OPEN ACCESS

INQUIRY ARTICLE

Performance and Learning in a Two-Stage Exam - Differential Performance Gains and Collaborative Dynamics in a Norwegian Geography Course

Gidske L. Andersen[1]*

1 Department of Geography, University of Bergen
* Gidske.Andersen@uib.no

Received: 2026-01-30; Accepted: 2026-02-18; Published 2026-05-19
Editor: Robert Kordts

Abstract

Two-stage exams are assessment forms that allow students to take a test individually, then to take it collaboratively in groups. The second stage connects testing and learning through social interaction, particularly peer dialogue, thereby enabling immediate, internal feedback to support self-regulation. Most published research has evaluated implementations in US and Canadian colleges, and no studies from Norway have been identified. This study evaluates the implementation of a two-stage exam in a geography course. Results are based on a comparison of grades between the two stages, complemented with responses from a questionnaire. Findings indicate that collaboration improves student performance, even to the extent that group results are better than the highest individual result. However, low-performing students tend to benefit the most. Groups collaborated effectively, as indicated by consensus-finding, with the most pronounced effect in the cohort with greater prior collaborative experience. Questionnaire responses suggest that this prior experience fostered a safe and trusting environment for exchanging ideas during peer dialogue. Altogether, the findings suggest that collaboration was effective and that positive interdependence existed, promoting learning. For instructors considering two-stage exams, the findings presented here suggest that this format offers pedagogical advantages and may be a timely alternative at a moment when AI challenges several other assessment forms that aim to align active learning and testing.

NORSOTL: NORSK TIDSSKRIFT FOR SCHOLARSHIP OF TEACHING AND LEARNING /
NORWEGIAN JOURNAL FOR THE SCHOLARSHIP OF TEACHING AND LEARNING
VOL 1 | UiB Læringskonferanse | 2026
Publisher: Research group "Teaching and Learning in Higher Education" at the University of Bergen & Research group "University Pedagogy" at the Western Norway University of Applied Sciences (HVL)

DOI: https://doi.org/10.15845/norsotl.v

Keywords: two-stage exam, collaborative assessment, group dynamics, motivation, geography

Introduction

Higher education has faced major challenges in recent years. While the pandemic came and went, Artificial Intelligence (AI) has arrived to stay. And while we were quickly forced to adapt to the pandemic, adapting to generative AI and Large Language Models appears to be both more difficult and slower. A special committee has been established to investigate how Norwegian higher education should adapt to AI (Malthe-Sørenssen-utvalget). How we assess students is particularly challenging, and the committee has already indicated that home exams should be avoided. AI, therefore, challenges all educators to think anew about assessment methods and to view new and old assessment formats with fresh eyes. Traditional school exams are not the way forward, even though a controlled environment undoubtedly offers an advantage when AI use should be excluded from assessments. The so-called two-stage exam[1] is an assessment method that could become more significant because it occurs in a controlled setting and includes elements of active learning.

The two-stage exam allows students to do the test individually and then collaboratively in groups. It is a widely popular format in STEM courses (Callaghan et al., 2025), but most published research has evaluated implementations in US and Canadian colleges, and no studies from Norway have been identified. It has been argued that this assessment form is a "low-hanging fruit" for instructors who want to introduce active learning without restructuring the whole course (Bruno et al., 2017). The study presented here evaluates the implementation of this assessment form in a first-semester geography course at a Norwegian university. It also reflects on whether it can be a sustainable addition to standard summative assessment forms in the age of AI, as well as on its contribution to long-term educational impact.

Learning through dialogue and social interaction

Dialogue is the common denominator for active learning and two-stage exams. Through dialogue, students receive immediate feedback, a process that promotes learning. In this process, the student compares their own knowledge with information from peers, and this internal feedback is a key mechanism in learning (Nicol, 2007; Nicol & Selvaretnam, 2022). Dialogue is a main component of active learning, in stark contrast to traditional teaching methods, where the student is a passive recipient of information, often through a monologue from the instructor (Biggs, 1999). Research has shown that teaching methods that promote active learning are twice as effective for learning as traditional teaching (Deslauriers et al., 2011). In an examination situation, however, dialogue is an unconventional tool, but this is precisely the strength of the two-stage exam. Instead of merely assessing the extent to which knowledge has been acquired, this form of assessment also facilitates learning during the assessment process itself.

During the second, collaborative stage of the exam, testing and learning are connected through social interaction, a key concept in two theories that have been used to explain the learning effect of two-stage exams, i.e. social constructivism and social

[1] In the literature, it is also referred to as a two-phase exam, a two-stage collaborative assessment, collaborative testing, collaborative group testing or a pyramid exam. Terms used interchangeably with these phrases include exam, examination, testing, and assessment.

interdependence (Efu, 2019; Mahoney & Harris-Reeves, 2019; Sibley & Ostafichuk, 2023; Zipp, 2007). Social constructivism explains how knowledge is constructed through social interaction. It builds on Vygotsky's concept of scaffolding learning (as cited in Sibley & Ostafichuk, 2023) in which a More Knowledgeable Other (MKO; teacher, peer, etc.) supports learners in attaining higher levels of knowledge and skills, with this support being gradually withdrawn as competence increases. Social interdependence, on the other hand, explains how individuals' outcomes are shaped by their own and others' actions (Johnson & Johnson, 2009). When a common goal exists, the theory predicts that collaboration improves, resulting in higher levels of achievement. This is an example of positive interdependence. Together, these theories help explain why social interaction in the second stage of the assessment can promote learning.

Pros and cons of two-stage exams

Empirical studies show that two-stage exams enhance both performance and learning (Bloom, 2009; Gilley & Clarkston, 2014; Jang et al., 2017). Additional benefits include improved collaboration skills (Levy et al., 2023), reduced test anxiety (as cited in Efu, 2019) and increased motivation (as cited in Jang et al., 2017). Moreover, students are generally positive about this format, and mention additional advantages like grade boost, making friends, and increased confidence (Rieger & Rieger, 2020).

Negative feedback on two-stage exams is relatively uncommon. However, one concern among students relates to who benefits from the collaborative stage, and it has therefore been pointed out that students need to be convinced that all, or at least most, students benefit from this assessment form for them to embrace it (Rieger & Rieger, 2020). Many studies have shown that results improve after the collaborative stage, to the point that the lowest group score can exceed the highest individual score (as summarised by Lee et al., 2022; Martin, 2018; Wieman et al., 2014). Still, studies are not unified in their conclusions about who benefits most. Some studies indicate that the benefit may be equally distributed across performance levels (Gilley & Clarkston, 2014; Mahoney & Harris-Reeves, 2019), while other studies show that low-performing students benefit the most (Bruno et al., 2017; Giuliodori et al., 2008). When low performers benefit most, it may be related to the challenge of free-riding. This is a known challenge in teamwork, in which some students benefit from others' efforts while contributing little themselves. Teams that struggle to collaborate in general are also less likely to reach consensus through constructive discussions, which is crucial to the success of two-stage exams (Rieger & Rieger, 2020; Sibley & Ostafichuk, 2023, p. 26). Further studies on group dynamics and composition to strengthen students' trust in the collaborative stage have therefore been recommended (Rieger & Rieger, 2020).

Objectives

This study evaluates the implementation of a two-stage exam in a first-semester geography course at the University of Bergen (UiB), Norway, the only course in the Faculty of Social Sciences that uses this assessment form. Based on the literature, students are expected to perform better in the second stage. Given that this can be confirmed, the study will first answer the following research questions:

- To what extent does group performance exceed individual performance?
- Compared to the performance level in the first stage, which students benefit the most in the collaborative stage?
- To what extent do groups reach consensus?

These quantitative findings will be complemented by a survey of students' self-

assessments of their performance, motivation, perceived learning, and experienced group dynamics during the collaborative stage. This analysis will be exploratory, with an underlying hypothesis that optimal collaboration in groups, characterised by safe and trustworthy peer dialogue, would lead to improved learning and result in more correct answers, as self-reported by an improvement in grade after the collaborative stage.

Materials and Methods

This study uses an uncontrolled, exploratory design to evaluate the implementation of a two-stage exam. The learning effect of two-stage exams is commonly evaluated by examining improved performance in the collaborative stage or by assessing retention with an additional individual test after that stage. However, retention tests require controlled experiments, which are often difficult to implement in small cohorts. The learning effects of test repetition may also confound results (Cooke et al., 2019).

Students' performance in the first stage depends on individual abilities and preparations. Results in the second stage depend on the effect of group learning (Zipp, 2007), and the improved performance after this stage is therefore used as an indicator of learning. However, less ideally, improved performance might depend on guessing or copy-pasting answers from other group members. It is not straightforward to determine which mechanism is at play in an uncontrolled design. Therefore, analysis of test results will be complemented with survey results to provide further insights into students' experiences during the collaborative stage.

The study uses test results from the 2023 and 2024 cohorts, referred to as H23 (n=60) and H24 (n=68), respectively, to answer the first set of research questions. It was important that respondents still had the test relatively fresh in mind. Therefore, the questionnaire was shared only with the H24 cohort. This was done in April 2025, while their test took place in October 2024. 43 of the H24 students responded.

Course context

The assessment examined is part of the first-semester 10-credit course GEO110, "Cartography and Thematic maps," at UiB. Its overall learning goal is to develop the competencies required to be a critical user of maps and geographical data. The course is organised into modules comprising online materials, lectures, seminars, and assignments. The two-stage exam assesses content from one module, supporting learning goals related to understanding cartographic concepts, remote sensing principles, global navigation satellite systems, and key characteristics of geographical data.

Students are encouraged to read recommended literature and watch video lectures. Quiz results help the instructor decide which topics to review during teaching. Teaching also includes collaborative activities like *think-pair-share*.

Before the first seminar, students were assigned to groups of three to five. These were also the groups for the exam, and students were therefore encouraged to work together in them during seminars. All seminars led to a hand-in. For H23, these were voluntary and individual. For H24, three mandatory group assignments were introduced prior to the exam to facilitate collaboration.

The two-stage exam

The two-stage exam consists of two identical online tests, the most common format. It is defined as computer-assisted assessments in which deployment and marking are automated (Boitshwarelo et al., 2017). Online tests have documented limitations, such as emphasising recall rather than higher-order thinking, and the risk that poorly designed

distractors can generate false knowledge (Boitshwarelo et al., 2017; Butler, 2018; Nicol, 2007). However, incorporating peer dialogue in the second stage can mitigate these drawbacks, as collaborative discussion fosters self-regulation and supports knowledge building through peer interaction and feedback (Nicol, 2007).

The two-stage exam took place mid-semester, around two weeks after the module's last lecture. Each test consisted of 30 questions and lasted 1 hour. Questions were like those on previous quizzes and included conventional multiple-choice questions, text/numeric entry, matching/pairing, drag-and-drop, hotspot, and graphic gap match (Inspera question types). Questions aimed at encouraging higher-order thinking rather than testing only knowledge recall. Because students share experiences across cohorts, question types were similar, but answer options differed between the H23 and H24 tests.

The exam has two parts: the individual *itest* and the collaborative *gtest*. Students discussed their answers in the pre-assigned groups in the collaborative part, but delivered their *gtest* results individually. Because of this individual responsibility in the second stage, each part counted equally. While equal weighting can increase the risk of grade inflation (Rieger & Heiner, 2014), it was considered low here because the results accounted for only 15% of the total grade.

Although the exam was summative, its unconventional format prevented it from being administered in a quiet school-exam setting. It was therefore organised by the instructor (the author) and two to three teaching assistants. This allowed observing group interactions during the *gtest*. Students were seated in assigned groups for the whole exam. Flexible rooms were organised to allow sufficient spacing, preventing disturbance and unwanted interaction during testing. After the *gtest*, the instructor presented the correct answers.

Questionnaire

The questionnaire (

Table 1) focused on aspects of learning, motivation and collaboration. The statements were formulated to reflect students' attitudes toward collaborative activities, whether such activities were motivating, and students' own perceptions of learning during the *gtest*. The statements were designed as part of a pedagogy course attended by the author and were built on previous experiences teaching GEO110. The response levels were Likert-type, using frequency, agreement, importance, or quality scales with five levels. Self-reported improvement on the *gtest* had three levels. The questionnaire also included one open field for reflections. The questionnaire consisted of only 14 statements to ensure respondents would complete it. It collected no personal information, and a self-evaluation of its ethical aspects was registered in Rette. Respondents were informed about how to withdraw their response.

Table 1. The questionnaire with *codes* used in the analyses, Likert-item *statements*, a keyword indicating its topical *focus*, the *level* scale used, and whether the scale needs to be *reversed* for interpretation. All scales had five levels, with the mid-level being neutral. Agreement scale: 1=strongly disagree, 5= strongly agree; Frequency scale: 1=Never, 5=Always; Quality scale: 1=Very poor, 5=Excellent; Importance scale: 1=Unimportant, 5=Very important

code	statement	focus	level	reverse
Q01	What role did mandatory group submissions early in the semester play in facilitating effective group collaboration?	cooperation	importance	0
Q02	In advance two stage exam, the collaboration in my group can best be described as:	cooperation	quality	0
Q03	During the *gtest*, we collectively arrived at the correct answer through discussion	learning	frequency	0
Q04	During the *gtest*, one person knew the answer, and the rest copied it without discussion	learning	frequency	1
Q05	When we discussed answer alternatives during the *gtest*, I felt that I learned something	learning	frequency	0
Q06	I felt safe and comfortable sharing answers, ideas and opinions with the rest of the group	learning	frequency	0
Q07	I have a positive attitude towards group work	motivation	frequency	0
Q08	During the *gtest*, everyone in my group chose the same answer	learning	frequency	0
Q09	I think the *gtest* positively influenced my overall result on the test	result	agreement	0
Q10	I believe others in my group benefited more from the *gtest* than I did.	result	agreement	1
Q11	The *gtest* was a waste of time	motivation	agreement	1
Q12	My group had one or more free-riders	cooperation	agreement	1
Q13	I missed guidelines for making group collaboration work effectively	cooperation	agreement	1
Q14	Knowledge of the *gtest* motivated us to make group collaboration effective from the start of the semester.	motivation	agreement	0
Response	Self-reported improvement on the *gtest* with 3 levels		lower, same or better result than the *itest*	

Data analysis

To answer the quantitative research questions, grades from the two tests were compared. Letter grades from official databases were converted to integers (A=5, F=0) and combined with group affiliation from the learning platform. Identifying information was deleted from the final datasets. The questionnaire was shared using Microsoft Forms.

Analyses were done in R (version 4.5.1; R Core Team, 2025) using the following libraries: the tidyverse package of libraries (Wickham et al., 2019), likert (Bryer et al., 2016), psych (Revelle, 2025), ggcorrplot (Kassambara, 2023), MASS (Venables & Ripley, 2002) and mgcv (Wood, 2011).

Improvement across stages and performance levels

To establish that grades were higher in the *gtest* than the *itest*, improved performance was assessed by comparing grade distribution. The null hypothesis is that there is no improvement in the second stage. The alternative hypothesis is that performance increased after the *gtest*. Paired observations within cohorts were tested using a paired Wilcoxon test, and differences among cohorts were tested using the Mann-Whitney test.

Previous studies have found that all group members benefit from successful group interaction, i.e. that the group result is better than the best individual result. To test this, three different measures of improvement (*gtest_result – itest_result*) were calculated for each group, using the lowest, the mean, and the highest grade on the *gtest,* respectively, compared to the highest grade of the *itest.* The null hypothesis is that there is no difference. This was tested by a Wilcoxon test, with the alternative hypothesis that there is a positive difference, i.e. the group effort is better than the best individual effort.

The pattern of improvement across performance levels was examined using a cross-tabulation of *itest* and *gtest* results. To further examine who benefits most, regression analysis was used with increased performance as the dependent variable and the *itest* result as the independent variable. The null hypothesis is that all students benefit equally. Both linear and non-linear regressions were computed to examine the type of relationship. Total improvement (TI) as well as relative improvement (RI) were modelled, defined respectively as:

TI = *gtest_result- itest_result*

RI = *gtest_result – itest_result)/(5-itest_result)*100, where 5 is the best integer grade (A)*

Relative improvement accounts for higher-performing students having less "room" for improvement, and it was calculated after Gilley & Clarkston (2014), but used integer grades.

Group consensus

If consensus decisions are made in the second stage, it will result in more similar results. With letter grades, the best indication of consensus-finding is fewer distinct grades among group members. The null hypothesis used to test this is that there is no reduction in distinct grades among group members after the *gtest*. This was tested using a Wilcoxon test across cohorts and stages, with the alternative hypothesis that there are fewer distinct grades after the *gtest*.

As groups in H24 had more organised group activities in advance of the exam than those in H23, there may have been an effect of the mandatory group work on their ability to reach consensus. The extent of consensus across cohorts was tested using a proportions test, with the alternative hypothesis that there are more consensus groups in H24.

The questionnaire

Responses from the questionnaire were visualised using Likert-item plots and analysed using a combination of qualitative and quantitative approaches to explore patterns in and correlations among responses. Correlations among items were tested using Kendall's Tau correlation, accounting for the ordinal nature of the Likert items (Harpe, 2015).

Results

Low performers benefit the most

Analysis of grade distribution confirmed that *gtest* results improved significantly compared to the *itest* (Figure 1 and Table 2). This was the case for both cohorts combined and for each cohort separately (p-value < 0,001, Table 2). The H24 cohort performed significantly better than H23 on both tests (Mann-Whitney *itest*: p-value < 0.001; *gtest*: p-value = 0.018), but B remained as the median grade for both cohorts after the *gtest*.

Across both cohorts, group performance significantly affected grades. The lowest group grade after the *gtest* was better than the highest on the *itest* (Table 2). However, this effect was driven by H23 results. For H24, only the best grade of the *gtest* was significantly better than the best grade of the *itest* (p-value = 0.012).

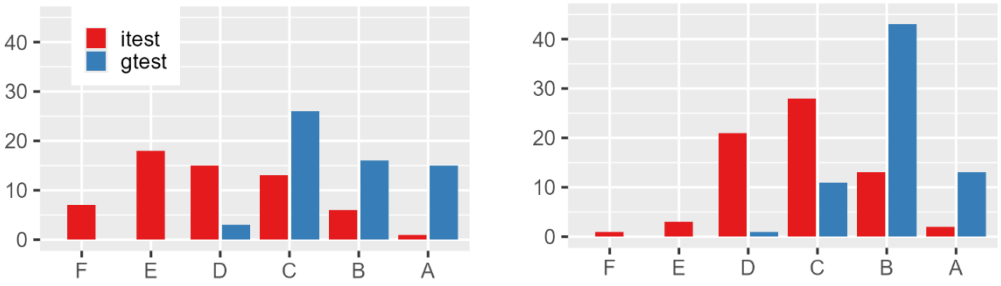

Figure 1. Grade distributions for H23 (left) and H24 (right).

Table 2. Results from the paired Wilcoxon tests, with description of the test, the alternative hypotheses, and p-value for tests including both cohorts (both) and each separately (H24 and H23).

Test description	alternative hypothesis	both	H24	H23
improvement: difference in grades (*itest* - *gtest*)	grades *gtest* > *itest*	<0.001	<0.001	<0.001
Consensus: distinct grades within groups	*gtest* distinct < *itest* distinct	<0.001	<0.001	0.004
Group performance 1	*gtest* min > *itest* max	0.026	0.232	0.021
Group performance 2	*gtest* mean > *itest* max	<0.001	0.055	<0.001
Group performance 3	*gtest* max > *itest* max	<0.001	0.012	<0.001

The cross-tabulation showed that most students (>50%) in both cohorts benefited, except for B students in H24, where only 30% improved their results (Figure 2). The regression analysis with total improvement as the dependent variable generally indicated that low-performing students benefited most (H24: slope = -0.77, adjusted R2 = 0.58, p-value < 0.001; H23: slope = -0.70, adjusted R2 = 0.52, p-value < 0.001). However, when inspecting relative improvement, RI, there were variations among cohorts (Figure 3). For H24, low-performing students still benefited most, but for H23, a U-shaped curve, rather than a

linear one, provided the best fit (ANOVA, *p*-value = 0.006), indicating that both low- and high-performing students gained more than mid-performing students.

Consensus increases

The number of distinct grades at the group level was significantly fewer for the *gtest* (2) than the *itest* (4) (Figure 4 and Table 2). This trend was strongest for H24, and the proportion of groups with the same grade for all group members was significantly higher in H24 than in H23 (75% versus 39%; p=0.01).

H24	F	E	D	C	B	A
F	0	0	0	**1**	0	0
E	0	0	0	0	**3**	0
D	0	0	1	6	**12**	2
C	0	0	0	3	**19**	6
B	0	0	0	1	**8**	4
A	0	0	0	0	1	1

H23	F	E	D	C	B	A
F	0	0	1	**3**	0	3
E	0	0	1	**12**	4	1
D	0	0	1	**8**	3	3
C	0	0	0	3	**8**	2
B	0	0	0	0	1	**5**
A	0	0	0	0	0	1

Figure 2: Cross-tabulation of grades from the *gtest* (columns/blue) and *itest* (rows/red) for H23 (left) and H24 (right). Diagonal cells are the number of candidates with similar results in both tests, while cells below the diagonal are the number of candidates with lower results after the *gtest*. Row-wise, cells in bold indicate the median grade after the *gtest* for the corresponding *itest* grade.

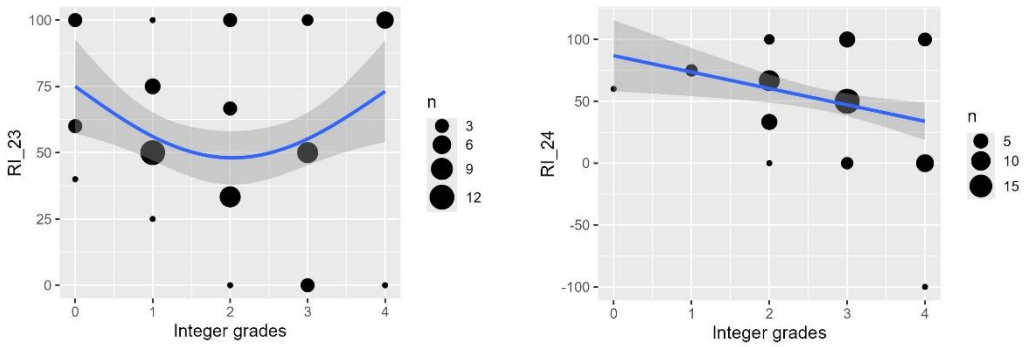

Figure 3: Relative improvement (RI) in grades for H23 (left) and H24 (right) given grades on the *itest* (x-axis). A non-linear model gave the best fit for H23 with the lowest modelled improvement for D-grade students. For H24, the relationship was linear and negative. Note the different scale on the y-axis.

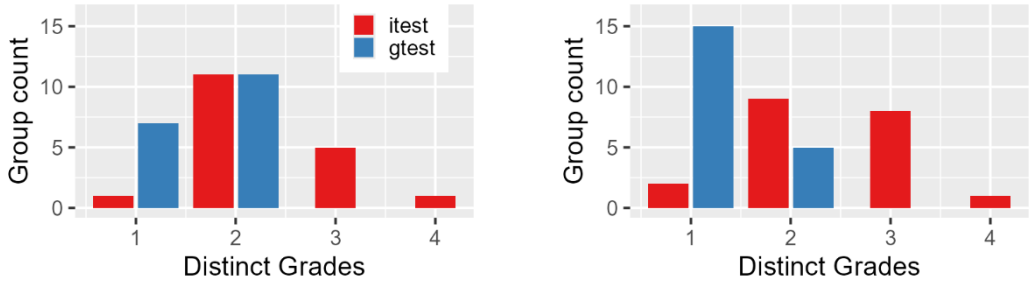

Figure 4. Distribution of distinct grades for H23 (left) and H24 (right) for the *itest* and *gtest*.

Relationship between motivation, learning, group work quality and gtest results

The responses generally indicated positive agreement with the statements (Figure 5). Only two statements had neutral as the dominant level (Q10 and Q4). Self-reported improvement following the *gtest* was consistent with the quantitative findings (Figure 6).

Responses regarding motivation (Q7, Q11, Q14) indicate that most students enjoyed group work (Q7: 70%) and found the *gtest* useful (Q11: 74%). The latter were more likely to report that their groups had no free-riders (Q11 vs Q12), were more confident in how to make collaborations work (Q11 vs Q13) and tended to perform better on the *gtest* than on the *itest* (Q11 vs Response).

While early knowledge about the *gtest* motivated around half the students to make collaboration effective from the beginning of the semester (Q14), 77% said that mandatory assignments were *very important* or *important* for the same purpose (Q1). These responses are positively correlated (Figure 5).

For the group work itself (Q1-2, Q12-13), 2/3 of respondents report that it functioned well or excellently before the test (Q2). A small group (16%) would have liked more guidelines on how to make group work effective (Q13), and about the same percentage reported that there were free-rider(s) in their group (Q12).

Responses concerning learning (Q3:6, Q8) revealed that most respondents (67%) *always* or *very often* had a positive learning experience during the *gtest* (Q5). Furthermore, 91% *always* or *often* felt confident and safe sharing thoughts during discussions, while only 9% *rarely* or *sometimes* had the same experience (Q6). Remarkably, there was no negative feedback on statements Q8 and Q3, and, respectively, 81% and 79% of respondents often or always experienced their group agreeing to and collectively arriving at the correct answer option. It was *never* or *rarely* the case that multiple-choice options were copied without discussion in the group (9%; Q4). Nevertheless, 49% reported that this was sometimes the case.

Significant correlations among *learning* statements (Figure 5) underscore that discussions in safe settings were important for consensus-finding and *feeling of learning* (Q3 vs Q5, Q6, Q8). Those who affirmed a feeling of learning (Q5) also had improved test results (Q9). Moreover, those who collectively chose the same answers (Q8) were more likely to find the *gtest* useful (Q11), to gain from it (Response), and to have sufficient guidelines for the group work (Q13).

Only 12% of the respondents found the *gtest* negative for their overall result (Q9). Furthermore, these respondents *strongly agreed/agreed* that other group members had a larger advantage, as suggested by a strong negative correlation (Q9 vs Q10). Hence, they more often found the *gtest* to be a waste of time (Q9 vs Q11) and did not improve their results after the *gtest* (Q10 vs Response).

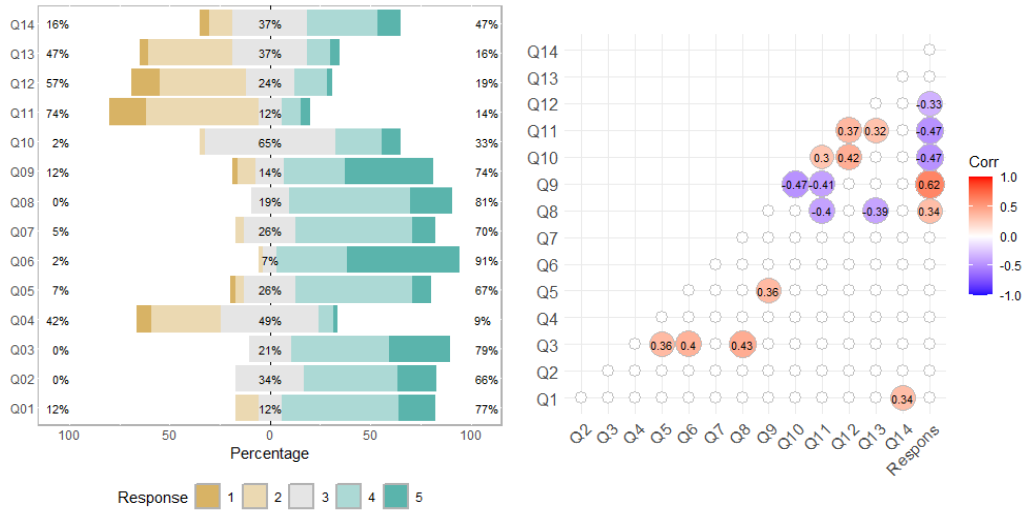

Figure 5. Left: Plot of Likert-items for the 14 statements (Q1-14). Response levels for each item are explained in Table 1. Numbers to the left are percentages for levels 1+2, and to the right, for levels 4+5. The percentage for the intermediate level is given on the grey symbol in the middle. Note that statements Q4 and Q10-13 have reverse levels: low levels indicate positive affirmation, and vice versa. Right: Visualisation of the Kendall Tau correlation matrix with only significant correlation coefficients visible. *p*-value level <= 0.05.

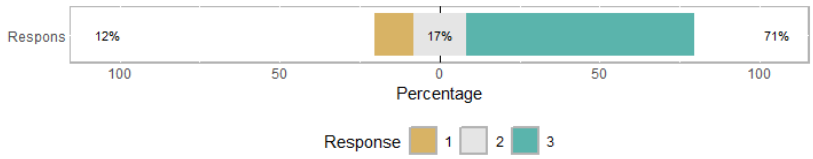

Figure 6. Plot of Likert-item *Response* with self-reported results after the *gtest* compared to the *itest*, 1 = lower result, 2 = same result and 3 = improved results.

The group of students who did not improve their grades after the *gtest* (non-improvers) had markedly different responses to Q5 and Q9-12 than those who did (improvers) (Figure 7). While most non-improvers *rarely* (17%) or only *sometimes* (42%) felt they learned something during the peer dialogue, only 23% of the improvers had the same experience (Q5). Non-improvers also agreed more often than improvers that free-riding was a problem (Q12: 36% vs 10%). This also made them more negative to the test itself (Q11) and its result (Q9).

Two open-field comments described different opinions about the relation between the test, group work and motivation:

> "I had read a lot beforehand, so I knew a lot from before, but the group work meant that I was able to answer the things I didn't know before. I liked the group work because I got along well with my group, and we did really well on the group part." Respondent 24 (translated by Google)

> «The fact that the multiple-choice test was divided into individual parts and group parts meant that I did not focus so much on practising for the exam. For my part, I think that the motivation for learning and preparing for the exam would have been greater if it were only individual.» Respondent 25 (translated by Google)

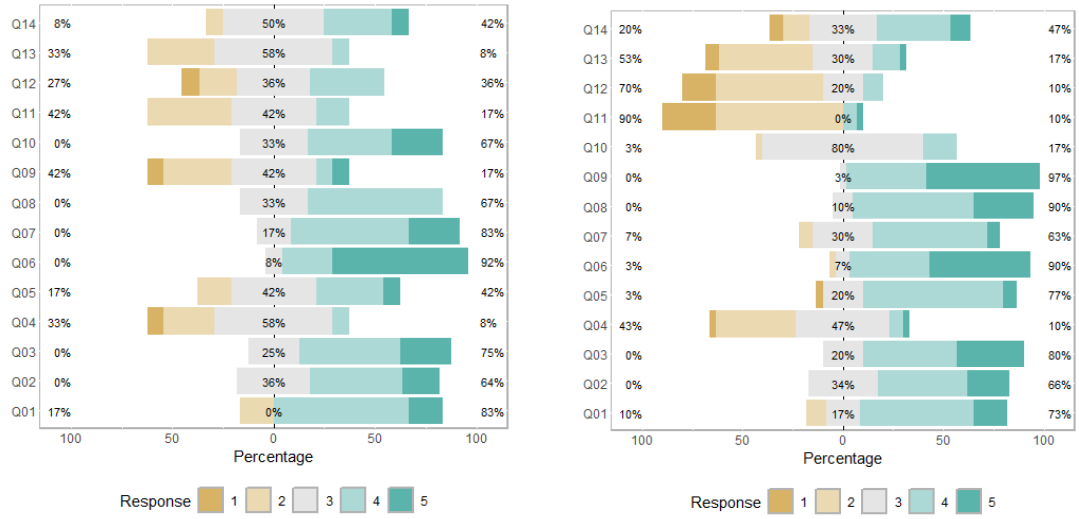

Figure 7: Likert-item responses for non-improvers (left) and improvers (right). Non-improvers are students who reported that their grade on the *gtest* was the same or lower than on the *itest*. Improvers performed better on the *gtest* than on the *itest*. Numbers to the left are percentages for levels 1+2, and to the right, for levels 4+5. The percentage for the intermediate level is given on the grey symbol in the middle. Note that statements Q4 and Q10-13 have reverse levels: low levels indicate positive affirmation, and vice versa.

Discussion

The study's results support previous findings that performance improves during the collaborative stage (cf. summary in Lee et al., 2022). Moreover, they align with findings summarised in Sibley & Ostafichuk (2023) that groups outperform individual achievements. The presence of possible MKOs within each group probably increased the likelihood of finding correct answers and of performing better. Because the test covered various topics, students might perceive different difficulty levels, increasing the chance that someone in the group is an MKO on at least one topic. This likely relates to respondent 24's comment. The group effect was strongest for H23, possibly because their initial *itest* median was grade D. A lower starting point offers greater potential for improvement and for acting as an MKO on some topics.

Generally, improved performance after the *gtest* suggests that most students benefit from this assessment form. But the findings also suggest that low performers consistently benefited more. This is in line with the findings of other studies (Bruno et al., 2017; Giuliodori et al., 2008). However, the pattern of relative improvement for H23 (Figure 3) is more nuanced. It also aligns with studies showing that benefits can be distributed equally across performance levels (Gilley & Clarkston, 2014; Mahoney & Harris-Reeves, 2019). This pattern is likely related to the presence of more low-performing students in H23, suggesting that some groups may have consisted solely of low performers. Groups with a mix of high- and low-performing students typically improve the most (e.g. see Bruno et al., 2017), whereas the outcomes of those with only low-performing students are more variable, contributing to the U-shaped pattern. While all students benefited from this assessment form, this study suggests that, most commonly, low-performing students benefited more than higher-performing students.

Consensus-finding is a key to success in two-stage exams and depends on well-functioning groups. In this study, consensus did indeed increase after the collaborative stage for both cohorts. Notably, the H24 cohort reached consensus more often than the

H23 cohort. Team functioning, related to both group composition and internal dynamics, may be one reason for this difference. Groups in both cohorts were teacher-formed, which has been shown to be more effective than student-created ones (as cited in Sibley & Ostafichuk, 2023). However, a key difference was the introduction of mandatory group-based learning activities prior to the H24 assessment. 77% of the H24 students found those important for helping their groups function effectively. The H23 cohort, on the other hand, had less collaborative experience in assigned groups. Research in Team-Based Learning has found that with as few as four cycles of readiness assurance tests (two-stage tests for formative purposes), teams have started making consensus-based decisions (as discussed in Sibley & Ostafichuk, 2023). Although three assignments in this course provided limited time for collaboration, they seem to have been sufficient for developing a team-feeling. According to Tuckman's (1965) model on stages in team development, teams typically go through the stages of forming, storming, norming, and performing before functioning effectively. The finding that 66% of respondents reported that their group functioned well or excellent before the assessment supports this interpretation that H24 groups had progressed to higher stages of team development. That group decisions were *always* or *often* reported as consensus-based (Q3; 79%) also aligns with quantitative results showing fewer distinct grades within groups. Furthermore, half of the students reported that early knowledge about the *gtest* motivated them to work effectively in groups. Altogether, these findings also suggest that many groups experienced positive interdependence in the collaborative stage.

Rieger & Rieger (2020) found that negative feedback is rare in two-stage exams, but when reported, it is often related to group dynamics. This is also the case in this study. A minority of respondents (12%) were dissatisfied with both their results and the *gtest* itself. These non-improvers agreed more often that there were free-riders in their groups. These free-riders, most likely among low-performing students, were probably disproportionately favoured during the collaborative stage. The free-riding attitude, reflected in the quote from respondent 25, typically leads to unfair results and frustration among students (Hall & Buzwell, 2013). It is likely that the reported dissatisfaction with the collaborative stage of the exam stems from this dynamic.

Reviews of social loafing in collective work indicate that it "is consistently obtained and moderate in magnitude" (Karau & Williams, 1993). Also, in this study, free-riding was moderate in magnitude. Still, reducing it further should be an aim. Zipp (2007) has suggested that fostering individual responsibility may reduce the risk of free-riding. This may be achieved by weighting the individual part more heavily, up to 90%, or by using more sophisticated regimes that account for group performance (Beatty, 2015; Callaghan et al., 2025; Rieger & Rieger, 2020; Zipp, 2007). The latter, however, may be difficult to incorporate into assessment platforms that automatically calculate test results. Delivering individually in stage two can also foster individual responsibility and prevent "unfair" out-voting of high performers (Zipp, 2007). This was the approach used in this study, but combining it with a heavier weighting of the individual stage seems more promising for reducing free-riding.

The evaluation of the implemented two-stage exam has shown improved performance. Nevertheless, improved performance alone cannot be interpreted as a sign of learning when the study design is exploratory and uncontrolled. Improved performance may result from exchanging correct answer alternatives, without prior peer dialogue known to facilitate learning. Even the fact that 67% of respondents always or often felt they had learned something during the collaborative stage is not a sufficient sign of actual learning (Deslauriers et al., 2019). The learning process is supported by internal feedback mechanisms, facilitated by comparison of knowledge through peer feedback and self-regulation (Nicol, 2007; Nicol & Selvaretnam, 2022). It is therefore reassuring that 79%

and 91% reported positively on indicators suggesting that learning took place, namely consensus-finding and feeling safe and comfortable during discussions. Lively discussions were indeed observed in all groups during the collaborative stage. Similar high engagement has rarely been experienced during teaching activities in this course. High engagement is often attributed to the high stakes involved with summative assessments (Rieger & Heiner, 2014). Altogether, these factors suggest that collaboration was effective and that positive interdependence existed, promoting learning.

Limitations

Given the study design and unknown cohort differences, the findings should be interpreted cautiously. Without testing retention, it is difficult to draw firm conclusions about the learning effect; however, retention testing also has its challenges (Cooke et al., 2019). Complementing the quantitative data with qualitative data was therefore important. Still, a richer questionnaire also including standardised questions, supplemented by individual or focus-group interviews, would have further strengthened the findings. This could be explored in follow-up studies. The quantitative analysis could have been further improved by basing it strictly on similar questions and using percentage scores rather than letter grades. This would have enabled direct comparison of the two cohorts, possibly giving more insight into the effect of prior mandatory assignments. This would also have allowed controlling for item quality, such as discrimination and difficulty, a recommended strategy for improving the quality and learning potential of online tests (Brown & Abdulnabi, 2017). Finally, a more comprehensive quantitative analysis of group behaviour could provide further insights into the quality of collaboration in the second stage (Beatty, 2015).

Long-term educational impact

With the use of generative AI, not only is active learning challenged, but the long-term educational impact is as well. The latter is an important goal of assessment and implies that the learning outcomes of courses and programmes become an integral part of students' competency toolkit, beyond the assessment itself.

There are several ways to address this challenge. In this broader context, two-stage exams offer the advantage of combining summative assessment in a controlled, AI-free environment with social interaction through peer dialogue. This combination can contribute to long-term educational impact by potentially increasing student motivation, fostering an inclusive student environment, and improving learning beyond purely cognitive aspects.

The push to study created by summative assessment is motivating for most students. It is also well known that exams may create anxiety about not performing well enough. However, this anxiety is often reduced or even eliminated by the collaborative aspect (Efu, 2019, Levy et al, 2023). Knowing that most students benefit from the collaborative stage can also be motivating (Jang et al., 2017). This may be especially encouraging for low-performing students, contributing to their self-confidence. For first-semester students, encountering this format early in their course of study can be reassuring.

The two-stage exam promotes social interaction. This aspect can be easily developed by encouraging learning activities within the assigned test groups from the start of the semester. This study indicates that such activities had a positive effect on creating a team feeling. This can further help build inclusive student environments. This is arguably very valuable for students in general, and particularly for first-semester students.

Two-stage exams integrate active learning into the assessment itself, addressing the paradox that, while active teaching and learning practices promote collaboration and peer

dialogue among students, these principles are often abandoned in assessment (Jang et al., 2017; Ley et al., 1995; Wieman et al., 2014; Zipp, 2007). By acknowledging that assessment is about more than grading, it supports learning beyond the strictly cognitive aspects. While the latter are the sole focus of, for instance, Bloom's taxonomy (Bloom et al., 1956), other learning frameworks, such as Significant Learning by Fink (2013), define learning as an overall experience. It includes learning related to aspects such as the Human dimension, Caring, and Learning how to learn, including, among others, self-learning and intrapersonal skills. These skills are as important as cognitive skills in securing lifelong learning, both independently and in collaboration.

Conclusion

This study found that the collaborative stage of two-stage exams benefited all students, with low-performing students benefiting the most. Although a minority of students reported challenges with free-riding, groups collaborated effectively, especially in the cohort with prior collaborative experience. Questionnaire responses suggest that this prior experience fostered a safe and trusting environment for exchanging ideas during peer dialogue in the collaborative stage. Altogether, the findings suggest that collaboration was effective and that positive interdependence existed, promoting learning.

Future AI-robust and sustainable assessment methods must utilise tools that help students recognise what learning is and that 'genuine' intelligence, in combination with social interaction, is fundamental for lifelong learning. Two-stage exams could be one of several such measures.

About the Author

Gidske L. Andersen is an associate professor in the Department of Geography, teaching GIS, remote sensing, cartography, and environmental geography.

Acknowledgement

Two anonymous referees and editors provided valuable feedback that enhanced the quality of this paper.

References

Beatty, I. D. (2015). Collaboration or copying? Student behavior during two-phase exams with individual and team phases. *Physics education research conference 2015*. PER conference, College Park, MD,

Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher education research & development*, *18*(1), 57–75. https://doi.org/10.1080/0729436990180105

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York.

Bloom, D. (2009). Collaborative test taking: Benefits for learning and retention. *College Teaching*, *57*(4), 216–220. https://doi.org/10.1080/87567550903218646

Boitshwarelo, B., Reedy, A. K., & Billany, T. (2017). Envisioning the use of online tests in assessing twenty-first century learning: a literature review. *Research and Practice in Technology Enhanced Learning*, *12*, 1–16. https://doi.org/10.1186/s41039-017-0055-7

Brown, G. T., & Abdulnabi, H. H. (2017). Evaluating the quality of higher education

instructor-constructed multiple-choice tests: Impact on student grades. Frontiers in Education. https://doi.org/10.3389/feduc.2017.00024

Bruno, B. C., Engels, J., Ito, G., Gillis-Davis, J., Dulai, H., Carter, G., Fletcher, C., & Böttjer-Wilson, D. (2017). Two-Stage Exams A Powerful Tool for Reducing the Achievement Gap in Undergraduate Oceanography and Geology Classes. *Oceanography*, *30*(2), 198–208. https://doi.org/10.5670/oceanog.2017.241

Bryer, J., Speerschneider, K., & Bryer, M. J. (2016). Package 'likert'. *Likert: Analysis and Visualization Likert Items (1.3. 5)[Computer software]. Available online at: https://CRAN. R-project. org/package= likert (accessed June, 2025)*.

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition*, *7*(3), 323–331. https://doi.org/10.1016/j.jarmac.2018.07.002

Callaghan, K., Kestin, G., Klales, A., McCarty, L., & Deslauriers, L. (2025). Active learning through flexible collaborative exams: Improving assessments across disciplines. *Active Learning in Higher Education*. https://doi.org/10.1177/14697874251344293

Cooke, J. E., Weir, L., & Clarkston, B. (2019). Retention following two-stage collaborative exams depends on timing and student performance. *CBE—Life Sciences Education*, *18*(2), ar12. https://doi.org/10.1187/cbe.17-07-0137

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, *116*(39), 19251–19257. https://doi.org/10.1073/pnas.1821936116

Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved Learning in a Large-Enrollment Physics Class. *Science*, *332*(6031), 862–864. https://doi.org/10.1126/science.1201783

Efu, S. I. (2019). Exams as learning tools: A comparison of traditional and collaborative assessment in higher education. *College Teaching*, *67*(1), 73–83. https://doi.org/10.1080/87567555.2018.1531282

Fink, L. D. (2013). *Creating significant learning experiences: An integrated approach to designing college courses*. John Wiley & Sons.

Gilley, B. H., & Clarkston, B. (2014). Collaborative testing: Evidence of learning in a controlled in-class study of undergraduate students. *Journal of College Science Teaching*, *43*(3), 83–91. https://doi.org/10.2505/4/jcst14_043_03_83

Giuliodori, M. J., Lujan, H. L., & DiCarlo, S. E. (2008). Collaborative group testing benefits high-and low-performing students. *Advances in Physiology Education*, *32*(4), 274–278. https://doi.org/10.1152/advan.00101.2007

Hall, D., & Buzwell, S. (2013). The problem of free-riding in group projects: Looking beyond social loafing as reason for non-contribution. *Active Learning in Higher Education*, *14*(1), 37–49. https://doi.org/10.1177/1469787412467123

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, *7*(6), 836–850. http://dx.doi.org/10.1016/j.cptl.2015.08.001

Jang, H., Lasry, N., Miller, K., & Mazur, E. (2017). Collaborative exams: cheating? Or learning? *American Journal of Physics*, *85*(3), 223–227. https://doi.org/10.1119/1.4974744

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher*, *38*(5), 365–379. https://doi.org/10.3102/0013189X09339057

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical

integration. *Journal of Personality and Social Psychology*, *65*(4), 681.

Kassambara, A. (2023). ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'. *R package version 0.1.4.1*.

Lee, T. R. C., Pye, M., Lilje, O., Nguyen, H. D., Hockey, S., De Bruyn, M., & Van den Berg, F. T. (2022). Two-stage Examinations in STEM: A Narrative Literature Review. *International Journal of Innovation in Science and Mathematics Education*, *30*(5). https://doi.org/10.30722/IJISME.30.05.005

Levy, D., Svoronos, T., & Klinger, M. (2023). Two-stage examinations: Can examinations be more formative experiences? *Active Learning in Higher Education*, *24*(2), 79–94. https://doi.org/10.1177/1469787418801668

Ley, K., Hodges, R., & Young, D. (1995). Partner testing. *Research and Teaching in Developmental Education*, 23–30. https://www.jstor.org/stable/42802445

Mahoney, J. W., & Harris-Reeves, B. (2019). The effects of collaborative testing on higher order thinking: Do the bright get brighter? *Active Learning in Higher Education*, *20*(1), 25–37. https://doi.org/10.1177/1469787417723243

Malthe-Sørenssen-utvalget (2026). *Notat: Foreløpige vurderinger*. Retrieved 31.01.2026 from https://malthesorenssenutvalget.no/notat-forelopige-vurderinger/

Martin, A. P. (2018). A Quantitative Framework for the Analysis of Two-Stage Exams. *International Journal of Higher Education*, *7*(4), 33–54. https://doi.org/10.5430/ijhe.v7n4p33

Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and Higher Education*, *31*(1), 53–64. ttps://doi.org/10.1080/03098770601167922

Nicol, D., & Selvaretnam, G. (2022). Making internal feedback explicit: harnessing the comparisons students make during two-stage exams. *Assessment & Evaluation in Higher Education*, *47*(4), 507–522. https://doi.org/10.1080/02602938.2021.1934653

R Core Team, R. (2025). R: A language and environment for statistical computing.

Revelle, W. (2025). psych: Procedures for psychological, psychometric, and personality research. *R package version 2.5.3*.

Rieger, G. W., & Heiner, C. E. (2014). Examinations that support collaborative learning: The students' perspective. *Journal of College Science Teaching*, *43*(4), 41–47. https://doi.org/10.2505/4/jcst14_043_04_41

Rieger, G. W., & Rieger, C. L. (2020). Collaborative assessment that supports learning. In *Active learning in college science: The case for evidence-based practice* (pp. 821–837). Springer.

Sibley, J., & Ostafichuk, P. (2023). *Getting started with team-based learning*. Taylor & Francis.

Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological Bulletin*, *63*(6), 384-399. https://doi.org/10.1037/h0022100

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth ed.). Springer.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wieman, C. E., Rieger, G. W., & Heiner, C. E. (2014). Physics exams that promote collaborative learning. *The Physics Teacher*, *52*(1), 51–53.

https://doi.org/10.1119/1.4849159

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Zipp, J. F. (2007). Learning by exams: The impact of two-stage cooperative tests. *Teaching Sociology*, *35*(1), 62–76. https://doi.org/10.1177/0092055X0703500105