger groups; fewer formative assessments and activities with feedback; less time for student-faculty interaction; and exams that emphasize simple questions to minimize disagreements between evaluators (Figure 2). Although possible, it is rare that simpler exam types are designed to assess learning at higher levels of cognitive complexity, which is what we care most about in higher education. Thus, we anticipate a dramatic reduction in the implementation of evidence-based teaching approaches, and a regressive shift in pedagogical practices that significantly erodes two decades of development in Norwegian higher education.

DOI: 10.5324/njsteme.v6i1.4873

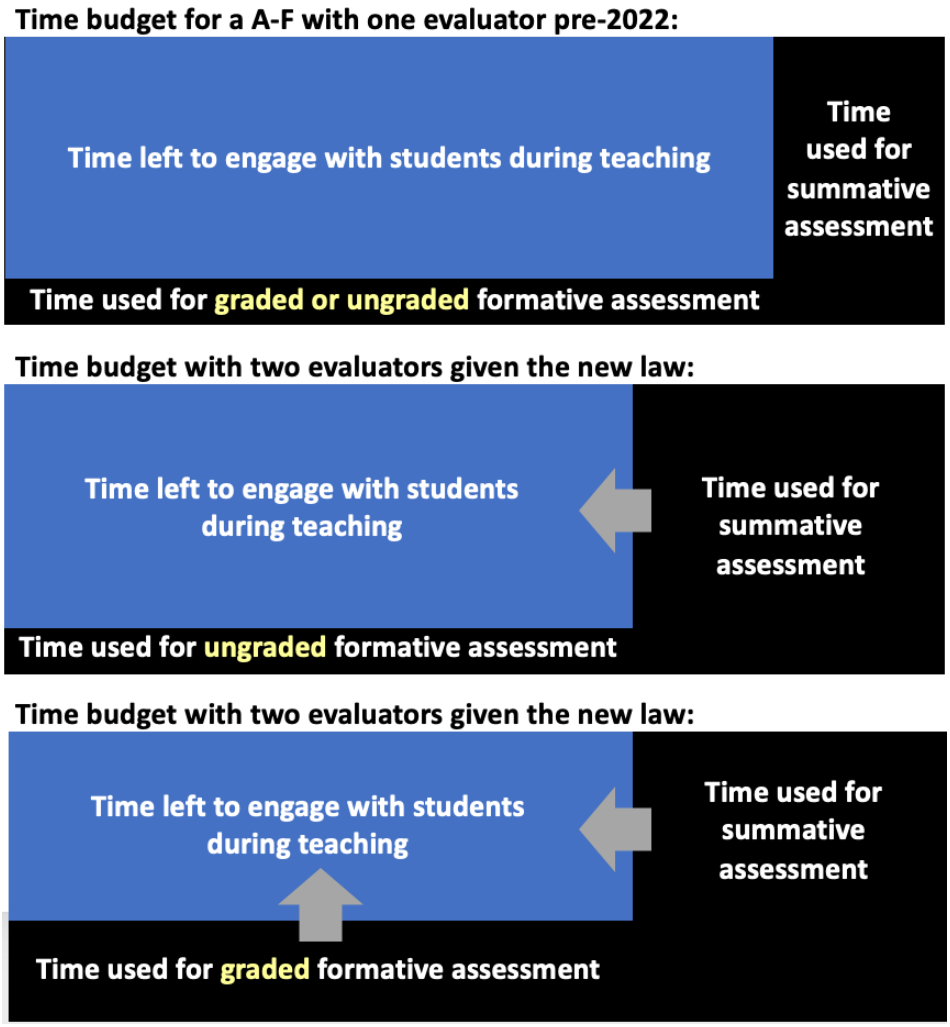

Figure 1. Visualizing the problem. Likely effects on teacher time budgets if the new legal requirement of two external evaluators for all A-F exams is met with business-as-usual teaching practices. Critically, as we use more time for assessment outside of class, and assuming finite time for teaching-related activities, the loss of time to engage with students is palpable. Note that this effect will vary with different types of courses and assessments, with some activities associated with a dramatic increase in effort, whereas other will be less affected.

DOI: 10.5324/njsteme.v6i1.4873

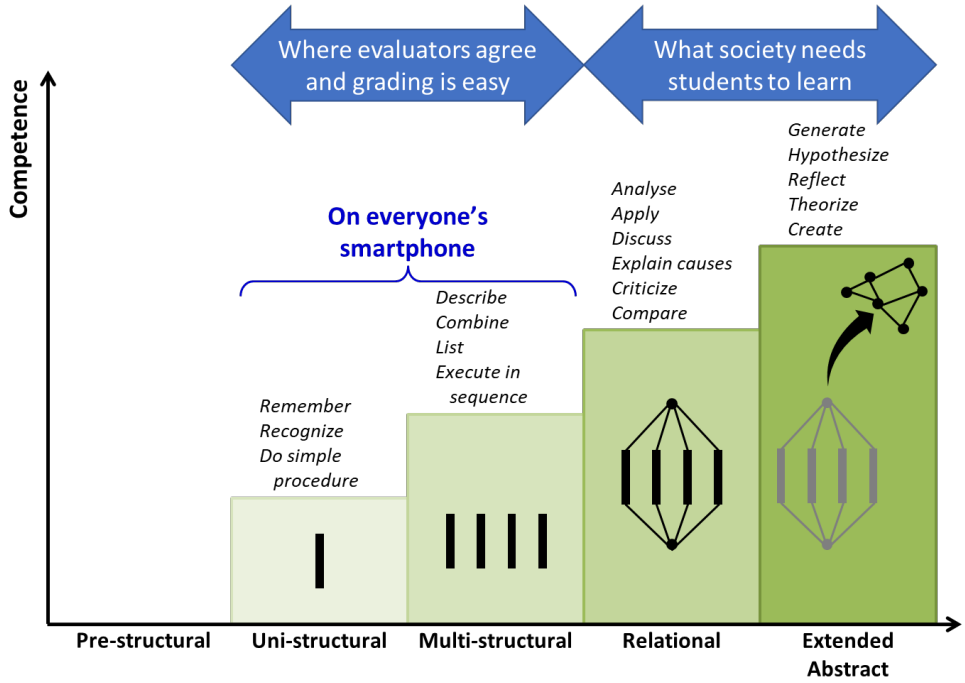

Figure 2. When two evaluators are required for every graded work, institutions may be tempted to sacrifice assessing higher-level thinking to save resources. This figure is a modified version of John Biggs' Structure of the Observed Learning Outcome (SOLO) taxonomy. The purpose of the SOLO taxonomy is to categorize learning activities by degree of cognitive complexity. Though students must have lower-levels of knowledge in place to be competent in their discipline or profession (the uni-structural or multi-structural levels), in higher education we expect that most program and course learning outcomes will be at higher levels. John Biggs also argues that assessments should assess at these higher levels of complexity. Unfortunately, assessments that are quick and easy to grade, and simple to come to consensus on across evaluators, are *more likely* to fall in the low-complexity end of the SOLO taxonomy.

In the rest of this paper, we offer an array of assessment strategies for maintaining educational quality given that the new law is likely to 'steal' time and resources away from formative teaching and learning activities, and more complex and sophisticated forms of end-of-term summative assessment. Our approach is two-fold: (1) we argue first the benefit of increasing the use of pass/fail courses, before (2) we describe concrete assessment strategies for A-F graded courses.

A general critique of grading: Why moving to pass/fail solutions can be a win

The intention behind the new law is to ensure that students receive fair grades that accurately reflect their performance and, presumably, their learning. There is a significant body of literature that critiques grading on multiple grounds. We do not have time to engage these arguments in depth, but will name some of the most serious critiques:

- grades make teaching and learning transactional (Stommel, 2020). They enhance students' instrumental orientation to education, where the goal becomes to score high grades, and reduce their internally driven motivation to learn (Boud, 1990; Shepard et al., 2018b);
- grades incentivize wrongly – including weighting the teacher's judgment and control over the student's growing capacity to make evaluative judgments about quality (Stommel 2020). They generate a preference for easier tasks, and nudge students towards shallow thinking, or 'surface learning' (Boud & Falchikov, 2006; Leenknecht et al., 2021; Shepard et al., 2018a, 2018b; Wiggins, 1991; Jørgensen & Bråten, 2019);
- grades provide meagre feedback (Kohn, 1994);
- grades don't typically correspond well to learning (rather they correspond to how well students follow instructions) – and even when they adequately capture something related to predetermined learning outcomes, they don't reflect the subjective, idiosyncratic nature of learning (Stommel, 2020);
- grades reward competitiveness and disincentivize collaboration (White & Fantone, 2010);
- grades have inconsistent meanings – there are too many variables, biases and pitfalls for grades to ever be genuinely fair (Stommel, 2020).

In contrast, a shift to pass/fail courses can lend itself to deeper learning, more intrinsically motivated students, more openness to collaboration, lower stress and anxiety, and reduced incentive for academic dishonesty (Bloodgood et al., 2009, Reed et al., 2011; Rohe et al., 2006; White & Fantone, 2010). In recognition of these benefits, the Law faculty at the University of Bergen recently transitioned to pass/fail courses for 36 of 60 credits (ETCS) in the first year of legal study. Medical education at the University of Oslo was entirely pass/fail from 1996 until 2014 (Frich et al., 2014). In addition, even in A-F graded courses, the existing course structure in Norwegian higher education encourages the use of pass/fail *formative* assessment (Dahl et al., 2009; Dahl, 2006) by formally separating the summative assessment (typically called a 'course exam', regardless of its form) from 'obligatory assignments' students must complete before they can sit for their exam.

We do not want to dismiss entirely that there can be genuine arguments in favor of grades as well. For example, the introduction of grades in legal education in Norway was tied in large part to activism by organized students from farming families who struggled to break into government and the judiciary, where positions were dominated by the sons of the wealthy elite. Meritocracy is the core argument in favor of grades: that objective measures of performance can level the playing field for students from marginalized groups. The idea that grades actually serve meritocratic ends has also been critiqued, however, especially on the grounds that they do relatively little to resolve equity issues, also in Norwegian legal education (see Guinier, 2015; Hansen & Strømme, 2021; and Mijs, 2016 for discussion). In this paper we will suggest strategies for transitioning at least some courses in a study program to pass/fail.

DOI: 10.5324/njsteme.v6i1.4873

TRANSFORMATIVE ASSESSMENT ALTERNATIVES IN PASS/FAIL COURSES

Our first recommendation for Norwegian academics and institutions is to increase the use of pass/fail courses for pedagogical as well as practical reasons. The Norwegian system allows for courses to be offered on a pass/fail basis – the current limited use of pass/fail is due to institutional, instructor and student expectations rather than law or regulation. Pedagogically speaking, pass/fail courses vault over the critiques of grading and open up opportunities for engaging in the creation of knowledge products with learning rather than sorting as the central goal. Obligatory activities become the basis for assessing student learning and performance. In this short paper, we are not able to discuss course design in depth, but we wish to highlight several new opportunities as well as a few practical considerations in the process of transforming assessment from A-F grading to pass/fail.

Making feedback meaningful, and possible: Research on assessment strongly suggests that meaningful feedback is critical to learning, that feedback should come quickly after performance, and that students should be asked to engage actively with feedback in order to 'close the feedback loop' and ensure that feedback is not ignored (Biggs & Tang, 2010). For example, students benefit greatly from the opportunity to evaluate and address feedback, and revise their work – whether or not it will be graded – much like academics' work often benefits from feedback from reviewers. Peer learning (see also peer assessment, peer feedback, peer review, and peer grading) is a well-tested pedagogical strategy (see, for example, Liu & Carless 2006, Reinholtz 2016, Sadler & Good 2006). For example, at Cleveland Clinic Lerner College of Medicine, developing competencies rather than collecting grades is at the core of all assessments, and there is a strong focus on students' developing a habit of growth from peer feedback (Altahawi et al., 2012; Dannefer, 2013). For peer review to work well, students need to develop metacognitive skills (the ability to think about their thinking and have awareness about their own level of understanding and their learning process). They need explicit guidance from the instructor and opportunities to practice. Thus, adding peer review to a course does demand some time, particularly in-class time for guiding students on how to evaluate and comment constructively on each other's work. But peer-review can require much less instructor time than instructor feedback, which is difficult to find time for even in pass/fail courses. At a minimum, instructors can leverage peer assessment in ways that ultimately reduce the amount of material the instructor must evaluate to determine final grades (Aahlberg and Lorås 2018). We recommend Saundra McGuire's work on teaching students how to learn (McGuire & McGuire, 2015), with its emphasis on strategies for meta-cognition and self-reflection that can be adopted for in-class peer-review.

Group work: Many skills and general competencies that are in high demand in the world of work (e.g., creativity, interdisciplinarity, leadership, TNS Gallup, 2015; adaptability, entrepreneurship, OECD, 2018; communication, collaboration, Støren et al., 2019) are best developed through groupwork. Furthermore, students can often benefit from learning or teaching challenging content in groups of peers, because slightly-more-advanced novices can be better than experts at recognizing the cognitive challenges that confused novices face (Lockspeiser et al., 2008). In pass/fail courses, group work can be sophisticated and extensive, while concerns about free-riders are lessened in a lower-competition situation. The pedagogy

of Team-Based Learning (TBL) can be a useful approach; it is designed to avoid common pitfalls that often cause students to dislike groupwork, and can help instructors foster well-functioning groups (e.g. Sibley & Ostafichuk, 2014, chapter 5).

Creating lasting products: As students do not leave a pass/fail course with a grade, what do they leave with? Although learning is clearly the central goal, there are also opportunities for designing assignments that can serve as lasting products: authentic assessments that students can showcase for future employers (such as a digital portfolio, a video, a website, a conference-style poster as well as the more traditional research paper or essay). *Authentic* assessments are ones where the assessment situation mimics the true performance situation (Wiggins, 1991) — that is to say: how might the students put the knowledge and skills they have gained into practice in a future workplace? Examples of authentic assessment vary by discipline, and could include: writing a lab report based on the student's own inquiry experience, developing a public-service announcement that distills contemporary knowledge about a health issue, critically evaluating an expert report, or making a video review of an art installation.

Imagine transforming an entire educational degree: at the end of a Bachelor degree, instead of having a transcript with opaque grades (what do they really mean? What did the students have to do to earn them?), students have a digital portfolio of meaningful lasting products they developed in their courses. This sounds radical, and perhaps it is, but might be a logical next step given how higher education is being offered in ever more disciplines (nursing, daycare teacher) where classical, theoretical understanding is balanced by practical skills or where employers in the new knowledge economy desire to hire for both knowledge and creativity.

What does a "pass" mean? A grade of pass should represent a meaningful degree of competence in relation to course learning outcomes; how high one sets a pass standard can vary by course but must be consistent within a given course. Passing a course does not need to mean students can just barely scrape through; for life-critical skills such as brain surgery or operating a nuclear power plant, the threshold for pass should probably correspond to an A. Rather, we can think that passing a course suggests that we are certifying that students have *adequately* met the learning outcomes. Pass can mean different things in different contexts, just like letter grades already do; we don't expect 1st year BA students to be able to do 1st year MA-level work, but they can still earn As in the first year, which means not all As have the same meaning. Ideally, and especially when we set high standards for passing pass/fail courses, students should be able to revise and redo (one, or some) assignments, and courses should be designed such that weak understanding the first time a student works with new knowledge doesn't torpedo their performance in the entire course. Note that under Norwegian law regulating higher education, pass/fail work is not anchored to letter grades. To tell students "you need to produce C-level work to pass this course" is not a useful way to 'ungrade' in any case; instead, we need to describe criteria for passing in a meaningful way. Read about specifications grading below for inspiration.

Study program design and pass/fail courses: It is critical to think about assessment design across a whole study program, both to ensure meaningful learning experiences and due to the incentive structure that A-F grades carry with them. If students take parallel courses with a mix of pass/fail and A-F grading, they will be incentivized to prioritize the A-F courses, to

the detriment of learning in the pass/fail courses. Further, it is worth reflecting on at what point in a study program's trajectory it is meaningful (even fair) to assign letter grades. Imagine, for example, on the bachelor level, that the first two years were fully pass/fail – with solid structures for feedback and 'closing the feedback loop' – and the third year courses all received A-F grades. In this model, students do not receive letter grades at early points when their individual academic maturity is likely to have high variability. However, by the third year they are expected to have reached a level of academic performance that is a fair reflection of their knowledge, skills and competences as they are closer to completion of a degree. Reducing the number of courses in a study program that are graded on an A-F scale also reduces the number of evaluators across the study program as a whole. Because time and availability of academics to act as evaluators is limited, freeing up time early in a study program allows for a reallocation of resources towards teaching activities and formative assessment with feedback, and also towards the fewer letter-graded courses in the latter semesters of the program.

Finally, we need to accept that part of our job (already) is handling student expectations about how we teach and how they learn. Students expect grades, and many have internalized grades as a reflection of their self-worth. Academics who have let go of grading describe that at first students feel uncomfortable in an academic landscape where their familiar currency has been removed, but that through ongoing discussions of the benefits of going gradeless and by helping students manage their own expectations, many students come to value and even prefer the grade-free approach (e.g., Altahawi et al., 2012).

How can I reduce the assessment workload in my graded course?

What if you wish to, or are required to, retain an A-F grading scale in your course? Here, we offer assessment strategies for course design that aim to give students both meaningful learning activities, feedback, and grading, without pulling all course resources towards grading. Our suggestions range from improving very familiar assessment forms to adopting assessment forms radically different from today's typical practices in Norway. Most of our suggestions build on the use of obligatory pass/fail assignments on the road towards a summative assessment graded on the A-F scale; others don't stipulate obligatory assignments, but we recommend them nonetheless. Note that the principles we presented for pass/fail courses, such as using peer review for feedback, having students close the feedback loop, designing authentic assessments, and using groupwork, are also important in graded courses, though may need to be scaled differently to save time and resources for grading. Furthermore, all of the approaches we suggest can be used without grades in a pass/fail course. In Table 1 we give an overview of our suggested strategies, which are detailed further below. We realize Table 1 is incomplete, but represents a range of options designed to allow for evidence-based assessment without imposing an onerous burden on instructor time.

Table 1. For teachers: transform A-F grading to use less time and fewer resources. All of these strategies can also be used in a pass/fail course.

Strategy	Example	Key elements of course design

Auto-scoring	Multiple choice	Simplest to grade; however, this has many pitfalls, primarily that good MCQs, MTFQs etc. can be difficult to write.
Less material to grade	Compact portfolio & other compact assessments	The graded part of the portfolio becomes significantly slimmer; Multiple obligatory pass/fail tasks, and only a slender but meaningful final summative product for A-F grading.
Fewer items to grade	Team-based learning	This is a highly structured pedagogical approach designed to disarm the negative sides of dysfunctional groups and unleash the power of teamwork.
Here-and-now grading	Oral exams	Oral examination might actually save time. The break-even point can be as high as 80-100 students.
Radically 'ungrading' within A-F scale	Competency-based approaches	More radical restructuring of assessment practices, such as specifications grading.

Auto-scoring assessment options

There are several possible options for auto-scoring student work, of which multiple-choice questions are the most popular. Because MCQs can be scored by computer, the role of the second evaluator here will likely be in ensuring quality, fit, and appropriateness of the assessment tasks or questions. MCQ exams have rightfully earned a bad reputation (Martinez, 1999; Masters et al., 2001; Stanger-Hall, 2012), because they typically test low levels of cognitive complexity in Bigg's or Bloom's taxonomies (see Fig. 1) (Melovitz Vasan et al., 2018; Momsen et al., 2010; Simkin & Kuechler, 2005). However, writing good MCQs that make use of higher cognitive levels is a skill that can be learned (Collins, 2006; Crowe et al., 2008; Kim et al., 2012). At the same time, it is worth noting that writing good MCQs is both time-consuming and difficult, and may only be worthwhile if you can reuse them, which means ensuring your questions won't find their way onto the internet. Furthermore, constrained-choice tests lack authenticity, as the task of filling in bubbles from an array of predetermined options in a high-stress timed situation is typically very unlike the complex skills of our disciplines and students' future jobs (Boud & Falchikov, 2006; Wiggins, 1991).

Multiple true/false questions (MTFQs) are a more sophisticated type of true/false questioning in which students evaluate each possible answer as true or false (Brassil and Couch 2019, Couch et al 2018). Although still subject to the same "inauthentic" critique as MCQs, MTFQs make it more difficult for students to simply "guess" a correct answer, and give instructors a more nuanced understanding of student mastery (and misconceptions).

Auto-scoring is also possible with other formats, including fill-in-the-blank (Medawela et al., 2018), click-and-drag (e.g. with mathematical proofs a la Poulsen et al., 2022), and click-on-target (LaDue and Shipley 2018) question types. In other words, MCQs are not the only option in the auto-scoring toolkit, nor is auto-scoring itself an "all or nothing" proposition;

rather, an instructor can opt for exams that are partly auto-scored, and partly evaluated manually.

Less material to grade: Compact portfolio assessment

A visual artist produces a large body of work over time, with multiple iterations of artistic experimentation. When it comes time to display her work, the artist makes a selection of her best and most interesting pieces, and typically writes an artistic statement explaining her artistic process. Portfolio assessment in education, as originally conceived, allows for similar student autonomy (selecting works to produce and include) and reflection – which we know enhance motivation and learning (Klenowski, 2002). Students package their portfolio to demonstrate that they have met course learning outcomes, and often are expected to write a reflective statement describing their selection process and how their included works demonstrate their learning. The Norwegian adoption of portfolio assessment ("mappevurdering") was spreading in the 1990s, but has in most instances been limited to an instrumental approach simply for permitting multiple student works to be bundled and assessed with a single letter grade. Bundling is not the only reason to use portfolios, and we encourage colleagues to think more holistically about bringing in opportunities for students to select, judge, and reflect in the portfolio process. For example, in many science courses, a final lab report could take the form of a portfolio, with discrete products (cleaned dataset, summary figures and tables, public abstract) accompanied by student reflections.

How can portfolio assessment be meaningful under the two-evaluator requirement? The selection of items for assessment can be significantly more limited, as the final portfolio need not include all products created in the course of the semester. Thus the portfolio can be scaled down significantly to make grading more manageable without compromising learning activities. The items selected can be chosen by the students or by the instructor, or in combination. In this case, we strongly recommend that a short (2-5 page) reflective learning essay or statement be a component in the final portfolio.

Less material to grade: Other compact summative assessments

Imagine scaling the portfolio down even further: the work graded on an A-F grade is *only* the reflective learning essay. Most student work in the semester comprises obligatory pass/fail assignments. Students receive feedback on at least some of the work – on an individual basis (by the instructor, peers or teaching assistants) or on a large-group level by the instructor as appropriate (e.g., the instructor shares overall trends and tendencies they see in the class's work). To 'close the feedback loop', we recommend that no matter which assessment strategy you select, it involves students having the opportunity to revise at least one assignment. For this course design to be pedagogically justifiable, it is important that the pass level for obligatory assignments is meaningful; that is, students need to produce quality work to pass, rather than cursory work or simply effort.

In this scenario, the final summative assessment is similar to the artist statement in the portfolio: students are graded from A-F on a relatively short reflective learning essay where they explore what they have learned in the process of the course, with particular attention to the products they have created via assignments and how they responded to the feedback they received in their revision(s). Jesse Stommel (2020) calls this a "process letter." The reflective

learning essay would be graded by two evaluators. A course structured in this way should be designed such that the writing of a reflective learning essay is an activity that meets the course's intended learning outcomes. For example, a course might include general competency outcomes such as: students are able to evaluate quality in academic work in the discipline; or students can gauge and adjust their own performance in relation to feedback and their own growth as scholars. These outcomes are then met in the final assignment, while other course learning outcomes are met through obligatory activities that would be more time-intensive to grade, *and* that are important for students to receive feedback on.

This approach does not necessarily need to culminate in a reflective learning essay. The instructor can choose a different summative assessment form (whether written or altogether different such as an oral exam or presentation) – *what matters is that the product that is graded A-F is compact enough that it is relatively quick to grade*. Time spent with student work is shifted largely to activities throughout the semester; less time is used by two evaluators evaluating and grading final student work – though it is important that the final product be meaningful and related to intended learning outcomes of the course, and ideally also to the other work students have produced throughout the course.

Fewer items to grade: Team-based learning

Team-based learning (TBL) is a modular, scaffolded implementation of group learning (Michaelsen & Sweet, 2008), one in which students are held accountable—both individually and as group members—for out-of-class preparation. Specifically, students prepare for a class session by reading, viewing recorded lectures or tutorials, etc. and then, in class, they take a learning readiness quiz (or *readiness assurance test*)—individually at first, and then while discussing with group members. The second test often uses the Immediate Feedback Assessment Form (or IF-AT – currently only available in online forms such as at www.intedashboard.com; Cotner et al., 2008a, 2008b), but this is not critical. Quizzing is followed by in-class application exercises, which can take many forms. In "classic TBL," students are graded on their individual and group performance on these quizzes, as a substantial part of the overall course grade. We recommend Sibley and Ostafichuk's (2014) book, which offers practical step-by-step guidance.

Although we expect that the two-evaluator requirement will be removed for multiple choice questions and other automatically graded assignments, we suggest TBL adopters modify their approach so that the obligatory quizzes are subject to pass/fail grading, combined with a graded, summative assessment. The benefits of this approach are that students are motivated to prepare for class work ahead of time, learning is enhanced by group-member contributions, and the instructor acts as a *facilitator of learning*, rather than as the *central figure*, an *expert sharing knowledge*, typically in a lecture format.

Here-and-now grading: Oral exams

It is often easier to discuss and test higher order cognitive skills during an oral exam than for many of the written alternatives. An oral exam can therefore serve as an ideal summative assessment in a course with pass/fail assignments and feedback along the way. Oral exams have always had the requirement of two evaluators and have served as a time-saving option primarily for smaller courses. But now that two evaluators are needed for written exams,

too, oral exams may save time for larger class sizes than before. In our hypothetical (and simplified) example (Fig. 3), the break-even point nearly doubled, from 47 to 81 students.

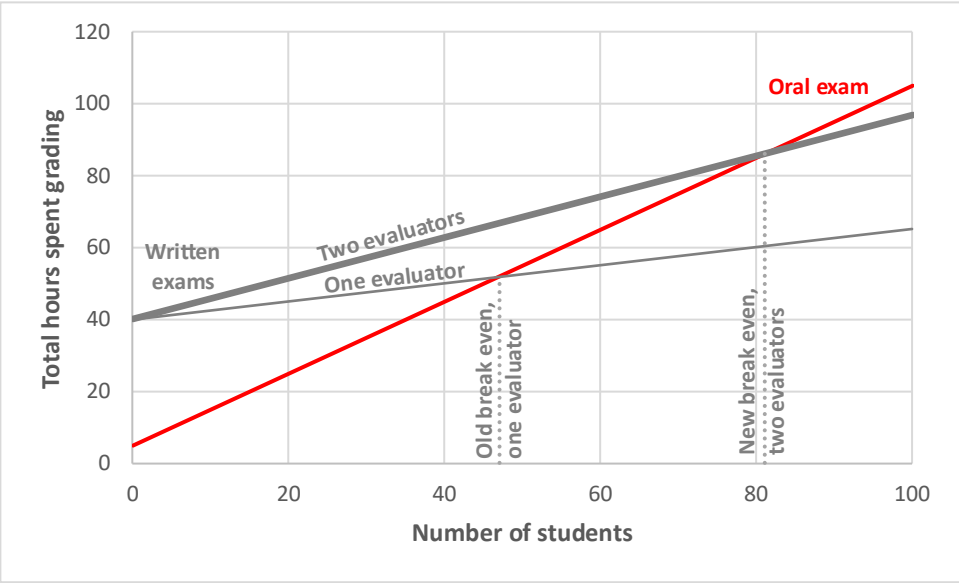

Figure 3. A new break even for oral exams? Oral exams have always had the requirement for two evaluators and effort invested remains unchanged by the proposed new law. But now that written exams also require two evaluators, oral exams should be evaluated as a feasible alternative. Oral exams are an efficient way to test higher cognitive skills in larger courses (81 students in this example) than before (47 students). Assumptions for oral exams (written exams in parentheses): Preparations 5 hrs (40 hrs), grading time per student 25 min (15 min), discussion/comparison among evaluators per student 5 min (2 min). Time for discussion/comparison was set to 0 min for the case of one evaluator for written exams.

Radically 'ungrading' within the A-F scale: Competency-based approaches

There are a number of variations on competency-based (sometimes called mastery-based) grading with different characteristics (Zimmerman & Dibenedetto, 2008). To put it very simply, these approaches involve students completing a set of tasks or assignments to a pass level – with pass set high, at a level where the instructor could argue that the students have achieved a reasonable level of competence over the learning outcomes embedded in each assignment. *The final grade is then contingent upon the level of difficulty of the collection of assignments a student has completed*. These approaches all involve transparency around what students have to do, what characterizes good-enough quality work, and why they need to do it.

In *specifications ('specs') grading*, developed by Linda Nilson (2015), the instructor develops clear specifications for what characterizes good-enough (passing) work for each assignment – similar to a rubric, but with only a single level of description. For each assignment or task, the student either meets expectations (pass) or does not. So far, this is not very different from obligatory assignments in the Norwegian system in a pass/fail course – except that each assignment has specifications communicated clearly to students and used as a guidepost for assessing whether the student work meets expectations.

The instructor creates bundles of assignments, with each bundle corresponding to a letter grade, and students can choose whether they are aiming for an A (complete the A bundle

wherein all assignments in that bundle meet expectations), a B (complete the B bundle), and so forth. The A bundle would have either *more* or *harder* assignments of the same kind (quizzes, problem sets, blog posts, commenting on peers' work, etc.) or have *additional* assignments of a different type than what students completing the B bundle would need to complete successfully. One option for technical implementation in the Norwegian system is to let the A and B bundles be represented by different course codes, so that each is graded pass/fail with one evaluator according to the set expectation. In courses with extensive feedback, this is a transparent way to communicate to students that emphasis is on feedback, not on fine-tuning grades. Flexibility could be incorporated through rules that permit students to change path mid-semester.

The third significant element in specs grading is that students may receive a number of tokens that they can 'cash in' to revise or re-do an assignment that did not meet expectations, or to submit an assignment late – the instructor determines how the token system works, and students determine how they wish to use their tokens strategically to help them succeed in the course. In this type of competency-based approach, specs grading, one could consult an external evaluator to decide on the list of assignments that make up the A bundle, B bundle, and so on. Competency grading does not currently fall under any of the described categories in university regulations (forskrift) that we know of – but university regulations can be updated. We argue that competency-based systems—pedagogically meaningful, fruitful for student learning, and gaining in popularity abroad—are worth putting into practice in pass/fail courses, and it is worth discussing how bundled assignments can be structured in A-F courses. One possible strategy might be that each bundle corresponds to a different course code – the A-bundle of assignments has a different course code than the B-bundle, and each course is assessed pass/fail. This is similar to how a course that students attend together can be coded differently for master students and bachelor students, or for different amounts of course credit.

A call to action

As we write this, the 2021 revision to the law on assessment in higher education is pending, but we are already seeing impacts as universities and university-colleges prepare for the changes. For example, the Norwegian University of Science and Technology (NTNU) initially wrote portfolio assessment completely out of its institutional regulations. Like the other critics of the absolute requirement for two evaluators, we fear that this change in law will lead to a whole cascade of changes that undermine the intentions of the Quality Reform – which has had some mixed results, but unquestionably has led to a more student- and learning-oriented approach to Norwegian higher education.

As of this writing, the status of the new Norwegian law is uncertain. Should the proposed law be implemented, we hope the above suggestions will be helpful to our colleagues as they grapple with how to balance grading demands with other professional duties. Future research will clarify the law's impacts, addressing critical questions such as: do two evaluators lead to more reliable grading (is this truly more "fair")? What aspects of an instructor's time budget are most impacted by the new grading demands—teaching, research, or service? How does this law impact the feasibility of existing grading deadlines (e.g., within three weeks of a course's conclusion)? Should the proposed law *not* be enacted, we still offer the above

suggestions to our colleagues to encourage them to revisit their assessment strategies, and ask themselves whether there is cause for change.

Viewed in the most positive light, this new requirement could act as a call for action on the part of instructors and administrators. This call demands that as well as continuing to argue against detailed regulation of our sector, we raise questions about what we mean by *quality* in education. As part of this call, we could have some meaningful discussions about the overall *purpose* of assessment, and whether our current assessment strategies are about tests and grading, or lifelong learning. Do we use assessment to motivate students to take responsibility for their own learning, and is our assessment authentic—does it fit the practices of our disciplines? Finally, do we need a *societal shift* in how assessment is perceived across the educational spectrum—a shift from an emphasis on test-taking, superficial learning, and grades with limited meaning, to an emphasis on student-centered teaching, practical assessments, and lifelong learning? We urge our colleagues not to allow our sector to undo all the work we have done over 20 years to transform teaching and learning practices in Norwegian higher education, but to meet the challenge with a willingness to make fundamental changes that are in the best interests of our students.

REFERENCES

Aalberg, T., & Lorås, M. (2018, August). Active learning and student peer assessment in a web development course. In Norsk IKT-konferanse for forskning og utdanning. https://ojs.bibsys.no/index.php/NIK/article/view/522

Altahawi, F., Sisk, B., Poloskey, S., Hicks, C., & Dannefer, E. F. (2012). Student perspectives on assessment: Experience in a competency-based portfolio system. Medical Teacher, 34 (3) , 221-225. doi:10.3109/0142159x.2012.652243

Beutel, D., Adie, L., & Lloyd, M. (2016). Assessment moderation in an Australian context: processes, practices, and challenges. Teaching in Higher Education, 22 (1), 1-14. doi:10.1080/13562517.2016.1213232

Biggs, J., & Tang, C. (2010). Teaching for quality learning at university (4th ed.). Maidenhead, UK: Open University Press.

Bloodgood, R. A., Short, J. G., Jackson, J. M., & Martindale, J. R. (2009). A change to pass/fail grading in the first two years at one medical school results in improved psychological well-being. Academic Medicine, 84 (5), 655-662. doi:10.1097/ACM.0b013e31819f6d78

Bloxham, S. (2009). Marking and moderation in the UK: false assumptions and wasted resources. Assessment & Evaluation in Higher Education, 34 (2), 209-220. doi:10.1080/02602930801955978

Bloxham, S., Hughes, C. & Adie, L. (2016). What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. Assessment & Evaluation in Higher Education, 41 (4), 638-653. doi:0.1080/02602938.2015.1039932

Bloxham, S. & Price, M. (2015). External examining: fit for purpose? Studies in Higher Education, 40 (2), 195-211. doi:10.1080/03075079.2013.823931

Bonsaksen, T., Thørrisen, M. M., Sveen, U., Kjeken, I., Aas, R. W., & Lund, A. (2018). Grade correspondence between internal and external examiners of occupational therapy students' Bachelor theses. Uniped, 41 (3), 319-330. doi:10.18261/issn.1893-8981-2018-03-11

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

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. Assessment and Evaluation in Higher Education, 31(4), 399–413. doi:10.1080/02602930600679050

Brassil, C. E., & Couch, B. A. (2019). Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: a Bayesian item response model comparison. International Journal of STEM Education, 6(1), 1-17. doi:10.1186/s40594-019-0169-0

Collins, J. (2006). Education techniques for lifelong learning: Writing multiple-choice questions for continuing medical education activities and self-assessment modules. Radiographics, 26(2), 543–551. doi:10.1148/rg.262055145

Cotner, S., Baepler, P., & Kellerman, A. (2008a). Scratch this! The IF-AT as a yechnique for stimulating group discussion and exposing misconceptions. Journal of College Science Teaching, 37 (4), 48–53. https://www.nsta.org/journals/journal-college-science-teaching/journal-college-science-teaching-marchapril-2008/scratch

Cotner, S., & Ballen, C. J. (2017). Can mixed assessment methods make biology classes more equitable? PLoS ONE, 12 (12), e0189610. doi:10.1371/journal.pone.0189610

Cotner, S. H., Fall, B. A., Wick, S. M., Walker, J. D., & Baepler, P. M. (2008b). Rapid feedback assessment methods: Can we improve engagement and preparation for exams in large-enrollment courses? Journal of Science Education and Technology, 17 (5), 437–443. doi:10.1007/s10956-008-9112-8

Couch, B. A., Hubbard, J. K., & Brassil, C. E. (2018). Multiple–true–false questions reveal the limits of the multiple–choice format for detecting students with incomplete understandings. BioScience, 68(6), 455-463. doi:10.1093/biosci/biy037

Crowe, A. J., Dirks, C., & Wenderoth, M. P. (2008). Biology in Bloom: Implementing Bloom's taxonomy to enhance student learning in biology. CBE-Life Sciences Education, 7 (4), 368 –381. doi:10.1187/cbe.08-05-0024

Dahl, B., Lien, E., & Lindberg-Sand, Å. (2009). Conformity or confusion? Changing higher education grading- cales as a part of the Bologna Process. Learning and Teaching, 2 (1), 39-79. doi:10.3167/latiss.2009.020103

Dahl, T. I. (2006). When precedence sets a bad example for reform: Conceptions and reliability of a questionable high stakes assessment practice in Norwegian universities. Assessment in Education: Principles, Policy and Practice, 13 (1), 5–27. doi:10.1080/09695940600563579

Dannefer, E. F. (2013). Beyond assessment of learning toward assessment for learning: Educating tomorrow's physicians. Medical Teacher, 35 (7), 560-563. doi:10.3109/0142159x.2013.787141

Frich, J., Lundin, K. E. A., Os, I. (2014). Karaktersystemet – avveining mellom ulike hensyn. Tidsskrift for Den Norske Legeforening, 134 (1), 14-15. doi:10.4045/tidsskr.13.1338

Guinier, L. (2015). The tyranny of the meritocracy: Democratizing higher education in America. Boston, MA: Beacon Press.

Hansen, M. N., & Strømme, T. B. (2021). Historical change in an elite profession—Class origins and grades among law graduates over 200 years. *British Journal of Sociology*, *72*(3), 651–671. doi:10.1111/1468-4446.12852

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. Review of Educational Research, 58 (1), 47-77. doi:10.3102/00346543058001047

Jørgensen, C., & Bråten, H. (2019). Can ‘ Passed with Distinction ’ as a New Grading Scale Favour the Transition towards Formative Assessment ? Nordic Journal of STEM Education, 3 (1), 7–11. doi:10.5324/njsteme.v3i1.2992

Kim, M. K., Patel, R. A., Uchizono, J. A., & Beck, L. (2012). Incorporation of Bloom’s taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. American Journal of Pharmaceutical Education, 76 (6), 114. doi:10.5688/ajpe766114

Klenowski, V. (2002). Developing portfolios for learning and assessment: Processes and principles. London, UK: RoutledgeFalmer.

Kohn, A. (1994). Grading – the issue is not how but why. Educational Leadership, 52 (2), 38-41. https://www.ascd.org/el/articles/grading-the-issue-is-not-how-but-why

Kohn, A. (2006). The trouble with rubrics. English Journal, 95 (4), 12-15. doi:10.2307/30047080

LaDue, N. D., & Shipley, T. F. (2018). Click-on-diagram questions: A new tool to study conceptions using classroom response systems. Journal of Science Education and Technology, 27(6), 492-507. doi:10.1007/s10956-018-9738-0

Leenknecht, M., Wijnia, L., Köhlen, M., Fryer, L., Rikers, R., & Loyens, S. (2021). Formative assessment as practice: the role of students’ motivation. Assessment and Evaluation in Higher Education, 46 (2), 236–255. doi:10.1080/02602938.2020.1765228

Lockspeiser T.M., O’Sullivan P., Teherani A., & Muller J. (2008). Understanding the experience of being taught by peers: the value of social and cognitive congruence. Adv Health Sci Educ Theory Pract, 13(3), 361-72. doi:10.1007/s10459-006-9049-8.

Martinez, M. E. (1999). Cognition and the question of test item format. Educational Psychologist, 34 (4), 207–218. doi:10.1207/s15326985ep3404_2

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. Journal of Nursing Education, 40 (1), 25-32. doi:10.1016/j.nedt.2006.07.006

McGuire, S. Y. & McGuire, S. (2015). Teach students how to learn: Strategies you can incorporate into any course to improve student metacognition, study skills, and motivation. Sterling, VA: Stylus Publishing.

Medawela, R. S. H. B., Ratnayake, D. R. D. L., Abeyasinghe, W. A. M. U. L., Jayasinghe, R. D., & Marambe, K. N. (2018). Effectiveness of “fill in the blanks” over multiple choice questions in assessing final year dental undergraduates. Educación Médica, 19(2), 72-76. doi:10.1016/j.edumed.2017.03.010

Melovitz Vasan, C. A., DeFouw, D. O., Holland, B. K., & Vasan, N. S. (2018). Analysis of testing with multiple choice versus open-ended questions: Outcome-based observations

in an anatomy course. Anatomical Sciences Education, 11 (3), 254–261. doi:10.1002/ase.1739

Michaelsen, L. K., & Sweet, M. (2008). The essential elements of Team-Based Learning. New Directions for Teaching and Learning, 2008 (116), 7–27. doi:10.1002/tl.330

Mijs, J.J.B. (2016). The unfulfillable promise of meritocracy: Three lessons and their implications for justice in education. Social Justice Research, 29, 14-34. doi:10.1007/s11211-014-0228-0

Minarechová, M. (2012). Negative impacts of high-stakes testing. Journal of Pedagogy, 3 (1), 82–100. doi:10.2478/v10159-012-0004-x

Momsen, J. L., Long, T. M., Wyse, S. A., & Ebert-May, D. (2010). Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. CBE Life Sciences Education, 9 (4), 435–440. doi:10.1187/cbe.10-01-0001

Liu, N. & Carless, D. (2006). Peer feedback: the learning element of peer assessment. Teaching in Higher Education, 11:3, 279-290. doi:10.1080/13562510600680582

Nichols, S. L., & Berliner, D. C. (2007). The pressure to cheat in a high-stakes testing environment. In E. M. Anderman & T. B. Murdock (Eds.), Psychology of academic cheating (pp. 289–311). Elsevier Academic Press. doi:10.1016/B978-012372541-7/50016-4

Nilson, L. (2015). Specifications grading: Restoring rigor, motivating students, and saving faculty time. Sterling, VA: Stylus Publishing.

OECD. (2018). Higher education in Norway: Labour market relevance and outcomes. doi:10.1787/9789264301757-en

Poulsen, S., Viswanathan, M., Herman, G. L., & West, M. (2022). Evaluating proof blocks problems as exam questions. ACM Inroads, 13(1), 41-51. doi:10.1145/3514213

Reed, D. A., Shanafelt, T. D., Satele, D. W., Power, D. V., Eacker, A., Harper, W., Moutier, C., Durning, S., Massie, F. S., Thomas, M. R., Sloan, J. A., & Dyrbye, L. N. (2011). Relationship of pass/fail grading and curriculum structure with well-being among preclinical medical students: A multi-institutional study. Academic Medicine, 86 (11), 1367–1373. doi:10.1097/ACM.0b013e3182305d81

Reinholz, D. (2016). The assessment cycle: a model for learning through peer assessment. Assessment & Evaluation in Higher Education, 41:2, 301-315. doi:10.1080/02602938.2015.1008982

Regjeringen. (2001). Gjør din plikt – Krev din rett. St. meld. nr. 27 (2000-2001). Kirke-, utdannings- og forskningsdepartementet. https://www.regjeringen.no/no/dokumenter/stmeld-nr-27-2000-2001-/

Regjeringen. (2021). Prop. 111L (2020-2021). Kunnskapsdepartementet. https://www.regjeringen.no/no/dokumenter/prop.-111-l-20202021/

Rohe, D. E., Barrier, P. A., Clark, M. M., Cook, D. A., Vickers, K. S., & Decker, P. A. (2006). The benefits of pass-fail grading on stress, mood, and group cohesion in medical students. Mayo Clinic Proceedings, 81 (11), 1443–1448. doi:10.4065/81.11.1443

Rye, J. F. (2014). Konsistente karakterer? Uniped, 37 (3), 63-77. doi:10.3402/uniped.v37.22654

Sadler, P.M. & Good E. (2006). The impact of self- and peer-grading on student learning. Educational Assessment, 11:1, 1-31. doi:10.1207/s15326977ea1101_1

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018a). Classroom assessment principles

to support learning and avoid the harms of testing. Educational Measurement: Issues and Practice, 37 (1), 52–57. doi:10.1111/emip.12195

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018b). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. Educational Measurement: Issues and Practice, 37 (1), 21–34. doi:10.1111/emip.12189

Sibley, J., & Ostafichuk, P. (2014). Getting started with Team-Based Learning. Sterling, VA: Stylus Publishing.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: what is the connection? Decision Sciences Journal of Innovative Education, 3 (1), 73–98. doi:10.1111/j.1540-4609.2005.00053.x

Smith, C. (2012). Why should we bother with assessment moderation? Nurse Education Today, 32 (6), e45-e48. doi:10.1016/j.nedt.2011.10.010

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. CBE Life Sciences Education, 11 (3), 294–306. doi:10.1187/cbe.11-11-0100

Stecher, B. (1998). The local benefits and burdens of large-scale portfolio assessment. International Journal of Phytoremediation, 21 (1), 335–351. doi:10.1080/0969595980050303

Stommel, J. (2020). How to ungrade. In Susan D. Blum (Ed.) Ungrading: Why rating students undermines learning (and what to do instead) (pp. 25-41). Morgantown, WV: West Virginia University Press.

Støren, L. A., Reiling, R. B., Skjelbred, S. E., Ulvestad, M. E., Carlsten, T. C., & Olsen, D. S. (2019). Utdanning for arbeidslivet. NIFU report 2019:3. doi:11250/2589732

TNS Gallup. (2015). NTNU Arbeidsgiverundersøkelsen 2015. https://innsida.ntnu.no/documents/portlet_file_entry/10157/NTNUs+arbeidsgiverunders%C3%B8kelse+2015.pdf/c0e492a2-1b3b-43a8-b13d-42c672ed9d6a

Von der Embse, N., Barterian, J., & Segool, N. (2013). Test anxiety interventions for children and adolescents: A systematic review of treatment studies from 2000-2010. Psychology in the Schools, 50 (1), 57–71. doi:10.1002/pits.21660

von der Embse, N., Jester, D., Roy, D., & Post, J. (2017). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. Journal of Affective Disorders, 227, 483–493. doi:10.1016/j.jad.2017.11.048

White, C. B., & Fantone, J. C. (2010). Pass-fail grading: Laying the foundation for self-regulated learning. Advances in Health Sciences Education, 15 (4), 469–477. doi:10.1007/s10459-009-9211-1

Wiggins, G. (1990). The case for authentic assessment. Practical Assessment, Research and Evaluation, 2 (2), 1990–1991. doi:10.7275/ffb1-mm19

Zimmerman, B. J., & Dibenedetto, M. K. (2008). Mastery learning and assessment: implications for students and teachers in an era of high-stakes testing. Psychology in the Schools, 45 (3), 206–216. doi:10.1002/pits.20291