Two-stage Examinations in STEM: A Narrative Literature Review

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

Corresponding Author: Timothy R. C. Lee (t.lee@sydney.edu.au)

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 usable 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 (Rempe et al., 2021) or ‘letting down the team’ (Bentley et al., 2021), frustration with group members (Rempe 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 et al., 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 videocall 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


## Appendix

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

<table>
<thead>
<tr>
<th>Reference</th>
<th>Subject Area</th>
<th>Did the questions in the group component appear in the individual component?</th>
<th>Question type</th>
<th>Proportion of time for group component (%)</th>
<th>Group component marks (%)</th>
<th>Group component higher mark?</th>
<th>Student Rating</th>
<th>Improved understanding or retention (either compared to other cohort, or in subsequent assessment)</th>
<th>Students generally reported a reduction in anxiety and/or stress?</th>
<th>How were groups chosen?</th>
<th>Number of Students (Number of students included in analysis, if applicable)</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stearns (1996)</td>
<td>Research Methods and Statistics</td>
<td>Yes</td>
<td>MCQ</td>
<td>50</td>
<td>~33</td>
<td>Yes</td>
<td>Not tested</td>
<td>Yes</td>
<td>Not tested</td>
<td>Student selected</td>
<td>8 and 25</td>
<td>Class sizes declined from 12 and 26 students respectively</td>
</tr>
<tr>
<td>Yuretich, Khan, Leckie and Clement (2001)</td>
<td>Introductory Oceanography</td>
<td>Some</td>
<td>MCQ</td>
<td>60</td>
<td>25</td>
<td>Yes</td>
<td>Positive</td>
<td>Yes</td>
<td>Not tested</td>
<td>Neighbours in the classroom</td>
<td>~600 in each cohort</td>
<td>Group mark not counted if it would lower a student’s grade.</td>
</tr>
<tr>
<td>Rao, Collins and DiCarlo (2002)</td>
<td>Cardiovascular Physiology</td>
<td>Yes</td>
<td>Fill in the blanks; MCQ; SAQ; true/false</td>
<td>50</td>
<td>20</td>
<td>Yes</td>
<td>Positive</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Instructor assigned</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Cortright, Collins, Rodenbaugh and DiCarlo (2003)</td>
<td>Exercise Physiology</td>
<td>Yes</td>
<td>MCQs; fill in the blanks; short answer essays</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Yes</td>
<td>Positive</td>
<td>Not tested</td>
<td>Not stated</td>
<td>Instructor assigned</td>
<td>38</td>
<td>Groups were mostly pairs</td>
</tr>
<tr>
<td>Zipp (2007)</td>
<td>Introductory Sociology</td>
<td>Yes</td>
<td>MCQ</td>
<td>~72</td>
<td>Variable, ~13, based on a weighting system</td>
<td>Yes</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Instructor selected to maximise diversity</td>
<td>122</td>
<td>Weighting system designed to not excessively benefit lower-performing students.</td>
</tr>
<tr>
<td>Giuliodori, Lujan and DiCarlo (2008)</td>
<td>Veterinary Physiology</td>
<td>Yes</td>
<td>MCQ</td>
<td>~43</td>
<td>0</td>
<td>Positive</td>
<td>Not tested</td>
<td>Yes</td>
<td>Students chose</td>
<td>65</td>
<td>Groups were pairs</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Subject Area</td>
<td>Did the questions in the group component appear in the individual component?</td>
<td>Question type</td>
<td>Proportion of time for group component (%)</td>
<td>Group component marks (%)</td>
<td>Group component higher mark?</td>
<td>Student Rating</td>
<td>Improved understanding or retention (either compared to other cohort, or in subsequent assessment)</td>
<td>Students generally reported a reduction in anxiety and/or stress?</td>
<td>How were groups chosen?</td>
<td>Number of Students (Number of students included in analysis, if applicable)</td>
<td>Notes</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-----------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------</td>
<td>-------------------------------------------</td>
<td>---------------------------</td>
<td>----------------------------</td>
<td>-----------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>---------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Eaton (2009)</td>
<td>Two Introductory Geoscience courses</td>
<td>Yes for one course, no for the other</td>
<td>MCQ</td>
<td>~37</td>
<td>Improvement in performance across multiple two-stage exams</td>
<td>Not tested</td>
<td>Yes</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Alphabetically in one course, neighbours in the other.</td>
<td>400-500 and 50-100+</td>
<td></td>
</tr>
<tr>
<td>Plotnick, Varelas and Fan (2009)</td>
<td>The Physical World</td>
<td>Yes</td>
<td>Pre-test included MCQ and constructed response</td>
<td>40</td>
<td>20</td>
<td>Positive</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Random suggested</td>
<td>131 over three years</td>
<td></td>
</tr>
<tr>
<td>Yu, Tsiknis and Allen (2010)</td>
<td>Two Computer Science courses</td>
<td>Yes</td>
<td>SAQ, programming code, computations</td>
<td>~38</td>
<td>Weighted average, score of group component pegged to individual if lower</td>
<td>Positive</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Instructor assigned to maximise diversity of major, and performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macpherson, Lee and Stooples (2011)</td>
<td>Earthquakes and Natural Disasters</td>
<td>Yes</td>
<td>Short to medium length essay</td>
<td>50</td>
<td>Weighted between whole exam and the subset of questions where they were the individual lead</td>
<td>Positive</td>
<td>Yes</td>
<td>Yes on the final exam, for students with the lowest performance in individual pre-test</td>
<td>Not tested</td>
<td></td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Leight, Saunders, Calkins, and Withers (2012)</td>
<td>Introductory Biology</td>
<td>Yes</td>
<td>MCQ, multiple correct, true/false and sequencing problems</td>
<td>50</td>
<td>Bonus mark scheme</td>
<td>Positive</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Students chose</td>
<td>~250</td>
<td></td>
</tr>
<tr>
<td>Gilley and Clarkston (2014)</td>
<td>Natural Disasters</td>
<td>Yes</td>
<td>MCQ</td>
<td>50</td>
<td>Not tested</td>
<td>Yes</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Students chose</td>
<td>98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Study involved individual retests as well as group retests.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Subject Area</th>
<th>Did the questions in the group component appear in the individual component?</th>
<th>Question type</th>
<th>Proportion of time for group component (%)</th>
<th>Group component marks (%)</th>
<th>Group component higher mark?</th>
<th>Student Rating</th>
<th>Improved understanding or retention (either compared to other cohort, or in subsequent assessment)</th>
<th>Students generally reported a reduction in anxiety and/or stress?</th>
<th>How were groups chosen?</th>
<th>Number of Students (Number of students included in analysis, if applicable)</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ives (2014)</td>
<td>Introductory Physics; Fluids, Waves and Energy</td>
<td>Some</td>
<td>MCQ</td>
<td>Not given</td>
<td>15, score of group component pegged to individual if lower</td>
<td>Not stated</td>
<td>Not tested</td>
<td>Results were mixed</td>
<td>Not tested</td>
<td>Students chose</td>
<td>679</td>
<td></td>
</tr>
<tr>
<td>Rieger and Heiner (2014)</td>
<td>Introductory Physics</td>
<td>Some</td>
<td>MCQ and SAQ</td>
<td>~33</td>
<td>15</td>
<td>Yes</td>
<td>Positive</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Either preformed, instructor assigned or students chose</td>
<td>178</td>
<td></td>
</tr>
<tr>
<td>Fengler and Ostafchuk (2015)</td>
<td>Mechanical Design</td>
<td>Yes</td>
<td>MCQ</td>
<td>50</td>
<td>Not stated</td>
<td>Yes</td>
<td>Positive</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Instructor selected to maximise diversity</td>
<td>750 over six years</td>
<td>Also tested two-stage exams were the second stage is individual but open book.</td>
</tr>
<tr>
<td>Knierim, Turner and Davis (2015)</td>
<td>Introductory Geology</td>
<td>Yes</td>
<td>MCQ or similar (true/false, matching)</td>
<td>Not stated</td>
<td>25, score of group component pegged to individual if lower</td>
<td>Yes</td>
<td>Not tested</td>
<td>Yes</td>
<td>Not tested</td>
<td>Not stated</td>
<td>~200</td>
<td></td>
</tr>
<tr>
<td>Green, Cates, White, and Farchione (2016)</td>
<td>Anatomy</td>
<td>Yes</td>
<td>Practical tests, SAQ</td>
<td>~45</td>
<td>25</td>
<td>Yes</td>
<td>Not tested</td>
<td>No</td>
<td>Not tested</td>
<td>Not stated</td>
<td>207</td>
<td></td>
</tr>
<tr>
<td>Lindsley, Morton, Pippitt, Lamb and Colbert-Getz (2016)</td>
<td>Foundational Sciences</td>
<td>Yes</td>
<td>MCQ</td>
<td>~37</td>
<td>10</td>
<td>Yes</td>
<td>Not tested</td>
<td>Yes for concepts that were originally answered incorrectly</td>
<td>Not tested</td>
<td>Randomly</td>
<td>104</td>
<td>Group-stage answers did not need to be identical across group members.</td>
</tr>
<tr>
<td>Vogler and Robinson (2016)</td>
<td>Undergraduate Educational Psychology</td>
<td>Yes</td>
<td>MCQ</td>
<td>&gt;50</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Positive</td>
<td>No in a followup test 2 weeks later, yes for a followup 2 months later</td>
<td>Only a minority of students found the two-stage format made them more nervous</td>
<td>Randomly, stratified by sex and year level</td>
<td>51 and 39</td>
<td>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.</td>
</tr>
<tr>
<td>Reference</td>
<td>Subject Area</td>
<td>Did the questions in the group component appear in the individual component?</td>
<td>Question type</td>
<td>Proportion of time for group component (%)</td>
<td>Group component marks (%)</td>
<td>Group component higher mark?</td>
<td>Student Rating</td>
<td>Improved understanding or retention (either compared to other cohort, or in subsequent assessment)</td>
<td>Students generally reported a reduction in anxiety and/or stress?</td>
<td>How were groups chosen?</td>
<td>Number of Students (Number of students included in analysis, if applicable)</td>
<td>Notes</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>---------------</td>
<td>------------------------------------------</td>
<td>---------------------------</td>
<td>----------------------------</td>
<td>-----------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>---------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Cao and Porter (2017)</td>
<td>Computer Science</td>
<td>Similar</td>
<td>Problem solving</td>
<td>~26</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Not tested</td>
<td>Yes for mid-performing students, no for low and high performing students</td>
<td>Not tested</td>
<td>Groups varied in performance heterogeneity</td>
<td>278 (247) in two sections</td>
<td>More homogenous groups of students showed greater performance increases on the final exam than more heterogeneous groups.</td>
</tr>
<tr>
<td>Fournier, Courret, Ramsey and Caulkins (2017)</td>
<td>Introductory Human Anatomy</td>
<td>Yes</td>
<td>MCQ</td>
<td>Not stated</td>
<td>15</td>
<td>Yes</td>
<td>Positive</td>
<td>No</td>
<td>In general, no.</td>
<td>Students chose</td>
<td>444</td>
<td></td>
</tr>
<tr>
<td>Bruno, Engels, Ito, Gillis-Davis, Dulai, Carter, Fletcher and Böttjer-Wilson (2017)</td>
<td>Oceanography and Geology (five courses)</td>
<td>Yes</td>
<td>Variable, including MCQ and essay</td>
<td>Multiple courses, ranged from 15 to 50</td>
<td>Yes</td>
<td>Anecdotally, positive</td>
<td>Not tested</td>
<td>Anecdotally, yes</td>
<td>Randomly</td>
<td>Randomly</td>
<td>289 in seven sections</td>
<td></td>
</tr>
<tr>
<td>McCurdy, Volterman, Shieff, Zeadin, Dunn, De Melo and Helli (2017)</td>
<td>Introductory Human Biochemistry</td>
<td>Yes</td>
<td>MCQ</td>
<td>40</td>
<td>15</td>
<td>Yes</td>
<td>Positive</td>
<td>Not tested (Comparison of the effect of immediate feedback on the group part of the exam was tested)</td>
<td>Yes</td>
<td>Randomly</td>
<td>399 (343)</td>
<td></td>
</tr>
<tr>
<td>Deng and Luo (2018)</td>
<td>Introductory Computing</td>
<td>Yes</td>
<td>MCQ</td>
<td>~33</td>
<td>20</td>
<td>Not stated</td>
<td>Positive</td>
<td>Yes for moderate difficulty questions, no for easy and hard difficulty questions</td>
<td>Not tested</td>
<td>Students chose</td>
<td>127 (54)</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Subject Area</td>
<td>Did the questions in the group component appear in the individual component?</td>
<td>Question type</td>
<td>Proportion of time for group component (%)</td>
<td>Group component marks (%)</td>
<td>Group component higher mark?</td>
<td>Improved understanding or retention (either compared to other cohort, or in subsequent assessment)</td>
<td>Students generally reported a reduction in anxiety and/or stress?</td>
<td>How were groups chosen?</td>
<td>Number of Students (Number of students included in analysis, if applicable)</td>
<td>Notes</td>
<td></td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>-------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------------</td>
<td>-------------------------------------------</td>
<td>---------------------------</td>
<td>-----------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>-----------------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>-----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Levy, Svoronos and Klinger (2018)</td>
<td>Three courses in Statistics/ Econometrics</td>
<td>Yes</td>
<td>Not specified</td>
<td>~31</td>
<td>Up to 10, no penalty if group score was lower than individual</td>
<td>Yes</td>
<td>Positive</td>
<td>Not tested</td>
<td>Yes</td>
<td>Random, including one member from top 40% of class</td>
<td>899 in five cohorts</td>
<td></td>
</tr>
<tr>
<td>Martin (2018)</td>
<td>Two courses: Introductory Statistics and Evolutionary Biology</td>
<td>Yes</td>
<td>MCQ</td>
<td>As long as needed</td>
<td>Not stated</td>
<td>Yes</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Randomly at first test, subsequently assigned to include one each from the lowest and highest performance quartile</td>
<td>753 in four cohorts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shepherd (2018)</td>
<td>Clinical Toxicology</td>
<td>Yes, with additional explanations of answers for the group component</td>
<td>Fixed choice and constructed response</td>
<td>50</td>
<td>~33</td>
<td>Yes</td>
<td>Generally positive</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Students chose</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>Cooke, Weir and Clarkston (2019)</td>
<td>Introductory Biology</td>
<td>Isomorphic in one exam, identical in the other.</td>
<td>SAQ</td>
<td>20, not including an individual retest component</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Positive</td>
<td>Yes, long-term; no, short-term</td>
<td>Students chose</td>
<td>158 (125)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ford (2019)</td>
<td>Human Physiology</td>
<td>Yes</td>
<td>Not stated</td>
<td>50</td>
<td>10</td>
<td>Yes</td>
<td>Not tested</td>
<td>Anecdotally, yes</td>
<td>Instructor selected to increase performance diversity</td>
<td>633 (225)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miller and James (2019)</td>
<td>Introductory Astronomy</td>
<td>Yes</td>
<td>MCQ</td>
<td>~33</td>
<td>25</td>
<td>Yes</td>
<td>Positive</td>
<td>Yes</td>
<td>Instructor selected for diverse majors and year levels and to exclude male-majority groups.</td>
<td>360 in four sections</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Group stage was conducted as a class, with all groups working on the same problem at the same time.

Students with adjustments for extra time started the individual component early, so that they could participate in the group component.

Two-stage exams had a more beneficial effect than flipped lectures on retention.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Subject Area</th>
<th>Did the questions in the group component appear in the individual component?</th>
<th>Question type</th>
<th>Proportion of time for group component (%)</th>
<th>Group component marks (%)</th>
<th>Group component higher mark?</th>
<th>Student Rating</th>
<th>Improved understanding or retention (either compared to other cohort, or in subsequent assessment)</th>
<th>Students generally reported a reduction in anxiety and/or stress?</th>
<th>How were groups chosen?</th>
<th>Number of Students</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newton, Rajakaruna, Kulak, Albabish, Gilley and Ritchie (2019)</td>
<td>Two courses: Biochemistry and Exercise Physiology</td>
<td>Yes</td>
<td>MCQ and long answer</td>
<td>~29</td>
<td>20</td>
<td>Yes</td>
<td>Not tested</td>
<td>Yes</td>
<td>Not tested</td>
<td>Students chose</td>
<td>64 (56) and 102 (94)</td>
<td></td>
</tr>
<tr>
<td>Chen and Kinniburgh (2019)</td>
<td>Introductory Statistics</td>
<td>Yes</td>
<td>MCQ</td>
<td>~47</td>
<td>15</td>
<td>Yes</td>
<td>Positive</td>
<td>Yes</td>
<td>Yes</td>
<td>Randomly</td>
<td>68 in two sections</td>
<td></td>
</tr>
<tr>
<td>Shaffer (2020)</td>
<td>Material and Energy Balances</td>
<td>Yes</td>
<td>SAQ with calculations</td>
<td>40</td>
<td>20</td>
<td>Yes</td>
<td>Positive</td>
<td>No</td>
<td>Not tested</td>
<td>Instructor selected to maintain gender balance, mixture of performance and preference for studying on the same days</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>Weicker (2020)</td>
<td>Algorithms and Data Structures</td>
<td>No</td>
<td>SAQ</td>
<td>~33</td>
<td>~32</td>
<td>No</td>
<td>Anecdotally, positive</td>
<td>Not tested</td>
<td>Not tested</td>
<td>Randomly, but dynamically changing groups during the group component were allowed</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>Bentley, Attardi, Faul, Melo and Palmer (2021)</td>
<td>Anatomy for Medical Radiation Sciences</td>
<td>Yes</td>
<td>MCQ</td>
<td>~33</td>
<td>25</td>
<td>Yes</td>
<td>Positive</td>
<td>No</td>
<td>Majority, yes</td>
<td>Randomly</td>
<td>97 (86) and 99 (81)</td>
<td></td>
</tr>
<tr>
<td>Khong and Tanner (2021)</td>
<td>Essential Proteomics</td>
<td>Yes</td>
<td>SAQ</td>
<td>~33</td>
<td>15, score of group component pegged to individual if lower</td>
<td>Yes</td>
<td>Positive</td>
<td>Not tested</td>
<td>No</td>
<td>Existing assessment groups</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Subject Area</td>
<td>Did the questions in the group component appear in the individual component?</td>
<td>Question type</td>
<td>Proportion of time for group component (%)</td>
<td>Group component marks (%)</td>
<td>Group component higher mark?</td>
<td>Student Rating</td>
<td>Improved understanding or retention (either compared to other cohort, or in subsequent assessment)</td>
<td>Students generally reported a reduction in anxiety and/or stress?</td>
<td>How were groups chosen?</td>
<td>Number of Students (Number of students included in analysis, if applicable)</td>
<td>Notes</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------</td>
<td>---------------------------------------------</td>
<td>---------------------------</td>
<td>-----------------------------</td>
<td>----------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>----------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Kinnear (2021)</td>
<td>Three undergraduate Mathematics units</td>
<td>Yes; Yes; Similar (SAQ converted to MCQ)</td>
<td>SAQ; MCQ; SAQ and MCQ</td>
<td>Unweighted; 50; 30 (score of group component pegged to individual if lower)</td>
<td>Yes; Yes; Yes</td>
<td>Positive; not tested; mixed</td>
<td>Yes; inconclusive; no</td>
<td>Yes; not tested; no</td>
<td>Students chose; Instructor assigned; Students chose</td>
<td>47; 24; 301 (254)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rempel, Dirks and McGintie (2021)</td>
<td>General (Introductory) Chemistry</td>
<td>Yes</td>
<td>Not stated</td>
<td>~28-50 (score of group component pegged to individual if lower)</td>
<td>Yes</td>
<td>Positive</td>
<td>Not tested</td>
<td>Yes</td>
<td>Not tested</td>
<td>Students chose</td>
<td>55 (49); 46 (39); 60 (41)</td>
<td></td>
</tr>
<tr>
<td>Walker and Robinson (2022)</td>
<td>Undergraduate Educational Psychology</td>
<td>Yes</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Not tested</td>
<td>Not tested</td>
<td>No in either short or long term</td>
<td>Not tested</td>
<td>Randomly</td>
<td>49; 44; 208</td>
<td>Study designed as a follow up to Vogler and Robinson (2016)</td>
</tr>
</tbody>
</table>