

Does format of group stage matter? Consensus vs. Non-consensus

By

Anita Ayame

A Thesis submitted to the Department of Biological Sciences, University of Manitoba,

In partial fulfillment of the requirements for the course

BIOL 4100 (Honours Thesis)

for the degree of

Bachelor of Science (Honours)

© March 2026

Abstract

Examinations remain a central feature of undergraduate Science, Technology, Engineering, and Mathematics education, shaping both performance and student experience. Two-stage examinations introduce a collaborative component into this setting, allowing students to engage in peer discussion immediately following individual assessment. Differences in group-stage structure, however, raise important questions about how collaborative interaction influences performance and affective outcomes.

A repeated-measures crossover design was implemented in BIOL 1020 at the University of Manitoba. Twenty-seven consenting students completed four two-stage assessments under alternating consensus and non-consensus formats. Consensus groups were required to submit a single agreed-upon response, whereas non-consensus groups discussed collaboratively but submitted individual answers. Collaborative benefit, defined as group-stage score minus individual-stage score, served as the primary outcome, alongside measures of anxiety reduction, motivation, and student experience.

Collaborative benefit was significantly higher under the consensus format relative to the non-consensus format, with a mean difference of 6.9 percentage points, while individual-stage performance remained unchanged. Performance gains remained stable across repeated assessments. Positive associations between collaborative benefit and measures of anxiety reduction, motivation, and student experience were also observed.

Group-stage structure shapes the effectiveness of collaborative assessment. Greater collaborative benefit observed under consensus conditions is consistent with the interpretation that requiring agreement may promote greater justification of reasoning, evaluation of alternative responses, and resolution of disagreement during the group stage. In contrast, non-consensus conditions

may allow participation without requiring convergence of reasoning, which could limit the extent of interaction required for answer revision. The selection of group-stage formats therefore represents a meaningful instructional decision in the design of assessment practices in undergraduate science education.

Acknowledgements

This work would not have been possible without the people who stood beside me, challenged me, and believed in me during moments when I struggled to believe in myself.

I would first like to express my deepest gratitude to Dr. Cassandra Debets. Thank you for taking a chance on me, for your patience, and for creating space for me to not only conduct research, but to truly express my thoughts, perspectives, and curiosity. At a time when I questioned my place in science, you helped restore my confidence and gave me renewed hope in continuing this path. That is something I will carry with me far beyond this thesis.

To Dr. Jane Waterman, thank you for your honesty, strength, and consistency. You held a standard that required me to rise, even on the days I felt like I could not. Your presence, especially in moments where showing up was not easy, reminded me that commitment is not about convenience, but about purpose. Knowing you would always show up made it impossible for me not to try.

I am also deeply grateful to Dr. Madison Egan for her care, thoughtfulness, and clarity. Your ability to guide with both kindness and direction created a space where I could grow without fear, and your feedback consistently pushed this work to be stronger.

I would like to extend my heartfelt thanks to Ms. Shanda Vitt, whose support has been foundational to my journey. Your proactive advocacy, care, and belief in my potential opened doors that I may not have been able to access on my own. You are the reason I was able to pursue the Honours program, and your impact on my academic path is something I will never forget.

I extend my appreciation to the Department of Biological Sciences at the University of Manitoba for providing not only the opportunity to complete this thesis, but also the understanding and

accommodations that made it possible during a particularly challenging period of my health.

That support did not go unnoticed.

To my family—Marvellous, Dillion, Fedora, and especially my father, Shadrack, and my mother, Esther—there are no words strong enough to capture what your support has meant. You stood by me through uncertainty, exhaustion, and moments of doubt, never once wavering. In times when my health made everything feel unstable, you were my constant.

Finally, to Shane Cummings, thank you for walking this journey with me in a way few people could. Despite not being in the sciences, you read every version of this work, challenged my thinking, and somehow came to understand it as deeply—if not more deeply—than I did at times. Your patience, presence, and belief in me carried me through more than you know.

Table of Contents

Introduction	1
Methods.....	4
Study Context and Participants.....	4
Study Design	4
Performance Measures	5
Perception Measures.....	6
Statistical Analysis	7
Ethics Statements	8
Results	8
Individual Performance.....	8
Collaborative Benefit	8
Stability Across Assessments.....	9
Student Perceptions	11
Perception – Performance Relationships.....	15
Linear Mixed Model	15
Discussion.....	16
Effect of group-stage format on collaborative performance.....	16
Repeated exposure and performance gains.....	17
Association between student experience and collaborative performance	18
Mechanisms underlying format differences	19
Individual performance and baseline knowledge	21
Stability of effects across assessments	21
Recommendations	23
Implications for assessment design.....	25
Conclusions and future directions	25

List of Tables

Table 1. Group-stage format by assessment and cohort	5
Table 2. Survey Items Included in Analysis and Operational Definitions	6
Table 3. Descriptive statistics for individual-stage and group-stage performance across the four assessments.	10
Table 4. Mean perception scores by assessment and group-stage format for anxiety reduction, motivation, and student experience.....	13
Table 5. Mean perception scores by group and group-stage format.	14
Table 6. Kendall's tau correlations between collaborative benefit and student perceptions.	15

List of Figures

Figure 1. Individual-stage and group-stage score distributions across the four assessments under rotating group-stage formats	9
Figure 2. Distribution of student responses to perception items under the consensus and non-consensus group-stage formats.	12

Introduction

Social learning theory explains that students construct knowledge through interactions, discussions, and shared problem solving and that they do not learn in isolation (Bandura 1977; Vygotsky 1978). In this view, learning is a social process. Research on observational learning demonstrates that people learn new behaviors by watching and imitating others (Bandura 1977), while sociocultural theory emphasizes that interaction and dialogue can extend a learner's potential understanding (Vygotsky 1978). Peer dialogue allows students to articulate their reasoning, challenge ideas, consider different perspectives, and refine their understanding of subject material in the classroom (Meaders and Vega 2022).

Active learning strategies extend principles of social learning by engaging students directly in the learning process. Large-scale evidence demonstrates that active learning improves student performance and reduces failure rates in STEM courses (Freeman et al. 2014). Collaborative problem-solving and peer discussion are central components of these approaches, reinforcing the role of interaction in shaping learning outcomes. Despite these advances in instructional practice, assessment methods have not always evolved in parallel.

Traditional summative examinations remain one of the most common methods used to assess student learning in higher education (Brown and Knight 1994; Struyven et al. 2005). Individual examinations typically evaluate performance under timed conditions and emphasize recall, application, and problem-solving. Although individual exams provide accountability, high-pressure testing environments can elevate test-related anxiety and limit opportunities for immediate feedback or collaborative reasoning (Hembree 1988; Putwain 2007). Concerns about anxiety and limited formative feedback have motivated the development of alternative assessment structures that preserve rigor while supporting student learning and experience.

Two-stage examinations represent one alternative assessment structure that can transform traditionally summative assessments into more formative learning experiences (Levy et al. 2018). In a two-stage exam, students first complete an assessment independently and then immediately retake the same assessment in small groups (Lindsley et al. 2016). The two-stage structure preserves individual accountability while introducing collaborative discussion. Empirical studies report that two-stage exams improve immediate performance (Gilley and Clarkston 2014), enhance retention under certain conditions (Cooke et al. 2019), and reduce perceived test anxiety (Fournier et al. 2017). Student perception research further indicates that collaborative testing is often viewed positively, with reported increases in engagement and classroom satisfaction (Khong and Tanner 2020; Rempel et al. 2023).

Despite documented advantages, outcomes associated with two-stage examinations vary across contexts. Some investigations report short-term performance gains without consistent long-term retention effects (Leight et al. 2012; Kinnear 2021). Differences in question type, assessment weighting, and classroom implementation may influence effectiveness (Ives et al. 2016; Shaffer 2020). Variation in implementation suggests that specific structural features of the group stage may influence collaborative outcomes.

One such structural feature that has received limited direct comparison is the format of the collaborative group stage which may be implemented as either a consensus or non-consensus approach. In a consensus format, group members are required to agree on a single answer that is submitted collectively on behalf of the entire group (Cortright et al. 2003; McCurdy et al. 2017). Requiring agreement may encourage justification of reasoning and resolution of misconceptions prior to submission. In a non-consensus format, students discuss questions collaboratively but submit individual responses. Maintaining individual submission after discussion may preserve

accountability while reducing the need for full resolution of disagreement. Direct comparisons of consensus and non-consensus formats remain limited in the literature (Shaffer 2020; Kinnear 2021).

Clarifying whether group-stage format influences collaborative performance is important for instructional design. Demonstrating superior performance under one format would indicate that group-structure requirements meaningfully shape collaborative reasoning during assessment. Conversely, comparable performance across formats would suggest that instructors may select implementation strategies based on logistical or pedagogical preference without substantial performance consequences.

The research objective of this study was to examine whether group-stage format in two-stage examinations influences student performance within an introductory biology course.

I hypothesized that the group-stage format influences collaborative performance between the individual and group stages of two-stage examinations. Specifically, I predict that students in consensus groups are expected to achieve greater collaborative benefit compared to those in non-consensus groups, reflecting more effective collaboration and shared reasoning.

I further hypothesize that collaborative performance varies across assessments with repeated exposure to the two-stage format. I predict that performance gains are expected to remain stable or increase slightly over time, indicating adaptation and improved coordination across repeated assessments.

I also hypothesize that student perceptions of the assessment experience are associated with collaborative performance. I predict that more positive self-reported experiences, including reduced anxiety and increased motivation, are expected to align with greater collaborative

benefit, indicating that affective factors are linked to performance during collaborative assessment.

Collectively, these hypotheses examine how group-stage format, repeated exposure, and student experience interact to shape collaborative performance in two-stage examinations.

Isolating the role of group-stage format within a controlled crossover design contributes to understanding how specific structural features of collaborative examinations influence both performance and student experience in undergraduate science education.

Methods

Study Context and Participants

The study was conducted in BIOL 1020: Biology 1 at the University of Manitoba during the six-week Summer 2024 term in Winnipeg, Manitoba. The course covered topics in cell biology, evolution, and genetics. All enrolled students completed four two-stage in-class tests that were comprised of 15 multiple choice questions, each worth 4% of the final course grade.

Participation in the research component was voluntary and not linked to course grades. Of the 122 enrolled students, 27 students consented to the use of their performance and survey data for research purposes (22% response rate). All analyses were conducted on consenting participants only. The compressed six-week term and the voluntary nature of participation likely contributed to the reduced sample size.

Study Design

A crossover repeated-measures design was used to compare two group-stage formats: consensus and non-consensus. All consenting students experienced both formats across the four assessments.

Students were assigned to two groups (Group X and Group Y), each with their own alternating format sequences based on student number (Table 1). Randomization was used to determine order. Student numbers were listed in order and assigned a random number by a random number generator in Excel. For each test, students first completed the test independently under timed conditions (25 minutes). Immediately following the individual stage, students completed the same assessment in groups of three or four students (12 minutes). Groups were reassigned at each assessment to reduce familiarity effects. Groups of two or five were not permitted, to maintain consistency in group interaction dynamics.

Table 1. Group-stage format by assessment and cohort.

Assessment	Group X Format	Group Y Format
Test 1	Consensus	Non-consensus
Test 2	Non-consensus	Consensus
Test 3	Consensus	Non-consensus
Test 4	Non-consensus	Consensus

The individual stage contributed 85% of the total test grade, and the group stage contributed 15%. If a group-stage score was lower than the corresponding individual-stage score, the individual score was retained. Students who missed the group stage for a given assessment were excluded from analyses for that assessment only but remained included in analyses of other completed assessments.

Performance Measures

Individual-stage and group-stage scores were recorded as percentages of correct responses for each assessment.

Collaborative benefit was calculated as the unstandardized difference between group-stage and individual-stage scores:

$$\text{Collaborative Benefit} = \text{Group Score} - \text{Individual Score}$$

Collaborative benefit served as the primary outcome variable for evaluating format effects.

Perception Measures

Immediately following each assessment, students were invited to complete a brief survey measuring perceptions of stress, motivation, and overall experience during the two-stage examination. Survey items were adapted from previously published studies of collaborative testing (Fournier et al. 2017; Shaffer 2020; Khong and Tanner 2020).

Table 2. Survey Items Included in Analysis and Operational Definitions

Construct	Survey Item	Scale
Anxiety Reduction	I feel less anxious about examinations after taking two-stage exams.	5-point Likert (1 = Strongly Disagree to 5 = Strongly Agree)
Motivation	I feel motivated about my studies after the two-stage exam.	5-point Likert
Student Experience	The two-stage exam improved my student experience.	5-point Likert

Responses were recorded on a five-point Likert scale (1 = strongly disagree to 5 = strongly agree). Analyses focused on three affective constructs: anxiety reduction, motivation, and overall student experience of the assessment format. Selection of anxiety reduction and

motivation reflects their established roles as affective responses to assessment contexts. Overall student experience was included as a global evaluative measure of the two-stage examination format (Table 2). Survey items assessing communication behaviors and peer interaction were excluded from inferential analysis to maintain focus on affective outcomes relevant to performance. Survey response summaries were calculated using available responses rather than restricting analyses to participants with complete survey data across all assessments. Consequently, the sample size differed across perception items and assessments. The full survey instrument is provided in Appendix A.

Statistical Analysis

All statistical analyses were conducted in R (version 4.3).

Descriptive statistics were calculated for individual-stage performance, group-stage performance, collaborative benefit, anxiety reduction, motivation, and student experience. Collaborative benefit was calculated as the unstandardized difference between group-stage and individual-stage scores (Group – Individual), providing a direct estimate of performance change attributable to the group stage.

Because all participants experienced both consensus and non-consensus formats, mean collaborative benefit was compared between formats using a paired-samples t-test.

Differences in performance across the four assessments were examined using repeated-measures analysis of variance (ANOVA), with assessment number (1–4) treated as a within-subject factor.

Associations between collaborative benefit and affective measures (anxiety reduction, motivation, and student experience) were examined using Kendall's tau correlation, as perception data were ordinal.

A linear mixed-effects model was used to examine the effects of group-stage format and assessment number on collaborative benefit. Collaborative benefit was specified as the dependent variable, with group-stage format (consensus, non-consensus) and test (1–4) included as fixed effects. Participant ID was included as a random intercept to account for repeated observations within individuals.

Ethics Statements

The study was approved by the University of Manitoba Research Ethics Board (Protocol HE2023-0171). Only coded data were analyzed.

Results

Individual Performance

Individual-stage scores did not differ significantly across tests, $F_{(3, 78)} = 1.928, p = 0.132$, with the lowest mean score observed in Test 2 (62.5 ± 21.1) and the highest in Test 1 (72.8 ± 16.1) (Table 3). Participant-level mean individual-stage scores were 67.0 ± 16.9 under the consensus format and 68.1 ± 16.0 under the non-consensus format, with no significant difference between formats, $t_{26} = -0.436, p = 0.666$.

Collaborative Benefit

Collaborative benefit was calculated as the unstandardized difference between group-stage and individual-stage performance (group – individual). Participant-level mean collaborative benefit was 20.1 ± 15.1 under the consensus format and 13.2 ± 15.8 under the non-consensus format.

Collaborative benefit was significantly higher under the consensus format, $t_{26} = 2.946, p = 0.0067$, with a mean difference of 6.9 percentage points (95% CI: 2.1, 11.7) and a paired Cohen's d of 0.567.

Stability Across Assessments

Descriptive statistics for individual-stage and group-stage performance across the four assessments are presented in Table 3. Individual-stage mean scores ranged from 55.6% in Group Y under the consensus format on Test 2 to 74.4% in Group X under the non-consensus format on Test 4. Group-stage mean scores ranged from 72.6% in Group Y under the consensus format on Test 2 to 91.5% in Group X under the consensus format on Test 3 (Figure 1).

Across the four assessments, group-stage scores were significantly higher than individual-stage scores overall, $F(1, 26) = 37.18, p < 0.001$. Individual-stage performance did not differ significantly across assessments, $F(3, 78) = 1.928, p = 0.132$. Group-stage performance differed significantly across assessments, $F(3, 78) = 3.183, p = 0.029$. Collaborative benefit did not differ significantly across assessments, $F(3, 78) = 1.508, p = 0.219$.

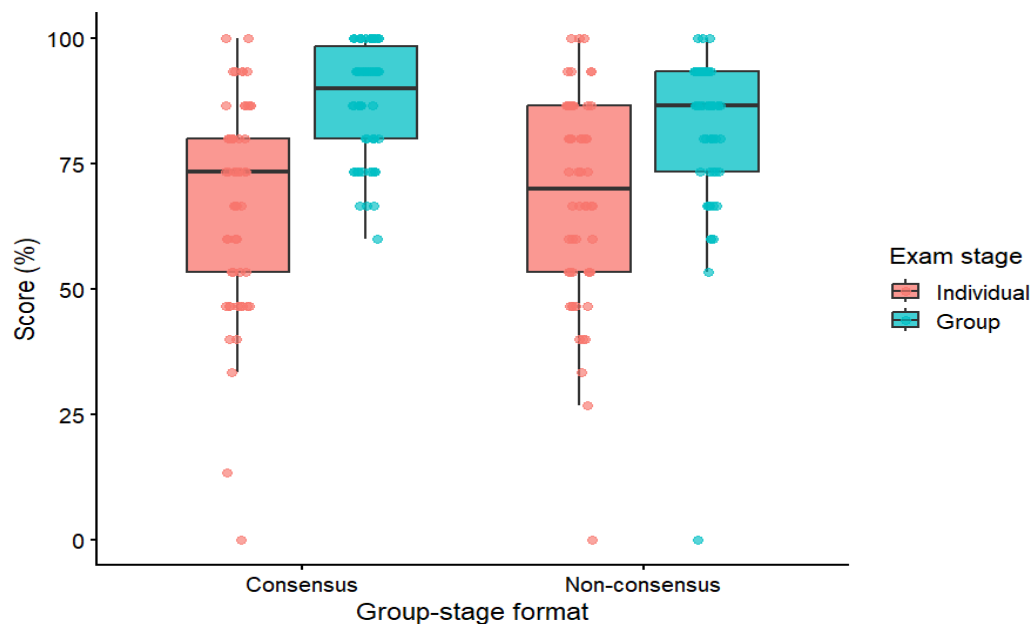


Figure 1. Figure 1. Individual-stage and group-stage score distributions under consensus and non-consensus group-stage formats. Points represent student scores, and boxplots show score distributions. Scores are presented as percentages.

Table 3. Descriptive statistics for individual-stage and group-stage performance across the four assessments.

Assessment & Format	Individual Stage Mean (SD)	Group Stage Mean (SD)	Benefit Mean (SD)
Group X (n=18)			
Test 1	0.737	0.907	0.170
Consensus	(0.170)	(0.089)	(0.210)
Test 2	0.659	0.819	0.159
Non – Consensus	(0.195)	(0.133)	(0.180)
Test 3	0.700	0.915	0.215
Consensus	(0.194)	(0.109)	(0.153)
Test 4	0.744	0.804	0.059
Non - Consensus	(0.245)	(0.213)	(0.162)
Group Y (n=9)			
Test 1	0.711	0.867	0.156
Non - Consensus	(0.149)	(0.088)	(0.173)
Test 2	0.556	0.726	0.170
Consensus	(0.236)	(0.097)	(0.198)
Test 3	0.570	0.770	0.200
Non – Consensus	(0.198)	(0.149)	(0.191)
Test 4	0.593	0.859	0.267
Consensus	(0.255)	(0.052)	(0.226)

Student Perceptions

Under the consensus format, 60% of responses were positive for anxiety reduction, 70% were positive for motivation, and 70% were positive for student experience, whereas 19%, 9%, and 5% of responses, respectively, were negative. Under the non-consensus format, 60% of responses were positive for anxiety reduction, 58% were positive for motivation, and 64% were positive for student experience, whereas 16%, 14%, and 9% of responses, respectively, were negative (Figure 2). Mean perception scores by assessment and group-stage format are presented in Table 4, and group-specific mean perception scores by format are presented in Table 5. Mean perception scores were generally higher under the consensus format compared to the non-consensus format across anxiety reduction, motivation, and student experience (Tables 4–5). Differences between formats were consistent across most assessments, with modest variation in magnitude. Group-level comparisons showed a similar pattern, with higher scores under the consensus format for most measures, although differences were small and not uniform across all comparison.

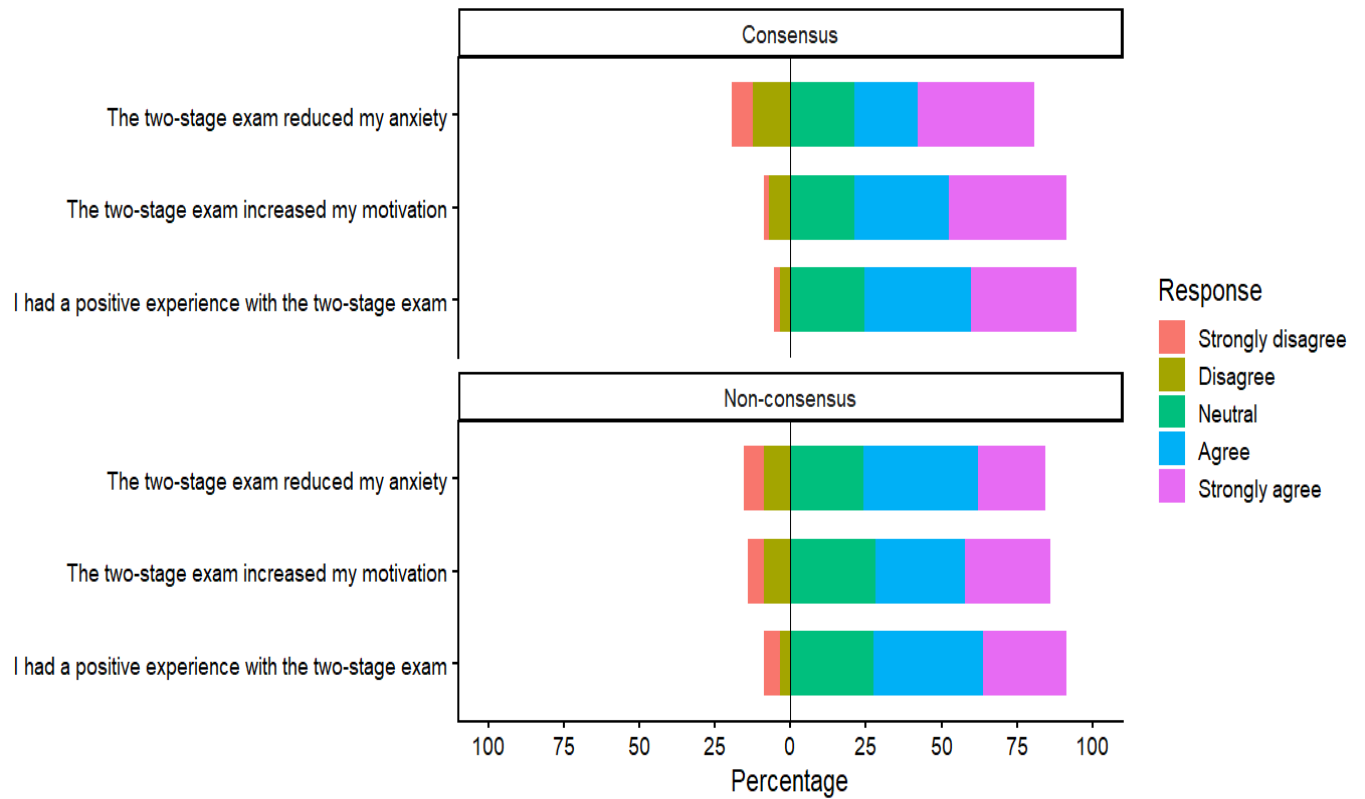


Figure 2. Distribution of student responses to perception items under the consensus and non-consensus group-stage formats. Bars show the percentage of responses in each category for anxiety reduction, motivation, and student experience, with groups combined within each format.

Table 4. Mean perception scores by assessment and group-stage format for anxiety reduction, motivation, and student experience.

Perception	Assessment	Consensus n	Consensus Mean (SD)	Non- consensus n	Non-consensus Mean (SD)
Anxiety Reduction	Test 1	21	3.524 (1.436)	14	3.214 (1.188)
Anxiety Reduction	Test 2	10	3.600 (1.174)	15	3.400 (1.352)
Anxiety Reduction	Test 3	17	4.118 (1.111)	14	3.714 (0.726)
Anxiety Reduction	Test 4	9	3.556 (1.424)	15	4.067 (1.100)
Motivation	Test 1	21	4.048 (0.921)	14	3.357 (1.151)
Motivation	Test 2	10	3.500 (1.354)	15	3.533 (1.125)
Motivation	Test 3	17	4.118 (0.857)	13	3.769 (1.013)
Motivation	Test 4	9	4.111 (1.167)	15	4.000 (1.254)
Student Experience	Test 1	21	4.048 (0.921)	14	3.500 (1.225)
Student Experience	Test 2	10	3.300 (1.059)	15	3.733 (0.961)
Student Experience	Test 3	17	4.294 (0.849)	14	3.929 (0.917)
Student Experience	Test 4	9	4.000 (0.866)	15	3.933 (1.163)

Note. Perception summaries were calculated using available responses for each survey item at each assessment. Sample sizes therefore varied across items, tests, and formats because not all participants completed every survey item at every assessment. Higher scores on the anxiety item reflect greater reported anxiety reduction.

Table 5. Mean perception scores by group and group-stage format.

Group	Format	Anxiety Reduction n	Anxiety Reduction Mean (SD)	Motivation n	Motivation Mean (SD)	Student Experience n	Student Experience Mean (SD)
X	Consensus	38	3.789 (1.318)	38	4.079 (0.882)	38	4.158 (0.886)
X	Non- consensus	30	3.733 (1.258)	30	3.767 (1.194)	30	3.833 (1.053)
Y	Consensus	19	3.579 (1.261)	19	3.789 (1.273)	19	3.632 (1.012)
Y	Non- consensus	28	3.464 (0.999)	27	3.556 (1.086)	28	3.714 (1.084)

Note. Values are averaged within group and format across the relevant assessments. For Group X, consensus-format values reflect Tests 1 and 3 and non-consensus values reflect Tests 2 and 4. Group Y followed the opposite sequence. Higher scores on the anxiety item reflect greater reported anxiety reduction.

Perception – Performance Relationships

All three perception variables were positively associated with collaborative benefit (Table 6).

Anxiety reduction and motivation showed the strongest associations, with slightly lower values observed for student experience. (Table 6).

Table 6. Kendall's tau correlations between collaborative benefit and student perceptions.

Perception	Kendall's τ	p-value
Anxiety reduction	0.226	< 0.001
Motivation	0.228	< 0.001
Student experience	0.212	< 0.001

Linear Mixed Model

A linear mixed-effects model with participant ID as a random intercept showed that collaborative benefit differed significantly by format, $F_{(1,77)} = 4.851, p = 0.031$. Collaborative benefit was lower in the non-consensus format than in the consensus format ($\beta = -0.062, SE = 0.028, t_{(77)} = -2.202, p = 0.031$). The effect of test was not significant, $F_{(3,77)} = 0.940, p = 0.426$. Relative to Test 1, the estimated coefficients were 0.018 for Test 2 ($SE = 0.039, p = 0.639$), 0.044 for Test 3 ($SE = 0.038, p = 0.240$), and -0.016 for Test 4 ($SE = 0.039, p = 0.674$).

Discussion

The present study examined whether consensus and non-consensus formats affect collaborative performance and student experience in two-stage examinations using a repeated-measures crossover design. Collaborative benefit was greater under the consensus format relative to the non-consensus format (Figure 1), whereas individual-stage performance did not differ between formats. Student perceptions of anxiety reduction, motivation, and overall experience were positive under both formats, with modest differences favoring the consensus format in several measures. Differences between formats were therefore expressed in the magnitude of collaborative performance gains rather than in individual performance.

Effect of group-stage format on collaborative performance

Evaluation of the study hypotheses indicates differential support across predicted outcomes. The hypothesis that group-stage format influences collaborative performance between the individual and group stages of two-stage examinations was supported, as students in the consensus format achieved greater collaborative benefit than those in the non-consensus format (Figure 1). Results from the linear mixed-effects model further support the observed difference in collaborative benefit between formats, with lower values under the non-consensus format after accounting for repeated measures within participants. Greater collaborative benefit under consensus formats is consistent with the prediction that requiring agreement promotes more effective collaboration and shared reasoning during the group stage. Requirement for agreement increases the need for explanation and justification of reasoning prior to submission, processes associated with improved conceptual understanding during collaborative assessment (Driessen et al. 2020). Evaluation of alternative responses further promotes deeper engagement with the

material, which is associated with improved performance in active learning environments (Freeman et al. 2014). Structured collaborative assessment formats that require resolution of disagreement, improve answer accuracy and learning outcomes across disciplines (Knierim et al. 2015; Koretsky et al. 2021). Interaction processes increase the likelihood of identifying and correcting misconceptions, resulting in greater improvement between individual and group-stage performance (Bruno et al. 2017).

Repeated exposure and performance gains

The second hypothesis proposed that collaborative performance would vary across assessments with repeated exposure, but this prediction was not supported. Collaborative benefit remained stable across the four tests (Table 3), indicating that performance gains did not increase over time despite repeated exposure to the two-stage format. Stability across assessments suggests that coordination and interaction patterns were established early and did not substantially change across subsequent assessments (Kinnear 2020). Absence of progressive improvement may also reflect the low-stakes nature of the assessments, which can reduce incentives for adaptation and limit effort invested in refining collaborative strategies across tests (Leight et al. 2012). Previous work support that collaborative testing often produces immediate performance benefits without consistent evidence of cumulative improvement across time, with retention effects depending on timing and student performance rather than increasing uniformly with repeated exposure (Leight et al. 2012; Cooke et al. 2019). Variation in outcomes across implementations further suggests that repeated exposure alone does not guarantee increased performance gains and may depend on assessment structure and instructional context (Newton et al. 2019).

Association between student experience and collaborative performance

The third hypothesis, which examined whether student perceptions of the assessment experience are associated with collaborative performance, was supported. More positive self-reported experiences, including reduced anxiety, increased motivation, and improved student experience, were associated with greater collaborative benefit (Table 6). Positive correlations between collaborative benefit and all three perception measures further support the relationship between affective experience and performance. Reduced anxiety during collaborative assessment has been widely reported in two-stage examination studies and is associated with increased willingness to engage in discussion and participate in group problem-solving (Fournier et al. 2017; Newton et al. 2019). Increased motivation and positive student experience have also been linked to higher levels of participation and interaction during collaborative learning activities, which contribute to improved group performance (Rieger and Heiner 2014; Khong and Tanner 2020).

Engagement, confidence, and willingness to participate in discussion influence the effectiveness of collaborative learning processes by increasing the likelihood of explanation, reasoning, and evaluation of alternative responses (Chi and Wylie 2014; Driessen et al. 2020). Greater participation in these interaction processes is associated with improved collaborative outcomes, providing a mechanism linking affective experience to performance during the group stage (Leight et al. 2012; Terry et al. 2017). Collectively, findings indicate that group-stage format and student experience influence collaborative performance, whereas repeated exposure does not produce measurable changes in performance gains within the timeframe of the present study. The observed associations between collaborative benefit and affective measures suggest

that differences between consensus and non-consensus formats may be partially mediated by variation in interaction quality.

Student perception data further contextualize differences between consensus and non-consensus formats. Both formats were associated with positive responses for anxiety reduction (Figure 2; Tables 4–5), consistent with evidence that two-stage examinations reduce test-related stress (Fournier et al. 2017). Motivation and overall student experience were generally higher under the consensus format across most comparisons (Tables 4–5). Requirement for a shared response may increase accountability and promote active participation, contributing to more favorable perceptions relative to the non-consensus format (Freeman et al. 2014). Increased accountability in group settings enhances engagement and responsibility for collective outcomes, contributing to deeper involvement in collaborative tasks (Millis and Cottell 1998; Oakley et al. 2004). Non-consensus formats may nevertheless support flexibility in individual reasoning and reduce pressure associated with reaching agreement, which may be advantageous in certain instructional contexts (Rieger and Heiner 2014; Struyven et al. 2005). Differences in perception were modest, indicating that both formats provide comparable affective benefits, with the consensus format offering additional advantages in engagement.

Mechanisms underlying format differences

Greater collaborative benefit under the consensus format indicates that the requirement for agreement alters the nature of interaction among students during the group stage. Improvements in group-stage performance observed in the present study are consistent with prior work demonstrating that collaborative two-stage examinations enhance performance in undergraduate biology courses, including laboratory-based settings where peer discussion

supports answer accuracy (Meaders and Vega 2022). Under the consensus format, justification of reasoning, resolution of disagreement, and evaluation of competing responses are required prior to submission. Interaction processes of this type improves conceptual understanding and answer accuracy during collaborative assessment and are consistent with models of knowledge construction through social interaction (Vygotsky 1978) and interactive engagement associated with deeper learning (Chi and Wylie 2014; Freeman et al. 2014; Driessen et al. 2020). In contrast, the non-consensus format permits discussion without requiring convergence, allowing individuals to retain initial responses without engaging in full evaluation of alternatives (Shaffer 2020).

Differences in collaborative benefit between formats therefore reflect differences in the extent to which interaction requires resolution of disagreement. Requirement for agreement effectively transforms the group stage from a setting of shared exposure to one of negotiated understanding, in which correctness must be collectively constructed rather than individually assumed (Gilley and Clarkston 2014; Vogler and Robinson 2016). Reaching a shared response requires students to justify reasoning and evaluate alternative answers, processes associated with improved answer accuracy during collaborative assessment (Driessen et al. 2020; Koretsky et al. 2021).

Our findings align with previous work demonstrating that structured peer interaction promotes conceptual change through explanation and evaluation of competing reasoning (Freeman et al. 2014), and that collaborative exam structures enhance learning outcomes across disciplines (Callaghan et al. 2025). Direct comparison of consensus and non-consensus formats remains limited in the literature, and findings provide evidence that group-stage format influences the effectiveness of collaborative interaction during assessment, with the consensus

format producing greater collaborative benefit than the non-consensus format, extending previous work on two-stage collaborative exams by directly comparing formats within the same cohort.

Individual performance and baseline knowledge

Absence of differences in individual-stage performance between consensus and non-consensus formats indicates that group-stage format does not influence baseline knowledge or preparation. Because both formats share an identical individual stage completed prior to group interaction, similarity in performance isolates the effect of format on the collaborative component of the assessment. Differences observed in collaborative benefit therefore arise from interaction processes during the group stage rather than from differences in prior learning. Comparable patterns have been reported in previous studies, where collaborative testing improves group performance without consistent effects on individual outcomes (Leight et al. 2012; Kinnear 2020). The study thus isolates the effect of format on collaborative processes, highlighting the importance of group-stage design.

Stability of effects across assessments

Stability of collaborative benefit across assessments indicates that differences between consensus and non-consensus formats were consistent over repeated exposure. Collaborative benefit did not vary across assessments, indicating that repeated exposure to the two-stage format did not produce cumulative increases in performance gains. The absence of variation across assessments suggests that the relative advantage of the consensus format was stable and not dependent on a specific test or time point, even though group-stage performance itself varied across assessments (Table 3). Stability in collaborative benefit may reflect carryover effects

between tests and variation in group composition, both of which can influence interaction dynamics and limit cumulative changes in performance (Oakley et al. 2004). In addition, collaborative processes may stabilize early, as students establish shared approaches to discussion and decision-making that persist across subsequent tasks. Research on collaborative learning shows that group interaction patterns often form quickly and remain consistent across repeated interactions, reducing the likelihood of progressive change over time (Millis and Cottell 1998; Oakley et al. 2004). Similar task structure and assessment conditions across tests may further constrain opportunities for adaptation, resulting in stable performance differences rather than increasing gains (Leight et al. 2012). Previous studies of two-stage examinations also report immediate performance benefits without consistent cumulative improvement across repeated assessments, with retention and performance effects varying depending on timing and context rather than increasing uniformly with exposure (Leight et al. 2012; Cooke et al. 2019). Structural features of the assessment, including the requirement for agreement in the consensus format, may therefore exert a consistent influence on interaction regardless of assessment timing, contributing to persistent differences in collaborative benefit (Koretsky et al. 2021; Kinnear 2020).

Limitations

Several limitations should be considered. The crossover design strengthens internal comparison between formats; however, carryover effects between assessments cannot be fully excluded (Shadish et al. 2002). Voluntary participation introduces potential self-selection bias, as participants may differ from non-participants in motivation or attitudes toward collaborative learning (Porter and Whitcomb 2005). The study was conducted within a single introductory biology course during a compressed six-week term, limiting generalizability to other contexts.

Perception measures were based on self-reported Likert-scale responses and may be influenced by response bias, although such measures are widely accepted in educational research (Norman 2010). Group composition varied across assessments due to reassignment, which may have influenced interaction dynamics. In addition, group size effects were not examined, although prior work suggests that group structure can influence participation and performance (Oakley et al. 2004).

Recommendations

Based on quantitative analysis of collaborative performance and student perception data, consensus-based group-stage formats are recommended for implementation in two-stage examinations in undergraduate science courses. Higher collaborative benefit observed under the consensus format, supported by statistical analysis and consistent across repeated assessments, indicates that requiring agreement during the group stage promotes more effective collaboration and improved performance outcomes, consistent with evidence that structured collaborative testing enhances learning and answer accuracy (Gilley and Clarkston 2014; Knierim et al. 2015; Koretsky et al. 2021). In cases where agreement is not readily achieved, consensus-based formats typically require groups to continue discussion until a shared response is reached or to adopt a majority-supported answer within time constraints. Although unresolved disagreement may occur in some instances, extended discussion increases exposure to alternative reasoning and supports identification of errors, even when full consensus is not immediately achieved (Knierim et al. 2015; Leight et al. 2012).

For instructors seeking to incorporate active learning strategies into existing assessment structures, consensus-based two-stage examinations represent a practical and efficient approach

to enhancing student engagement without requiring substantial changes to course design. The group stage can be implemented within existing assessment periods, allowing instructors to integrate collaborative learning while maintaining individual accountability through the individual-stage component (Bruno et al. 2017; Levy et al. 2018).

Group sizes of three to four students are appropriate for facilitating discussion while ensuring that all group members participate meaningfully (Oakley et al. 2004). Requiring a single shared response during the group stage encourages explanation, evaluation of alternative answers, and resolution of disagreement, processes that support conceptual understanding and error correction (Chi and Wylie 2014; Driessen et al. 2020). Maintaining a weighting structure that prioritizes individual performance while incorporating a smaller group component can further support student confidence and reduce concerns regarding grading fairness (Gilley and Clarkston 2014).

Although non-consensus formats may preserve individual accountability, the lower collaborative benefit associated with this format suggests reduced effectiveness in promoting interaction and knowledge construction (Leight et al. 2012; Vogler and Robinson 2016). Use of non-consensus formats may therefore be limited to contexts where logistical constraints or specific instructional objectives necessitate individual submission following discussion.

At the institutional level, adoption of consensus-based two-stage examination formats may contribute to improved student engagement, more positive assessment experiences, and enhanced collaborative learning in introductory courses (Fournier et al. 2017; Khong and Tanner 2020; Rempel et al. 2023). Integration of structured collaborative assessment strategies provides an accessible entry point for instructors seeking to expand active learning practices while supporting student performance and classroom interaction.

Implications for assessment design

Findings have implications for assessment design in courses using two-stage examinations, where structured implementation influences student outcomes, participation, and attendance (Knierim et al. 2015), and may contribute to reducing achievement gaps in STEM education (Bruno et al. 2017). Greater collaborative benefit under the consensus format indicates that requiring agreement enhances the effectiveness of group-stage interaction when the objective is to maximize performance gains. Consensus-based formats should therefore be preferred in most instructional contexts, particularly when the goal is to promote conceptual understanding through discussion.

Non-consensus formats may still be appropriate in contexts where emphasis is placed on maintaining individual accountability for reasoning and answer selection, particularly when instructors aim to assess independent understanding alongside collaborative exposure. However, the absence of a structured agreement may limit the depth of interaction and reduce opportunities for error correction during the group stage. Selection of group-stage format should therefore be guided by instructional priorities (Gilley and Clarkston 2014).

Conclusions and future directions

In summary, group-stage format influences collaborative performance in two-stage examinations. The consensus format was associated with greater collaborative benefit relative to the non-consensus format, while individual-stage performance remained unchanged, indicating that differences in outcomes arise from interaction processes rather than baseline knowledge. Requirement for agreement in the consensus format promotes explanation, evaluation of

alternative responses, and resolution of disagreement, thereby enhancing group-stage performance.

Findings support the use of consensus-based formats in instructional contexts where the goal is to maximize performance gains and promote conceptual understanding through discussion. Non-consensus formats may remain appropriate in contexts where emphasis is placed on maintaining individual accountability for reasoning and answer selection; however, the absence of required agreement may limit the depth of interaction and opportunities for error correction. Format selection should therefore reflect instructional priorities, with consensus-based formats offering greater benefits for learning and performance in most cases.

Group-stage format represents a specific design feature that shapes collaborative assessment outcomes. Effects commonly attributed to collaborative testing are not uniform, as variation in assessment structure produces distinct performance patterns. Future research should directly examine interaction processes during the group stage, including the frequency and quality of explanation, justification of reasoning, and resolution of disagreement, to clarify how these factors contribute to collaborative benefit. Experimental manipulation of group composition and size would further determine how peer dynamics influence performance under consensus and non-consensus formats. In addition, longitudinal studies assessing retention of course material beyond the immediate assessment period would help determine whether collaborative gains translate into sustained learning. Replication across diverse course types, class sizes, and disciplinary contexts would also strengthen understanding of the generalizability of group-stage format effects.

Literature Cited

- Bandura, A. 1977. *Social Learning Theory*. Englewood Cliffs, N.J. : Prentice Hall.
- Brown, S., and Knight, P. 1995. Assessing Learners in Higher Education. *British Journal of Educational Studies* **43**(2): 234. doi:<https://doi.org/10.2307/3121949>.
- Bruno, B., Engels, J., Ito, G., Gillis-Davis, J., Dulai, H., Carter, G., Fletcher, C., and Böttjer-Wilson, D. 2017. Two-Stage Exams: A Powerful Tool for Reducing the Achievement Gap in Undergraduate Oceanography and Geology Classes. *Oceanography* **30**(2). Oceanography Society. doi:<https://doi.org/10.5670/oceanog.2017.241>.
- Callaghan, K., Kestin, G., Klales, A., McCarty, L., and Deslauriers, L. 2025. Active learning through flexible collaborative exams: Improving assessments across disciplines. *Active Learning in Higher Education* **0**(0). doi:<https://doi.org/10.1177/14697874251344293>.
- Chi, M.T.H., and Wylie, R. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* **49**(4): 219–243. doi:<https://doi.org/10.1080/00461520.2014.965823>.
- Cooke, J.E., Weir, L., and Clarkston, B. 2019. Retention following Two-Stage Collaborative Exams Depends on Timing and Student Performance. *CBE—Life Sciences Education* **18**(2): ar12. doi:<https://doi.org/10.1187/cbe.17-07-0137>.
- Cortright, R.N., Collins, H.L., Rodenbaugh, D.W., and DiCarlo, S.E. 2003. STUDENT RETENTION OF COURSE CONTENT IS IMPROVED BY COLLABORATIVE-GROUP TESTING. *Advances in Physiology Education* **27**(3): 102–108. doi:<https://doi.org/10.1152/advan.00041.2002>.

- Driessen, E.P., Knight, J.K., Smith, M.K., and Ballen, C.J. 2020. Demystifying the Meaning of Active Learning in Postsecondary Biology Education. *CBE—Life Sciences Education* **19**(4): ar52. doi:<https://doi.org/10.1187/cbe.20-04-0068>.
- Fournier, K.A., Couret, J., Ramsay, J.B., and Caulkins, J.L. 2017. Using collaborative two-stage examinations to address test anxiety in a large enrollment gateway course. *Anatomical Sciences Education* **10**(5): 409–422. doi:<https://doi.org/10.1002/ase.1677>.
- Freeman, S., Eddy, S.L., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H., and Wenderoth, M.P. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences* **111**(23): 8410–8415. doi:<https://doi.org/10.1073/pnas.1319030111>.
- Gilley, B.H., and Clarkston, B. 2014. Collaborative Testing: Evidence of Learning in a Controlled InClass Study of Undergraduate Students. *Journal of College Science Teaching* **43**(3): 83–91. National Science Teachers Association. doi:<https://doi.org/10.2307/43632038>.
- Hembree, R. 1988. Correlates, Causes, Effects, and Treatment of Test Anxiety. *Review of Educational Research* **58**(1): 47. doi:<https://doi.org/10.2307/1170348>.
- Ives, J., Van, J., Nutifafa Kwaku Sumah, and Stang, J.B. 2016. Examining student participation in two-phase collaborative exams through video analysis. 2017 *Physics Education Research Conference Proceedings*. doi:<https://doi.org/10.1119/perc.2016.pr.038>.
- Khong, M.L., and Tanner, J.A. 2020. A collaborative two-stage examination in biomedical sciences: Positive impact on feedback and peer collaboration. *Biochemistry and Molecular Biology Education* **49**(1). doi:<https://doi.org/10.1002/bmb.21392>.

- Kinnear, G. 2020. Two-Stage Collaborative Exams have Little Impact on Subsequent Exam Performance in Undergraduate Mathematics. *International Journal of Research in Undergraduate Mathematics Education* **7**(1). doi:<https://doi.org/10.1007/s40753-020-00121-w>.
- Knierim, K., Turner, H., and Davis, R.K. 2015. Two-Stage Exams Improve Student Learning in an Introductory Geology Course: Logistics, Attendance, and Grades. *Journal of Geoscience Education* **63**(2): 157–164. doi:<https://doi.org/10.5408/14-051.1>.
- Koretsky, M.D., McColley, C.J., Gugel, J.L., and Ekstedt, T.W. 2021. Aligning classroom assessment with engineering practice: A design-based research study of a two-stage exam with authentic assessment. *Journal of Engineering Education* **111**(1): 185–213. doi:<https://doi.org/10.1002/jee.20436>.
- Leight, H., Saunders, C., Calkins, R., and Withers, M. 2012. Collaborative Testing Improves Performance but Not Content Retention in a Large-Enrollment Introductory Biology Class. *CBE—Life Sciences Education* **11**(4): 392–401. doi:<https://doi.org/10.1187/cbe.12-04-0048>.
- Levy, D., Svoronos, T., and Klinger, M. 2018. Two-stage examinations: Can examinations be more formative experiences? *Active Learning in Higher Education* **24**(2): 146978741880166. doi:<https://doi.org/10.1177/1469787418801668>.
- Lindsley, J.E., Morton, D.A., Pippitt, K., Lamb, S., and Colbert-Getz, J.M. 2016. The Two-Stage Examination. *Academic Medicine* **91**(10): 1384–1387. doi:<https://doi.org/10.1097/acm.0000000000001185>.

- Meaders, C.L., and Vega, Y. 2022. Collaborative Two-Stage Exams Benefit Students in a Biology Laboratory Course. *Journal of Microbiology & Biology Education* **24**(1). doi:<https://doi.org/10.1128/jmbe.00138-22>.
- Millis, B.J., and Cottell, P.G. 1998. Cooperative learning for higher education faculty. Oryx Press, Phoenix, Az.
- Newton, G., Ritchie, K., Albabish, W., Gilley, B., Rajakaruna, R., and Kulak, V. 2019. Research and Teaching: Two-Stage (Collaborative) Testing in Science Teaching: Does It Improve Grades on Short-Answer Questions and Retention of Material? *Journal of College Science Teaching* **048**(04). doi:https://doi.org/10.2505/4/jcst19_048_04_64.
- Norman, G. 2010. Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education* **15**(5): 625–632. doi:<https://doi.org/10.1007/s10459-010-9222-y>.
- Oakley, B., Felder, R., Brent, R., and Elhajj, I. 2004. Turning Student Groups into Effective Teams. *Journal of Student Centered Learning* **2**(1): 9–34. Available from [https://engr.ncsu.edu/wp-content/uploads/drive/1ofGhdOciEwloA2zofffqkr7jG3SeKRq3/2004-Oakley-paper\(JSCL\).pdf](https://engr.ncsu.edu/wp-content/uploads/drive/1ofGhdOciEwloA2zofffqkr7jG3SeKRq3/2004-Oakley-paper(JSCL).pdf).
- Porter, S.R., and Whitcomb, M.E. 2005. Non-response in student surveys: The role of demographics, engagement and personality. *Research in Higher Education*, **46**(2): 127–152. doi:10.1007/s11162-004-1597-2.

- Putwain, D.W. 2007. Test anxiety in UK schoolchildren: Prevalence and demographic patterns. *British Journal of Educational Psychology* **77**(3): 579–593.
doi:<https://doi.org/10.1348/000709906x161704>.
- Rempel, B., McGinitie, E., and Dirks, M. 2023. The Influence of Two-Stage Collaborative Testing on Peer Relationships: A Study of First-Year University Student Perceptions. *The Canadian journal for the scholarship of teaching and learning* **14**(2). University of Western Ontario Libraries. doi:<https://doi.org/10.5206/cjsotlracea.2023.2.14252>.
- Rieger, G., and Heiner, C. 2014. Examinations That Support Collaborative Learning: The Students' Perspective. *Journal of College Science Teaching* **043**(04).
doi:https://doi.org/10.2505/4/jcst14_043_04_41.
- Shadish, W.R., Cook, T.D., and Campbell, D.T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin, Boston, Mass.
- Shaffer, J.F. 2020. Student Performance on and Perceptions of Collaborative Two-Stage Exams in a Material and Energy Balances Course. *Chemical Engineering Education* **54**(2).
doi:<https://doi.org/10.18260/2-1-370.660-115564>.
- Struyven, K., Dochy, F., and Janssens, S. 2005. Students' perceptions about evaluation and assessment in higher education: a review. *Assessment & Evaluation in Higher Education* **30**(4): 325–341. doi:<https://doi.org/10.1080/02602930500099102>.
- Terry, T., Volterman, K., Shiell, R., Melec Zeadin, Dunn, K., Melo, J.D., and Helli, P. 2017. Enhancing Two-stage Collaborative Exams by Incorporating Immediate Feedback. *The Western Conference on Science Education* **1**(1): 9. [accessed 18 September 2025].

Vogler, J.S., and Robinson, D.H. 2016. Team-Based Testing Improves Individual Learning. *The Journal of Experimental Education* **84**(4): 787–803.

doi:<https://doi.org/10.1080/00220973.2015.1134420>.

Vygotsky, L.S. 1978. *Mind in Society: Development of Higher Psychological Processes*. *Mind in Society: Development of Higher Psychological Processes* **1**(1).

doi:<https://doi.org/10.2307/j.ctvjf9vz4>.

Appendix

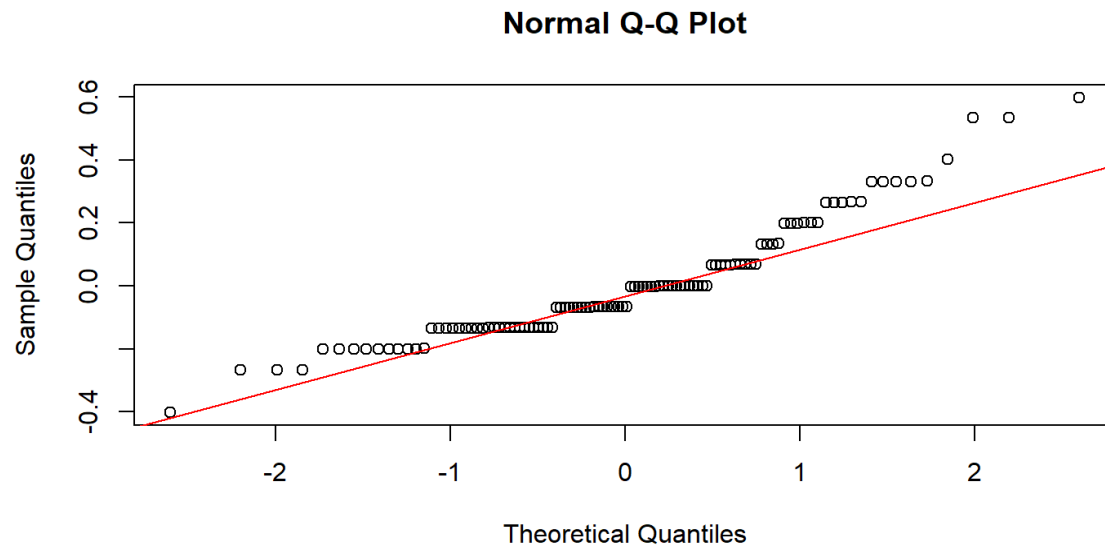
Appendix A. Student Perception Survey Questions

Scale: All questions were answered on a Likert scale: *Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree.*

Question	Item
1	I find the two-stage exam to be valuable.
2	I had an opportunity to practice my communication skills during the two-stage exam.
3	Two-stage exams improved my ability to communicate and collaborate with my peers.
4	I feel that my peers were respectful towards me during the two-stage exam.
5	The collaboration and communication of the two-stage examination enhanced my understanding of the course topics.
6	During the group portion of the two-stage examination, I explained/justified an answer to the group members.
7	The two-stage examination provided immediate feedback on my understanding of the course topics.
8	I feel less anxious about examinations after taking two-stage exams.
9	I feel motivated about my studies after the two-stage exam.
10	The two-stage exam improved my student experience.

Appendix B. Variables, operational definitions, and analyses

Variable	Type	Operationalization	Used in
Individual score	Continuous (%)	% Correct on individual stage	All directions
Group score	Continuous (%)	% Correct on group stage	All directions
Standardized performance gain	Continuous	$(\text{Group} - \text{Individual}) / (100 - \text{Individual})$; gain=0 when Individual=100	Dir. 1–2 (descriptive/sensitivity)
Group Benefit	Continuous	Group - Individual	Dir. 1-2-4
Format	Categorical	Consensus vs Non-consensus	Dir. 1–2–4
Test number	Categorical/Ordinal	Tests 1–4	Dir. 2
Sequence group	Categorical	X and Y (format sequence)	Dir. 3 (Design)
Stress	Ordinal (1–5)	Likert item	Dir. 4
Confidence	Ordinal (1–5)	Likert item	Dir. 4



Appendix C. Normal Q–Q plot of residuals for collaborative benefit scores.

Visual inspection indicated that data points generally followed the reference line, with only minor deviations at the tails. These results suggest that collaborative benefit was approximately normally distributed, supporting the use of parametric analyses.