Two-stage Examinations in STEM: A Narrative Literature Review

Timothy R. C. Lee[a], Matthew Pye[a], Osu Lilje[a], Hong Dao Nguyen[a], Samantha Hockey[a], Mark de Bruyn[a,b], and Francesca Trudy van den Berg[a]

Corresponding Author: Timothy R. C. Lee (t.lee@sydney.edu.au)
[a]School of Life and Environmental Sciences, The University of Sydney, Sydney NSW 2006, Australia
[b]School of Environment and Science, Griffith University, Brisbane QLD 4111, Australia

Keywords: examinations; feedback; group assessment; two-stage examinations

Abstract

Written, invigilated examinations are valued for their reliability, economy and academic integrity. Nevertheless, examinations are problematic. Final, summative examinations can disadvantage students who experience assessment anxiety, and students may not receive useable feedback. An alternative is the two-stage examination, where a traditional examination is followed by a group examination with similar questions. Students gain peer feedback on their examination performance, and can meaningfully apply this feedback. Use of this format in tertiary STEM education in universities has indicated that students prefer the format, although it has been little studied in Australia. Furthermore, its effects on reducing stress and fostering deeper learning are not well understood. The COVID-19 pandemic and switch to online learning has provided us with an opportunity to review our assessment practices and has led to a new willingness to test different examination formats. Here we provide a narrative review of the results of previous studies on two-stage examinations and, based on this and our experience teaching in large-cohort introductory biology courses at an Australian university, we propose a formula for employing them in this context.

Introduction

The advantages and disadvantages of examinations

Summative final examinations, an enduring feature of tertiary assessment, have both advantages and disadvantages. They have been thought to provide a reliable and valid measure of students' competence and understanding, while also providing an incentive for revision (van der Vleuten, 2000). Traditional final summative examinations (taken individually without communication, with no access to resources, limited time, and the same conditions and questions for all students) are also a relatively time-efficient and economical means of assessing many students. The usefulness of examinations rests tentatively on the assumption that performance in an examination correlates strongly with real command of the subject material (Kibble, 2017; Oppenheim, Jahoda, & James, 1967).

Examinations are not without disadvantages; they can provoke stress in students, (more commonly in female students than in male students), and this stress can have a negative effect on student performance (Chapell et al., 2005; DordiNejad et al., 2011). This makes examinations a potentially gender-biased assessment of student ability (Ballen, Salehi, & Cotner, 2017). Examinations might not provide long-term learning benefits for students; on the one hand, there is evidence to suggest that studying and/or sitting tests improves learning and retention compared to no test (Butler & Roediger, 2007; Vojdanoska, Cranney, & Newell, 2010), but on the other hand, students who do not receive feedback on examinations perform less well than students who do (Sato et al., 2018; Vojdanoska et al., 2010). Final examinations where feedback is not provided may therefore represent a lost opportunity for further learning (Fyfe, 2010).

An additional challenge for teachers is contract cheating in assessments, which is increasing in prevalence (Lancaster & Cotarlan, 2021; Newton, 2018). Some educators might feel that examinations are a more secure assessment than assignments, as students are relatively unlikely to take an exam for another student (Bretag et al., 2019). However, one Australian study has found that contract cheating occurs at a relatively high rate in exams, particularly in multiple-choice examinations, and that while teachers report detecting contract cheating most in assignments, students report engaging in these behaviours most in exams (Harper, Bretag, & Rundle, 2021). Online assessment poses particular challenges for detecting cheating as student behaviours are more difficult to observe (Dawson, 2020a).

The characteristics of effective feedback

Feedback is an essential component of any meaningful assessment, and students must have an opportunity to usefully apply the feedback they have received as the last step in the learning cycle (Al-Bashir, Kabir, & Rahman, 2016; Wilson, 2013). Despite this, feedback is consistently ranked low in student assessment of tertiary teaching. The Australian Quality Indicators for Learning and Teaching (QILT) national survey found that the question "*Teachers commented on your work in ways that help you learn*" was scored the lowest out of all 11 Teaching Quality questions, every year from 2013 to 2020 (Social Research Centre, 2013-2020).

Feedback on final examinations might be provided through different means, including feedforward processes (Scoles, Huxham, & McArthur, 2013), for example through formative pre-examination testing (Hope & Polwart, 2012). Although feedback is often regarded as an important aspect contributing to student learning, in the context of feedback after final examinations, students often only receive a grade without any specific feedback on any 'gaps' in knowledge or skill they displayed (Fyfe, 2010; Scoles et al., 2013).

Providing feedback on a final examination is challenging; to foster the best learning outcomes, this feedback needs to be actionable in a future assessment (Al-Bashir et al., 2016; Boud & Molloy, 2013), but this is difficult for assessments that occur at the very end of a semester's teaching cycle. Even if detailed feedback were provided on examination performance after the examination period, students may not find it useful. In a study by Fyfe (2010), individualized feedback was emailed to students four weeks after their final examination. Students were appreciative of feedback that was provided but felt that the feedback was not as useful as it could be because they could not recall the answers they had given to the questions. Immediate feedback can be provided in online examinations; with learning benefits increasing with greater amounts of detail (Wojcikowski & Kirk, 2013).

The two-stage examination: Addressing some deficiencies of the traditional examination method

A two-stage examination is a form of assessment combining a traditional written examination with a collaborative group-work component. The following pattern for running a two-stage examination from Wieman, Rieger, and Heiner (2014) is typical of the format:

1) Students undertake and submit an individual examination.
2) Students assemble into groups of 3-4, and each group is given a new paper (the group paper) to complete. Students communicate within their group and agree on answers to the questions on this paper together. Each group submits one paper.
3) The score each student receives is mostly from the individual component, and partly from the group component.

In this way, the evaluative properties of the traditional examination are retained, but with the addition of timely peer-feedback on performance (Wieman, 2017; Wieman et al., 2014). Students can improve their understanding by receiving immediate feedback from their peers, with minimal intervention by teachers (Cartney, 2010; Heller, Keith, & Anderson, 1992). Collaborative examinations provide an immediate opportunity for students to close the gap between what they know and what they should know, and encourages communication with peers about learning - all important aspects of good feedback (Cartney, 2010; Nicol & Macfarlane-Dick, 2006). Collaborative tests may also reduce the anxiety and stress associated with testing (Zipp, 2007). In a study by Guo and Li (2016), students reported that collaborative testing provided a more relaxed testing environment, and students who had previously been tested collaboratively performed better on the final examination than those who had sat traditional tests. For these reasons, two-stage examinations may alleviate the feedback and anxiety problems of traditional examinations, while retaining their advantages in the individual component.

COVID-19 disruption and a new approach to exams

The traditional examination format was disrupted globally in 2020, when the COVID-19 pandemic necessitated a new approach to prevent the transmission of the virus in examination halls. Examinations were cancelled, replaced with different assessments, or conducted online (e.g., Arenas, Calsamiglia, & Loviglio, 2021; Bhute, Campbell, Kogelbauer, Shah, & Brechtelsbauer, 2020; Clark, Callam, Paul, Stoltzfus, & Turner, 2020). There are several choices that can be made with online examinations: supervised or unsupervised (Dawson, 2020b; Hollister & Berenson, 2009), closed-book or open-book/open-web (Durning et al., 2016; Gehringer & Peddycord, 2013), and of varying duration (Petrović, Vitas, & Pale, 2017; Williams & Wong, 2009). Students generally prefer online to in-person examinations, for reasons including easier composition, being more relevant to the modern world, and convenience, but have concerns about academic integrity (reviewed in Butler-Henderson & Crawford, 2020). If two-stage examinations are to be implemented in an education landscape transformed by the pandemic, then these factors must be considered when planning to implement these examinations online.

In this paper, we provide a narrative literature review on two-stage examinations and identify trends in this research area. We assessed these studies to answer the following research questions:

- What are the parameters for two-stage examination setting (proportion of time and marks for group component, formation method for groups, type of questions, similarity of questions between individual and group component)?
- Is there an increase in group marks compared to individual marks?
- Do students respond positively to the format?
- Does the format improve retention or comprehension of the material in subsequent assessment, or comparing between cohorts?
- Do students report a reduction in stress or anxiety with exams in this format?

The authors of this narrative literature review have taught in large-cohort undergraduate biology units of study, and we are familiar with the logistical challenges of delivering these units. We use the results of our narrative review to suggest a practical format for implementing this type of examination in large-cohort courses similar to our teaching context. We also provide suggestions for facilitating these examinations in an online format, a widespread response to the COVID-19 pandemic.

Materials and Methods

The literature search was conducted on the 28th of June 2022 on the following platforms using the following terms:

ProQuest Education: "Two-stage exam*", all fields, scholarly journal, 24 results
"Two-stage cooperative test*", all fields, scholarly journal, 15 results
"Two-stage testing", all fields, scholarly journal, 70 results
"Pyramid exam*", all fields, scholarly journal, 21 results

Web of Science: "Two-stage exam*", all fields, 34 results
"Two-stage cooperative test*", all fields, 1 result
"Two-stage testing", all fields, 80 results
"Pyramid exam*", all fields, scholarly journal, 5 results

Semantic Scholar: "Two-stage exam*", 47 results
"Two-stage cooperative test*", 2 results
"Pyramid exam*", 3 results

We include only studies where a two-stage examination was implemented and reported on. Excluded here are studies where group components and individual components were separate, take home or open book examinations, and studies where the group component preceded the individual component. Studies from non-tertiary contexts, or non-STEM contexts, were also excluded. From this search, 31 studies were found that satisfied our criteria, and eight further studies were identified for inclusion from the reference lists of those 31 studies, making a total of 39 included studies.

Results

A compilation of several two-stage examination trials in STEM subjects from 1996 to 2022 is shown in Table 1 (see Appendix, page 84). Cohort sizes ranged from 11 to 679 students. The weighting of the group component (where stated) ranged from 0 to 50%, and the proportion of time for the group component (where stated) ranged from 20% of the test time to as long as needed. Group selection varied, from student selected to instructor selected or random. Question type varied, with multiple-choice and short answer questions most commonly employed, and true/false questions, essay questions and other question types employed more rarely. In most cases questions on the individual test reappeared on the group test. Where stated, students nearly always got higher marks on the group portion of the examination than on the individual portion, and students usually reacted positively to the format. Results concerning whether two-stage examinations reduce anxiety or stress were reported only in a minority of studies, and these results were mixed. Results about whether the two-stage exam format improves understanding compared to traditional testing, (either in a subsequent assessment or comparing cohorts), are mixed.

Discussion

The result that group component marks were almost always found to be higher than individual component marks is not surprising, (we might expect that multiple people working on a

problem will yield correct solutions more often than one person working alone). The very positive response from students was surprising, as were the mixed results related to improved learning.

Increase of group marks over individual: Given that some students are concerned that group work might lower their mark (Miller & James, 2019; Nordberg, 2008), this result could be reassuring. It also implies that students do gain immediate and useful feedback on their performance, as they can correct mistakes made in the individual part. This result is strikingly illustrated by Fengler and Ostafichuk (2015), who found that in a two-stage examination where the individual and group components were identical, 95% of individual student test scores were exceeded by the *lowest* group-test score.

Wieman et al. (2014) found that the result of the group section was usually better than the individual results of even the highest performing students. Wieman et al. (2014) suggested using the same questions in the group component as the individual component (perhaps with added questions), because it helped in providing targeted feedback on performance, and provided a starting point for discussion among the students (because they had all thought about the questions and had committed to an answer). Martin (2018) found a particularly interesting result in a study where identical questions were presented in the group stage as in the individual stage. In this study, where all students in a group got a particular question wrong, the group collectively got that question right in the group stage 57% of the time. A similar study by Kinnear (2021) found this occurred 17% and 29% of the time in two different settings. These results highlight the power of collaboration to correct misconceptions.

Student response: Although the response to two-stage examinations was generally positive in the studies reviewed here, students do not always respond positively to group work; the main complaint being that not all group members contribute equally, and that the marks of some students will be constrained by others (Nordberg, 2008). Yu, Tsiknis, and Allen (2010) reported that, overall, high-achieving students were more negative about two-stage examinations than low-achieving students. This problem could be addressed in two ways: firstly, by explaining early to students the theory and practice of the two-stage examination so that they understand its benefits (including the fact that research shows that group marks are rarely lower than individual marks), and secondly by setting the marking so that a student's mark cannot reduce as a consequence of the group part of the examination, because their mark is 'pegged' to the individual component (Bruno et al., 2017; Wieman et al., 2014). Low-performing students tend to benefit more from two-stage examinations than high- performing students, although both usually benefit (Bruno et al., 2017; Eaton, 2009; Giuliodori, Lujan, & DiCarlo, 2008). By maintaining a higher weighting for the individual part of the two-stage examination, high-performing students should not be adversely impacted.

Effect on improved learning: Cooke, Weir, and Clarkston (2019) made the point that comparing two-stage examination studies on whether they improve retention of material covered is difficult, as studies vary widely in their subject matter, implementation, and time span over which retention is tested. The studies included in this narrative review found benefits of two-stage examinations varied between the short and long term (Cooke et al., 2019; Vogler & Robinson, 2016) or between different cohorts of students (Cao & Porter, 2017) or between questions of different difficulty levels (Deng & Luo, 2018). Further studies in a variety of settings will be helpful in further elucidating the effects of two-stage examinations on learning.

Stress and anxiety: Although most of the studies included in this paper did not survey students

about stress or anxiety, those that did generally found that in two-stage examinations, students experienced less stress or anxiety than in traditional examinations. Where negative opinions were expressed, worry about giving incorrect opinions (Rempel, Dirks, & McGintie, 2021) or 'letting down the team' (Bentley, Attardi, Faul, Melo, & Palmer, 2021), frustration with group members (Rempel et al., 2021) and exhaustion (Bentley et al., 2021) were mentioned by students as factors that increased stress or anxiety with the format.

Employing two-stage examinations in large undergraduate courses

Although opinions differ on what size would be considered a 'large' class, by both staff and students, one definition is 240+ students (Cash, Letargo, Graether, & Jacobs, 2017). Although large cohorts require more resources to provide timely feedback, the use of two-stage examinations in large classes enables students to obtain feedback without the associated administration hours required by staff to provide feedback comments on an additional assessment. In the studies included in our narrative review, approaches to two-stage examinations in large-cohort units were diverse, although using multiple-choice questions was more common, with this question format (or similar) used in all studies with more than 240 students in a single cohort or section.

None of the two-stage examinations surveyed in our narrative literature review focused on conducting two-stage examinations online. Recently, recommendations have been provided by universities (Sutter, 2019; University of Guelph Office of Teaching and Learning, ND) and teachers have reported on their practice (Barshay, 2020; Stewart & Wickenden, 2020; Truchan, 2020) of conducting two-stage examinations online. Two-stage examinations can be conducted online either synchronously or asynchronously (University of Guelph Office of Teaching and Learning, ND); students could complete the group component synchronously in online video chat, or they could be given a set time to organise a meeting themselves to work on the questions (Barshay, 2020). These preliminary reports indicate that running two-stage examinations online is feasible, even synchronously for courses with large cohorts (Stewart & Wickenden, 2020).

The authors of this study have been involved in the delivery of large undergraduate biology courses (cohorts of 600-1750 students); based on our experiences teaching in those courses and this narrative review of the literature, we propose the following suggestions and considerations for implementing two-stage examinations in a large undergraduate introductory biology course:

1) Instructors should allocate groups to increase diversity. We feel the method described in Shaffer (2020) is a useful model; this method ensures a mixture of genders and assessment performance in each group, as well as encouraging students to work together prior to the exam to establish rapport.

2) Instructors should provide alternative group stage implementation for students who miss the examination. One solution is to have the student sit only an individual stage of the examination (e.g., Bruno et al., 2017), although the student may lose the benefit of feedback and the pooling of knowledge of several students (e.g., Khong & Tanner, 2021). Instead, students who miss the group component could be awarded the group mark from their assigned group, even though the student was not actually present. This would help maintain grade equity for those students, although the feedback benefit of the two-stage examination is still lost. Students with academic adjustments for whom group work would be disadvantageous could similarly be able to opt-out of the group component, but still receive the group mark from their assigned group. For students who

have extra time for the examination, we suggest the approach of Shepherd (2018) where these students start the individual component earlier, so that all students finish the individual component at the same time and can participate in the group component.

3) Questions should be multiple-choice rather than written answers. There is evidence to suggest that students engage in complex discussions around answers for even multiple-choice questions in two-stage examinations (Kinnear, 2021; Rieger & Heiner, 2014). Marking of multiple-choice type assessments is also well-suited to large classes; the majority of large-cohort studies surveyed here used this type (e.g., Eaton, 2009; Fournier, Couret, Ramsay, & Caulkins, 2017; McCurdy, Volterman, Shiell, Zeadin, & Dunn, 2017; Yuretich, Khan, Leckie, & Clement, 2001).

4) Ample time should be allocated. Students should not be time-stressed during the group stage of the examination, so as to maximize the benefit of the group discussion (Bentley et al., 2021). The approach of Martin (2018) to give students as much time as they need was applied with a large cohort; this is a useful model and is compatible with an asynchronous assessment format.

5) Marks should be 'pegged' so that students performing better in the individual component than the group component have their individual mark count as their group mark (e.g., Knierim, Turner, & Davis, 2015; Yuretich et al., 2001). This prevents a student being disadvantaged if they are unable to steer a group consensus to the correct answer.

6) For online two-stage exams, we suggest an asynchronous format is most practical for large cohorts (University of Guelph Office of Teaching and Learning, ND), providing 24 hours for groups to complete this section of the exam, as in the case of Barshay (2020).This would avoid technical issues that can arise with students needing to sort into groups in a videochat with the time pressures of an exam, an issue flagged by Stewart and Wickenden (2020). Contract cheating can be addressed by making sure students and staff are aware of the academic integrity requirements of the assessment (Dawson, 2020a; Spruin, 2022), monitoring for examination content on file-sharing websites (Hill, Mason, & Dunn, 2021), and perhaps blocking access to these websites (Spruin, 2022).

Conclusion

Research surveyed here indicates that two-stage examinations are positively received by students, and provide feedback to address misconceptions from the individual examination. Further research is needed to determine under what circumstances they help to reduce anxiety, and whether they can improve understanding of concepts in subsequent assessment. The benefits of two-stage examinations may be especially relevant in large classes where feedback is difficult to provide at scale. With the COVID-19 pandemic prompting a reconsideration of examinations generally, we encourage educators to consider this format as an option for their final examinations.

References

Al-Bashir, M. M., Kabir, M. R., & Rahman, I. (2016). The value and effectiveness of feedback in improving students' learning and professionalizing teaching in higher education. *Journal of Education and Practice, 7*(16), 38-41.

Arenas, A., Calsamiglia, C., & Loviglio, A. (2021). What is at stake without high-stakes exams? Students' evaluation and admission to college at the time of COVID-19. *Economics of education review, 83*, 102143. https://doi.org/10.1016/j.econedurev.2021.102143

Ballen, C. J., Salehi, S., & Cotner, S. (2017). Exams disadvantage women in introductory biology. *PloS one, 12*(10), e0186419-e0186419. https://doi.org/10.1371/journal.pone.0186419

Barshay, J. (2020). Proof Points: Improving college exams during remote learning. Retrieved from https://hechingerreport.org/proof-points-improving-on-college-exams-during-remote-learning/

Bentley, D. C., Attardi, S. M., Faul, J., Melo, V., & Palmer, C. (2021). Two-stage collaborative group testing does not improve retention of anatomy among students studying medical radiation technology. *Journal of Medical Imaging and Radiation Sciences, 52*, S96-S109.

Bhute, V. J., Campbell, J., Kogelbauer, A., Shah, U. V., & Brechtelsbauer, C. (2020). Moving to timed remote assessments: The impact of COVID-19 on year end exams in chemical engineering at Imperial College London. *Journal of Chemical Education, 97*(9), 2760-2767. https://doi.org/10.1021/acs.jchemed.0c00617

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education, 38*(6), 698-712.

Bretag, T., Harper, R., Burton, M., Ellis, C., Newton, P., Rozenberg, P., Saddiqui, S., & van Haeringen, K. (2019). Contract cheating: a survey of Australian university students. *Studies in Higher Education, 44*(11), 1837-1856. https://doi.org/10.1080/03075079.2018.1462788

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

Butler-Henderson, K., & Crawford, J. (2020). A systematic review of online examinations: A pedagogical innovation for scalable authentication and integrity. *Computers and education, 159*, 104024-104024. https://doi.org/10.1016/j.compedu.2020.104024

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4-5), 514-527. https://doi.org/10.1080/09541440701326097

Cao, Y., & Porter, L. (2017). *Impact of performance level and group composition on student learning during collaborative exams*. Paper presented at the Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education, Bologna, Italy.

Cartney, P. (2010). Exploring the use of peer assessment as a vehicle for closing the gap between feedback given and feedback used. *Assessment and evaluation in higher education, 35*(5), 551-564. https://doi.org/10.1080/02602931003632381

Cash, C. B., Letargo, J., Graether, S. P., & Jacobs, S. R. (2017). An analysis of the perceptions and resources of large university classes. *CBE life sciences education, 16*(2), ar33. https://doi.org/10.1187/cbe.16-01-0004

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97*(2), 268-274.

Chen, S., & Kinniburgh, S. (2019). A controlled experiment on two-stage exams in an introductory statistics course. *The international journal of assessment and evaluation, 26*(2), 1-12. https://doi.org/10.18848/2327-7920/CGP/v26i02/1-12

Clark, T. M., Callam, C. S., Paul, N. M., Stoltzfus, M. W., & Turner, D. (2020). Testing in the time of COVID-19: A sudden transition to unproctored online exams. *Journal of Chemical Education, 97*(9), 3413-3417. https://doi.org/10.1021/acs.jchemed.0c00546

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

Cortright, R. N., Collins, H. L., Rodenbaugh, D. W., & DiCarlo, S. E. (2003). Student retention of course content is improved by collaborative-group testing. *Advances in Physiology Education, 27*(3), 102-108. https://doi.org/10.1152/advan.00041.2002

Dawson, P. (2020a). The prevention of contract cheating in an online environment. *How to design assessments to prevent contract cheating.* Retrieved from https://www.teqsa.gov.au/sites/default/files/prevention-contract-cheating-in-online-environment-web.pdf?v=1587691121

Dawson, P. (2020b). Strategies for using online invigilated exams. *Academic integrity experts advice hub.* Retrieved from https://www.teqsa.gov.au/sites/default/files/strategies-for-using-online-invigilated-exams.pdf?v=1603758032

Deng, Q., & Luo, X. (2018). *PipE²: An innovative pipelining design for collaborative two-stage exams*. Paper presented at the Proceedings of the 19th Annual SIG Conference on Information Technology Education, Fort Lauderdale, Florida, USA. https://doi-org.ezproxy.library.sydney.edu.au/10.1145/3241815.3241850

DordiNejad, F. G., Hakimi, H., Ashouri, M., Dehghani, M., Zeinali, Z., Daghighi, M. S., & Bahrami, N. (2011). On the relationship between test anxiety and academic performance. *Procedia - Social and Behavioral Sciences, 15*, 3774-3778. https://doi.org/10.1016/j.sbspro.2011.04.372

Durning, S. J., Dong, T., Ratcliffe, T., Schuwirth, L., Artino, A. R., Jr., Boulet, J. R., & Eva, K. (2016). Comparing open-book and closed-book examinations: A systematic review. *Academic medicine, 91*(4), 583-

599. https://doi.org/10.1097/ACM.0000000000000977
Eaton, T. T. (2009). Engaging students and evaluating learning progress using collaborative exams in introductory courses. *Journal of Geoscience Education, 57*(2), 113-120.
Fengler, M., & Ostafichuk, P. M. (2015). *Success with two-stage exams in mechanical engineering.* Paper presented at the Canadian Engineering Education Association Conference, McMaster University.
Ford, D. (2019). A three-year retrospective analysis comparing student retention of human physiology concepts in flipped, lecture, and two-stage cooperative testing classrooms. *Journal of College Science Teaching, 49*(1), 59-63.
Fournier, K. A., Couret, J., Ramsay, J. B., & 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. https://doi.org/10.1002/ase.1677
Fyfe, G. (2010). *The final examination: a squandered opportunity for feedback to students or a poor use of time?* Paper presented at the ATN Assessment Conference 2010: Assessment, Sustainability, Diversity and Innovation, University of Technology, Sydney.
Gehringer, E. F., & Peddycord, B. W. (2013). Experience with online and open-web exams. *Journal of Instructional Research, 2*, 10-18.
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.
Giuliodori, M. J., Lujan, H. L., & DiCarlo, S. E. (2008). Collaborative group testing benefits high- and low-performing students. *Advances in Physiology Education, 32*, 274-278. https://doi.org/10.1152/advan.00101.2007
Green, R. A., Cates, T., White, L., & Farchione, D. (2016). Do collaborative practical tests encourage student-centered active learning of gross anatomy? *Anatomical Sciences Education, 9*(3), 231-237. https://doi.org/10.1002/ase.1564
Guo, Y., & Li, E. (2016). Collaborative testing in practical laboratories: an effective teaching-learning method in histology. *Journal of Vetinary Medical Education, 43*(1), 9-12. https://doi.org/10.3138/jvme.1114-108R2
Harper, R., Bretag, T., & Rundle, K. (2021). Detecting contract cheating: examining the role of assessment type. *Higher Education Research & Development, 40*(2), 263-278. https://doi.org/10.1080/07294360.2020.1724899
Heller, P., Keith, R., & Anderson, S. (1992). Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving. *American Journal of Physics, 60*(7), 627-636. https://doi.org/10.1119/1.17117
Hill, G., Mason, J., & Dunn, A. (2021). Contract cheating: an increasing challenge for global academic community arising from COVID-19. *Research and Practice in Technology Enhanced Learning, 16*(1), 24. https://doi.org/10.1186/s41039-021-00166-8
Hollister, K. K., & Berenson, M. L. (2009). Proctored versus unproctored online exams: Studying the impact of exam environment on student performance. *Decision sciences journal of innovative education, 7*(1), 271-294. https://doi.org/10.1111/j.1540-4609.2008.00220.x
Hope, S. A., & Polwart, A. (2012). Engagement with online pre-exam formative tests improves exam performance and feedback satisfaction. *Bioscience Education, 20*(1), 37-52. https://doi.org/10.11120/beej.2012.20000037
Ives, J. (2014). *Measuring the learning from two-stage collaborative group exams*. Paper presented at the Physics Education Research Conference, Minneapolis, USA.
Khong, M. L., & Tanner, J. A. (2021). A collaborative two-stage examination in biomedical sciences: Positive impact on feedback and peer collaboration. *Biochemistry and Molecular Biology Education, 49*(1), 69-79. https://doi.org/10.1002/bmb.21392
Kibble, J. D. (2017). Best practices in summative assessment. *Advances in Physiology Education, 41*(1), 110-119. https://doi.org/10.1152/advan.00116.2016
Kinnear, G. (2021). Two-stage collaborative exams have little impact on subsequent exam performance in undergraduate mathematics. *International Journal of Research in Undergraduate Mathematics Education, 7*(1), 33-60. https://doi.org/10.1007/s40753-020-00121-w
Knierim, K., Turner, H., & 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. https://doi.org/10.5408/14-051.1
Lancaster, T., & Cotarlan, C. (2021). Contract cheating by STEM students through a file sharing website: a Covid-19 pandemic perspective. *International Journal for Educational Integrity, 17*(1). https://doi.org/10.1007/s40979-021-00070-0
Leight, H., Saunders, C., Calkins, R., & 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. https://doi.org/10.1187/cbe.12-04-0048

Levy, D., Svoronos, T., & Klinger, M. (2018). Two-stage examinations: Can examinations be more formative experiences? *Active Learning in Higher Education, 0*(0), 1469787418801668. https://doi.org/10.1177/1469787418801668

Lindsley, J. E., Morton, D. A., Pippitt, K., Lamb, S., & Colbert-Getz, J. M. (2016). The two-stage examination: A method to assess individual competence and collaborative problem solving in medical students. *Academic medicine, 91*(10), 1384-1387. https://doi.org/10.1097/acm.0000000000001185

Macpherson, G. L., Lee, Y.-J., & Steeples, D. (2011). Group-examination improves learning for low-achieving students. *Journal of Geoscience Education, 59*(1), 41-45.

Martin, A. P. (2018). A quantitative framework for the analysis of two-stage exams. *International Journal of Higher Education, 7*(4), 33-54.

McCurdy, T. R., Volterman, K., Shiell, R., Zeadin, M., & Dunn, K. (2017). Enhancing two-stage collaborative exams by incorporating immediate feedback. *Discussions on University Science Teaching: Proceedings of the Western Conference on Science Education, 1*(1), Article 9.

Miller, S. T., & James, C. R. (2019). What is the impact of collaborative exams on learning and attitudes in introductory astronomy classes? *Journal of Astronomy & Earth Sciences Education, 6*(1), 17-30.

Newton, G., Rajakaruna, R., Kulak, V., Albabish, W., Gilley, B. H., & Ritchie, K. (2019). Two-stage (collaborative) testing in science teaching: Does it improve grades on short-answer questions and retention of material? *Journal of College Science Teaching, 48*(4), 64-73.

Newton, P. M. (2018). How common is commercial contract cheating in higher education and is it increasing? A systematic review. *Frontiers in Education, 3*. https://doi.org/10.3389/feduc.2018.00067

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218. https://doi.org/10.1080/03075070600572090

Nordberg, D. (2008). Group projects: more learning? Less fair? A conundrum in assessing postgraduate business education. *Assessment & Evaluation in Higher Education, 33*(5), 481-492. https://doi.org/10.1080/02602930701698835

Oppenheim, A. N., Jahoda, M., & James, R. L. (1967). Assumptions underlying the use of university examinations. *Higher Education Quaterly, 21*(3), 341-351.

Petrović, J., Vitas, D., & Pale, P. (2017). *Experiences with supervised vs. unsupervised online knowledge assessments in formal education.* Paper presented at the Proceedings of ELMAR-2017, Zadar, Croatia.

Plotnick, R. E., Varelas, M., & Fan, Q. (2009). An integrated earth science, astronomy, and physics course for elementary education majors. *Journal of Geoscience Education, 57*(2), 152-158.

Rao, S. P., Collins, H. L., & DiCarlo, S. E. (2002). Collaborative testing enhances student learning. *Advances in Physiology Education, 26*(1), 37-41. https://doi.org/10.1152/advan.00032.2001

Rempel, B. P., Dirks, M. B., & McGintie, E. G. (2021). Two-stage testing reduces student-perceived exam anxiety in introductory chemistry. *Journal of Chemical Education, 98*, 2527-2535.

Rieger, G. W., & Heiner, C. E. (2014). Examinations that support collaborative learning: the students' perspective. *Journal of College Science Teaching, 43*(4), 41-47.

Sato, B. K., Dinh-Dang, D., Cruz-Hinojoza, E., Denaro, K., Hill, C. F. C., & Williams, A. (2018). The impact of instructor exam feedback on student understanding in a large-enrollment biology course. *BioScience, 68*(8), 601-611. https://doi.org/10.1093/biosci/biy058

Scoles, J., Huxham, M., & McArthur, J. (2013). No longer exempt from good practice: using exemplars to close the feedback gap for exams. *Assessment & Evaluation in Higher Education, 38*(6), 631-645. https://doi.org/10.1080/02602938.2012.674485

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), 52-58.

Shepherd, G. G. (2018). Experience with implementing pyramid examinations in an elective pharmacy course. *Currents in Pharmacy Teaching & Learning, 10*(12), 1631-1635. https://doi.org/10.1016/j.cptl.2018.09.003

Social Research Centre. (2013-2020). Quality Indicators for Learning and Teaching (QILT) student experience surveys. Retrieved from https://www.qilt.edu.au/qilt-surveys/student-experience

Spruin, E. (2022). Mitigating the risk of contract cheating in UK higher education: A multi-level solution. *Journal of Higher Education Policy and Leadership Studies, 3*(1), 109-120. https://doi.org/10.52547/johepal.3.1.109

Stearns, S. A. (1996). Collaborative exams as learning tools. *College Teaching, 44*(3), 111-112.

Stewart, J. J., & Wickenden, J. (2020). *Successful implementation of two-stage exams in a remote-teaching environment*. Paper presented at the Online Teaching Series, The University of British Columbia. https://admin.video.ubc.ca/html5/html5lib/v2.86.1/mwEmbedFrame.php/p/113/uiconf_id/23449170/entry_id/0_owm4opqu?wid=_113&iframeembed=true&playerId=kaltura_player&entry_id=0_owm4opqu&flashvars

Sutter, A. M. (2019). Advanced teaching technique: Two-stage exams in class or online. *Agronomy and Horticulture Online Guides, Summer 2019*.

Truchan, H. (2020) *Q&A on collaborative or two-stage exams/Interviewer: R. Goc*. Northwestern University.
University of Guelph Office of Teaching and Learning. (ND). Two-stage exams. Retrieved from https://otl.uoguelph.ca/two-stage-exams
van der Vleuten, C. (2000). Validity of final examinations in undergraduate medical training. *British Medical Journal, 321*(7270), 1217. https://doi.org/10.1136/bmj.321.7270.1217
Vogler, J. S., & Robinson, D. H. (2016). Teab-based testing improves individual learning. *The Journal of Experimental Education, 84*(4), 787-803.
Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied cognitive psychology, 24*(8), 1183-1195. https://doi.org/10.1002/acp.1630
Walker, J. D., & Robinson, D. H. (2022). Does two-stage testing promote long-term individual learning? *The Journal of Experimental Education*. https://doi.org/10.1080/00220973.2022.2059750
Weicker, K. (2020). *Teaching cooperative problem solving*. Paper presented at the Proceedings of the 4th European Conference on Software Engineering Education, Seeon/Bavaria, Germany.
Wieman, C. (2017). *Improving how universities teach science : Lessons from the Science Education Initiative*. Cambridge, MA: Harvard University Press.
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
Williams, J. B., & Wong, A. (2009). The efficacy of final examinations: A comparative study of closed-book, invigilated exams and open-book, open-web exams. *British Journal of Educational Technology, 40*(2), 227-236. https://doi.org/10.1111/j.1467-8535.2008.00929.x
Wilson, A. (2013). Feedback as a transformative tool. In K. Coleman & A. Flood (Eds.), *Marking Time: Leading and managing the develpment of assessment in higher education*. Champagne, Illinois: Common Ground.
Wojcikowski, K., & Kirk, L. (2013). Immediate detailed feedback to test-enhanced learning: an effective online educational tool. *Medical Teacher, 35*(11), 915-919. https://doi.org/10.3109/0142159x.2013.826793
Yu, B., Tsiknis, G., & Allen, M. (2010). *Turning exams into a learning experience*. Paper presented at the Special Interest Group on Computer Science Education, Milwaukee, USA.
Yuretich, R. F., Khan, S. A., Leckie, R. M., & Clement, J. J. (2001). Active-learning methods to improve student performance and scientific interest in a large introductory oceanography course. *Journal of Geoscience Education, 49*(2), 111-119. https://doi.org/10.5408/1089-9995-49.2.111
Zipp, J. F. (2007). Learning by exams: the impact of two-stage co-operative tests. *Teaching Sociology, 35*, 62-76.

Appendix

Table 1. Comparison of studies from 1996-2022 on the implementation of two-stage examinations in tertiary STEM courses

Reference	Subject Area	Did the questions in the group component appear in the individual component?	Question type	Proportion of time for group component (%)	Group component marks (%)	Group component higher mark?	Student Rating	Improved understanding or retention (either compared to other cohort, or in subsequent assessment)	Students generally reported a reduction in anxiety and/or stress?	How were groups chosen?	Number of Students (Number of students included in analysis, if applicable)	Notes
Stearns (1996)	Research Methods and Statistics	Yes	MCQ	50	~33	Yes	Not tested	Yes	Not tested	Student selected	8 and 25	Class sizes declined from 12 and 26 students respectively
Yuretich, Khan, Leckie and Clement (2001)	Introductory Oceanography	Some	MCQ	60	25	Yes	Positive	Yes	Not tested	Neighbours in the classroom	~600 in each cohort	Group mark not counted if it would lower a student's grade.
Rao, Collins and DiCarlo (2002)	Cardiovascular Physiology	Yes	Fill in the blanks; MCQ; SAQ; true/false	50	20	Yes	Positive	Not tested	Not tested	Instructor assigned	16	
Cortright, Collins, Rodenbaugh and DiCarlo (2003)	Exercise Physiology	Yes	MCQs; fill in the blanks; short answer essays	Not stated	Not stated	Yes	Positive	Yes	Not tested	Not stated	38	Groups were mostly pairs
Zipp (2007)	Introductory Sociology	Yes	MCQ	~72	Variable, ~13, based on a weighting system	Yes	Not tested	Yes	Not tested	Instructor selected to maximise diversity	122	Weighting system designed to not excessively benefit lower- performing students.
Giuliodori, Lujan and DiCarlo (2008)	Veterinary Physiology	Yes	MCQ	~43	0	Yes	Positive	Not tested	Yes	Students chose	65	Groups were pairs

Reference	Subject Area	Did the questions in the group component appear in the individual component?	Question type	Proportion of time for group component (%)	Group component marks (%)	Group component higher mark?	Student Rating	Improved understanding or retention (either compared to other cohort, or in subsequent assessment)	Students generally reported a reduction in anxiety and/or stress?	How were groups chosen?	Number of Students (Number of students included in analysis, if applicable)	Notes
Eaton (2009)	Two Introductory Geoscience courses	Yes for one course, no for the other	MCQ	~37	~33, in one course questions answered correctly individually and then incorrectly in a group not penalised	Yes	Not tested	Improvement in performance across multiple two-stage exams	Not tested	Alphabetically in one course, neighbours in the other.	400-500 and 50-100+	
Plotnick, Varelas and Fan (2009)	The Physical World	Yes	Pre-test included MCQ and constructed response	40	20	Yes	Positive	Not tested	Not tested	Random suggested	131 over three years	
Yu, Tsiknis and Allen (2010)	Two Computer Science courses	Yes	SAQ, programming code, computations	~38	Weighted average, score of group component pegged to individual if lower	Yes for at least some questions	Positive	Results were mixed	Not tested	Students chose	37 and 59	
Macpherson, Lee and Steeples (2011)	Earthquakes and Natural Disasters	Yes	Short to medium length essay	50	Weighted between whole exam and the subset of questions where they were the individual lead	Yes	Positive	Yes on the final exam, for students with the lowerst performance in individual pre-test	Not tested	Instructor assigned to maximise divesrity of major, and performance	19	
Leight, Saunders, Calkins, and Withers (2012)	Introductory Biology	Yes	MCQ, multiple correct, true/false and sequencing problems	50	Bonus mark scheme	Yes	Positive	No	Yes	Students chose	~250	
Gilley and Clarkston (2014)	Natural Disasters	Yes	MCQ	50	15	Yes	Not tested	Yes	Not tested	Students chose	98	Study involved individual retests as well as group retests.

Reference	Subject Area	Did the questions in the group component appear in the individual component?	Question type	Proportion of time for group component (%)	Group component marks (%)	Group component higher mark?	Student Rating	Improved understanding or retention (either compared to other cohort, or in subsequent assessment)	Students generally reported a reduction in anxiety and/or stress?	How were groups chosen?	Number of Students (Number of students included in analysis, if applicable)	Notes
Ives (2014)	Introductory Physics; Fluids, Waves and Energy	Some	MCQ	Not given	15, score of group component pegged to individual if lower	Not stated	Not tested	Results were mixed	Not tested	Students chose	679	
Rieger and Heiner (2014)	Introductory Physics	Some	MCQ and SAQ	~33	15	Yes	Positive	Not tested	Not tested	Either preformed, instructor assigned or students chose	178	
Fengler and Ostafichuk (2015)	Mechanical Design	Yes	MCQ	50	Not stated	Yes	Positive	Not tested	Not tested	Instructor selected to maximise diversity	750 over six years	
Knierim, Turner and Davis (2015)	Introductory Geology	Yes	MCQ or similar (true/false, matching)	Not stated	25, score of group component pegged to individual if lower	Yes	Not tested	Yes	Not tested	Not stated	~200	Also tested two-stage exams were the second stage is individual but open book.
Green, Cates, White, and Farchione (2016)	Anatomy	Yes	Practical tests, SAQ	~45	25	Yes	Not tested	No	Not tested	Not stated	207	
Lindsley, Morton, Pippitt, Lamb and Colbert- Getz (2016)	Foundational Sciences	Yes	MCQ	~37	10	Yes	Not tested	Yes for concepts that were originally answered incorrectly	Not tested	Randomly	104	Group-stage answers did not need to be identical across group members.
Vogler and Robinson (2016)	Undergraduate Educational Psychology	Yes	MCQ	>50	Not stated	Not stated	Positive	No in a followup test 2 weeks later, yes for a followup 2 months later	Only a minority of students found the two-stage format made them more nervous	Randomly, stratified by sex and year level	51 and 39	In the first cohort, students taking tests individually took them twice, in line with students in the two-stage testing treatment. Students were provided with feedback in the second test in both treatments.

Reference	Subject Area	Did the questions in the group component appear in the individual component?	Question type	Proportion of time for group component (%)	Group component marks (%)	Group component higher mark?	Student Rating	Improved understanding or retention (either compared to other cohort, or in subsequent assessment)	Students generally reported a reduction in anxiety and/or stress?	How were groups chosen?	Number of Students (Number of students included in analysis, if applicable)	Notes
Cao and Porter (2017)	Computer Science	Similar	Problem solving	~26	Not stated	Not stated	Not tested	Yes for mid-performing students, no for low and high performing students	Not tested	Groups varied in performance heterogeneity	278 (247) in two sections	More homogenous groups of students showed greater performance increases on the final exam than more heterogenous groups.
Fournier, Couret, Ramsey and Caulkins (2017)	Introductory Human Anatomy	Yes	MCQ	Not stated	15	Yes	Positive	No	In general, no.	Students chose	444	
Bruno, Engels, Ito, Gillis-Davis, Dulai, Carter, Fletcher and Böttjer-Wilson (2017)	Oceanography and Geology (five courses)	Yes	Variable, including MCQ and essay	Not stated	Multiple courses, ranged from 15 to 50	Yes	Anecdotally, positive	Not tested	Anecdotally, yes	Random (1 course), assigned to maximise diversity (1 course), students chose (3 courses)	289 in seven sections	In one course, group-stage answers did not need to be identical. In one course, students could consult notes.
McCurdy, Volterman, Shiell, Zeadin, Dunn, De Melo and Helli (2017)	Introductory Human Biochemistry	Yes	MCQ	40	15	Yes	Positive	Not tested (Comparison of the effect of immediate feedback on the group part of the exam was tested)	Yes	Randomly	399 (343)	
Deng and Luo (2018)	Introductory Computing	Yes	MCQ	~33	20	Not stated	Positive	Yes for moderate difficulty questions, no for easy and hard difficulty questions	Not tested	Students chose	127 (54)	

Reference	Subject Area	Did the questions in the group component appear in the individual component?	Question type	Proportion of time for group component (%)	Group component marks (%)	Group component higher mark?	Student Rating	Improved understanding or retention (either compared to other cohort, or in subsequent assessment)	Students generally reported a reduction in anxiety and/or stress?	How were groups chosen?	Number of Students (Number of students included in analysis, if applicable)	Notes
Levy, Svoronos and Klinger (2018)	Three courses in Statistics/ Econometrics	Yes	Not specified	~31	Up to 10, no penalty if group score was lower than individual	Yes	Positive	Not tested	Yes	Random, including one member from top 40% of class	899 in five cohorts	
Martin (2018)	Two courses: Introductory Statistics and Evolutionary Biology	Yes	MCQ	As long as needed	Not stated	Yes	Not tested	Not tested	Not tested	Students chose	753 in four cohorts	Group stage was conducted as a class, with all groups working on the same problem at the same time.
Shepherd (2018)	Clinical Toxicology	Yes, with additional explanations of answers for the group component	Fixed choice and constructed response	50	~33	Yes	Generally positive	Not tested	Not tested	Randomly at first test, subsequently assigned to include one each from the lowest and highest performance quartile	37	Students with adjustments for extra time started the individual component early, so that they could participate in the group component.
Cooke, Weir and Clarkston (2019)	Introductory Biology	Isomorphic in one exam, identical in the other.	SAQ	20, not including an individual retest component	Not stated	Not stated	Positive	Yes, long-term; no, short-term	Not tested	Students chose	158 (125)	
Ford (2019)	Human Physiology	Yes	Not stated	50	10	Yes	Not tested	Yes	Anecdotally, yes	Instructor selected to increase performance diversity	633 (225)	Two-stage exams had a more beneficial effect than flipped lectures on retention.
Miller and James (2019)	Introductory Astronomy	Yes	MCQ	~33	25	Yes	Positive	Yes	Not tested	Instructor selected for diverse majors and year levels and to exclude male-majority groups.	360 in four sections	

Reference	Subject Area	Did the questions in the group component appear in the individual component?	Question type	Proportion of time for group component (%)	Group component marks (%)	Group component higher mark?	Student Rating	Improved understanding or retention (either compared to other cohort, or in subsequent assessment)	Students generally reported a reduction in anxiety and/or stress?	How were groups chosen?	Number of Students (Number of students included in analysis, if applicable)	Notes
Newton, Rajakaruna, Kulak, Albabish, Gilley and Ritchie (2019)	Two courses: Biochemistry and Exercise Physiology	Yes	MCQ and long answer	~29	20	Yes	Not tested	Yes	Not tested	Students chose	64 (56) and 102 (94)	
Chen and Kinniburgh (2019)	Introductory Statistics	Yes	MCQ	~47	15	Yes	Positive	Yes	Yes	Randomly	68 in two sections	
Shaffer (2020)	Material and Energy Balances	Yes	SAQ with calculations	40	20	Yes	Positive	No	Not tested	Instructor selected to maintain gender balance, mixture of performance and preference for studying on the same days	34	
Weicker (2020)	Algorithms and Data Structures	No	SAQ	~33	~32	No	Anecdotally, positive	Not tested	Not tested	Randomly, but dynamically changing groups during the group component were allowed	89	
Bentley, Attardi, Faul, Melo and Palmer (2021)	Anatomy for Medical Radiation Sciences	Yes	MCQ	~33	25	Yes	Positive	No	Majority, yes	Randomly	97 (86) and 99 (81)	
Khong and Tanner (2021)	Essential Proteomics	Yes	SAQ	~33	15, score of group component pegged to individual if lower	Yes	Positive	Not tested	No	Existing assessment groups	11	

Reference	Subject Area	Did the questions in the group component appear in the individual component?	Question type	Proportion of time for group component (%)	Group component marks (%)	Group component higher mark?	Student Rating	Improved understanding or retention (either compared to other cohort, or in subsequent assessment)	Students generally reported a reduction in anxiety and/or stress?	How were groups chosen?	Number of Students (Number of students included in analysis, if applicable)	Notes
Kinnear (2021)	Three undergradute Mathematics units	Yes; Yes; Similar (SAQ converted to MCQ)	SAQ; MCQ; SAQ and MCQ	Not given; not given; ~38	Unweighted; 50; 30 (score of group component pegged to individual if lower)	Yes; Yes; Yes	Positive; not tested; mixed	Yes; inconclusive; no	Yes; not tested; no	Students chose; Instructor assigned; Students chose	47; 24; 301 (254)	
Rempel, Dirks and McGintie (2021)	General (Introductory) Chemistry	Yes	Not stated	~28-50	15-30 (score of group component pegged to individual if lower)	Yes	Positive	Not tested	Yes	Not stated	55 (49); 46 (39); 60 (41)	
Walker and Robinson (2022)	Undergraduat e Educational Psychology	Yes	MCQ; MCQ; SAQ and MCQ	Not stated	Not stated	Not stated	Not tested	No in either short or long term	Not tested	Randomly	49; 44; 208	Study designed as a folloup to Vogler and Robinson (2016)