The effects of collaborative testing on higher-order thinking: Do the bright get brighter?

John W Mahoney and Brooke Harris-Reeves

Biographical/contact details of author(s)

John W. Mahoney, School of Exercise Science, Australian Catholic University, Brisbane, 4014, Australia. Tel: +61 7 38616076. Email: jwjmahoney@gmail.com

Brooke Harris-Reeves, School of Allied Health Sciences, Griffith University, Gold Coast, 4215, Australia. Email: b.harris-reeves@griffith.edu.au.

John Mahoney is a lecturer in the areas of sport, exercise, and health psychology, teaching into undergraduate and postgraduate programs in exercise science. His research interests are divided between learning and teaching in higher education and the benefits of sport for psychosocial development in youth.

Brooke Harris-Reeves is the program director for the Bachelor of Sport Development at Griffith University. She teaches into undergraduate courses in sport development and is interested in researching topics in teaching and learning with a particular focus on embedding employability frameworks within and across degrees.

Abstract

Collaborative testing has been shown to enhance student performance compared to individual testing. It is suggested that collaborative testing promotes higher-order thinking, but research has yet to explore this assumption directly. The aim of this study was to explore the benefits of collaborative testing on overall performance, as well as performance on higher-order thinking questions. It was hypothesised that, compared to individual test results, students would perform better overall and on higher-order thinking questions under collaborative testing conditions. It was expected that these differences would be equal when comparing students of different academic abilities (that is, “upper”, “middle”, and “lower” performers). Undergraduate students completed an individual followed by a collaborative test as part of summative assessment. Analyses revealed that, with the exception of upper performers, students performed better overall on the collaborative test. Additionally, regardless of their academic abilities, students performed better on the higher-order thinking questions under collaborative conditions. This improvement was equal across different academic abilities, suggesting that collaborative testing promotes higher-order thinking even when taking into account previous academic achievement. The acceptability and application of collaborative testing is discussed.

Keywords: social interdependence theory; active learning; learning assessment; Bloom’s Taxonomy

Collaborative testing enhances student performance and critical thinking

Collaborative (or cooperative) tasks are an effective approach for enhancing student outcomes (Zhao and Kuh, 2004; Cortright et al., 2005; Bloom, 2009; Johnson and Johnson, 1989; Johnson et al., 2007; Slavin, 1996; Springer et al., 1999). Such tasks involve students working in dyads or small groups to achieve set goals. The opportunity to collaborate allows

students opportunities to share perspectives, debate points of disagreement, question and understand each other's points of view, problem-solve complex dilemmas, and reach agreements (Zhao and Kuh, 2004; Bloom, 2009; Johnson and Johnson, 1989; Johnson et al., 2007). Student testing is one setting where collaborative tasks can be embedded. Where individual testing protocols may promote the use of rote learning strategies and other rudimentary approaches to knowledge retention, collaborative testing calls upon higher levels of abstraction and deeper understanding (Bloom, 2009; Gilley and Clarkston, 2014; Millis and Cottel, 1998; Slavin, 1996). In other words, collaborative testing is a strategy that encourages the active use of advanced cognitive systems important for higher-level learning (Cortright et al., 2005). Despite such conceptual grounds, researchers have not yet explored how collaborative testing influences higher-order thinking.

While appreciated by students (Herrmann, 2013), collaborative approaches are more valued if they are related to assessment (Machemer and Crawford, 2007). Embedding collaborative approaches into assessment tasks resonates with the two basic principles of cooperative learning, namely positive interdependence (that is, where collective efforts are essential to individual achievement) and individual accountability (that is, individuals are evaluated independently of each other; Millis and Cottel, 1998; Slavin, 1996). Arguably, collaborative assessment tasks are an ideal setting in which positive interdependence and individual accountability can be targeted. While embedding collaborative approaches within assessment tasks has its advantages, there are some limitations that warrant acknowledgment. There exists a cocktail of individual and group level motivations, personalities, and social phenomena that influence students' experiences in collaborative approaches and are difficult – if not impossible – to control or regulate. Specifically, students' motivations to perform (French and Kottke, 2013; Morgeson et al., 2005), the compatibility of group members' personalities (Driskell et al., 2006; Humphrey et al., 2011; O'Neill and Kline, 2008; Peeters et al., 2006a; Peeters et al., 2006b), attempts to save face and avoid embarrassment (Robinson et al., 2015; Micari and Pazos, 2014), social intimidation and conformity (Micari and Drane, 2011; Ilarda and Findlay, 2006; Kline, 1999), and conflict management (Mohammed and Angell, 2004; Vuopala et al., 2016) are all factors that have emerged as influencing students' experiences when participating in collaborative task. When this cocktail is 'unbalanced' there is a risk of encountering undesirable student outcomes such as freeriding, resentment towards peers, conflict, and distrust (confer, Herrmann, 2013), amongst many other factors. Nevertheless, the potential gains that can be made through the use of collaborative approaches, particularly collaborative testing, provides a rationale for employing and continuing to implement such strategies.

Collaborative testing is a strategy founded in social constructivism, specifically within social interdependence theory (Johnson and Johnson, 1989). According to this perspective, individuals are active agents in an ongoing process of learning that occurs as a result of interactions with others and the environment (Clark, 1998; Mishra, 2014; Johnson and Johnson, 1989). For example, a child who has no prior knowledge of mathematics can be taught simple addition (for example, $1 + 1 = 2$) through interactions with a teacher. This knowledge can be advanced over time to include more complex, linear knowledge (for example, summing fractions), as well as non-linear knowledge (for example, subtraction, multiplication, division). Such advances would again be dependent on interactions with knowledgeable social agents such as teachers and parents. As can be seen from this example, knowledge is non-existent prior to social interaction and frequently undergoes development and reconstruction. While this is a basic example, the same premise extends to all formal and informal learning environments, be they small or large in size (Mishra, 2014). The conceptual foundations of social constructivism inform a number of contemporary pedagogical approaches such as group work, peer marking, simulation training, of which collaborative testing is but one example (Giuliodori et al., 2008). In light of such knowledge, it can be argued that learning occurs through collaborative testing because students are afforded opportunities to actively develop, reconstruct, and advance their knowledge through interactions with others.

Researchers have supported the positive effects collaborative testing has on student knowledge in tertiary education settings (Johnson et al., 2007; Gilley and Clarkston, 2014; Springer et al., 1999). For example, Rao et al. (2002) found that university students

performed significantly better on multiple-choice, short-essay, and true/false tests when working collaboratively compared to individually. Another study by Eaton (2009) explored collaborative testing with two undergraduate university classes. In this study, students completed either four or five collaborative tests throughout a university semester. As with Rao et al.'s study, students in Eaton's experiment performed better on the collaborative tests. Eaton also advanced previous findings by reporting that the positive effects of collaborative testing on student performance were consistent across class level, as well as time. Another group of researchers (Cortright et al., 2002) found similar findings to Rao et al. (2002) and Eaton (2009). These researchers identified that knowledge was not only improved in university students under collaborative testing conditions, but was also retained after follow-up testing completed several weeks later.

In the aforementioned studies, students completed collaborative tests after individual tests. As such, improved performances may have been due to test frequency, practice effects, and/or reduced psychological factors that might otherwise negatively affect test performance. Indeed, this concern extends to the majority of research that has explored the association between collaborative testing and student performance (for example, Bloom, 2009; Kapitanoff, 2009). To address this concern, Gilley and Clarkston (2014) conducted a controlled study with a group of undergraduate university students. These authors presented a series of test questions to students before either sitting individual or collaborative tests. As with studies that have only explored the benefits of collaborative testing after individual testing, Gilley and Clarkston found that students in the collaborative test condition performed better than those in the individual test condition. Such findings, coupled with those above, highlight the meaningful and unique contributions collaborative testing has on student performance.

Researchers (for example, Eaton, 2009; Gilley and Clarkston, 2014) have also explored the effects of collaborative testing on student performances across achievement levels. In these studies, student performances improved under collaborative testing conditions compared to individual tests regardless of achievement level. That is, categorised based on results from individual tests, upper, middle, and lower achievers all demonstrated improved performances in the collaborative test conditions. However, others (Giuliodori et al., 2008) have reported that the effects of collaborative testing on student performance may be more pronounced in lower compared to upper achieving students. In all these studies, little attention has been paid to the types of questions students are asked. Specifically, it is not known if, through collaboration, students perform better on questions that address more basic knowledge, more complex knowledge, or both. Investigating the types of questions students perform better on through collaboration would provide further insight into the benefits of collaborative testing protocols, especially when considering different achievement levels.

Bloom, in his taxonomy of education objectives (see, Krathwohl, 2002), posited that student learning is structured from basic to complex and concrete to abstract. He detailed six categories within this structure that included – from basic and concrete to complex and abstract – knowledge, comprehension, application, analysis, synthesis, and evaluation. The first three were of principal interest to the current study. Knowledge concerns the recall of information and understanding of basic concepts; comprehension refers to the integration of information and the ability to explain ideas; and application concerns the practical use of information beyond conceptual knowledge. Morrison and Free (2001) illustrated that multiple-choice test questions could be designed around Bloom's taxonomy as a method for assessing different levels of student learning and knowledge. These authors suggested that designing questions around Bloom's Taxonomy is an effective approach for discriminating between more and less knowledgeable students. One question that arises from the works of Morrison and Free is how collaborative testing affects student performance on higher-order thinking questions (that is, questions designed around upper structures of Bloom's Taxonomy such as application). It is likely that upper achieving students already perform well on such questions and that collaboration will only benefit lower achieving students. However, if the principles of social constructivism hold, upper achieving students should also benefit from collaboration when answering higher-order thinking questions. This benefit might not be as pronounced as lower achieving students, but should still be present because of the opportunity to share perspectives, debate points of disagreement, and draw a consensus.

In light of the above, the aim of this study was to explore the effects of collaborative testing on overall student performance, as well as when responding to higher-order thinking questions. It was hypothesised that students' performances, both overall and when responding to higher-order thinking questions, would improve under the collaborative test condition compared to and regardless of individual test performance. Additionally, attempts were made to garner students' perspectives about their experiences following the collaborative test. While garnering students' perspectives was a peripheral focus of the study, it had the potential to offer insight into the positive and negative aspects of collaborative testing beyond test performance alone.

Method

Participants

Participants were 168 ($M_{age} = 22.43$ years, $SD = 2.54$) male ($n = 97$) and female ($n = 71$) students enrolled in undergraduate health science degrees (exercise science = 124; psychology = 25; health sciences = 9; dietetics and nutrition = 7; other = 3) at an Australian university. Students were in various years of their degrees (first year = 7; second year = 50; third year = 106; fourth year = 5) and completing a subject titled, sport and exercise psychology (see below for a description of this subject).

Subject Structure and Content

The subject was delivered over a 12-week semester in the academic year (2016) and covered topics related to sport and exercise psychology. Each week, students attended a 2-hour lecture and a 2-hour tutorial. The subject was separated into three modules, each spanning 4 weeks. The first module focused on foundation knowledge relevant to sport and exercise psychology (for example, personality, motivation), the second addressed psychology in performance settings (for example, attention and concentration, mental toughness), and the third module concerned the effects of physical activity on psychological health (for example, mental health, exercise adherence). Students completed examinations following each module. That is, three examinations; the first in week 5, the second in week 9, and the third following week 12, during the University's examination period. The examination that followed the second module was coordinated as an individual and collaborative test and, hence, was the principal focus of this study. As part of the subject, students also completed a group research report. They were allocated to groups of 3-5 students during the first week of classes and continued to work together until the end of semester.

Measures

Overall academic performance. Students' overall academic performances were measured as the number of correctly answered examination questions. Students were presented with 30 multi-choice questions that pertained to the content presented across the four lectures delivered in module two. Higher scores represented better academic performance, with 30 being the highest possible score (that is, 1 mark per correct answer). Students received two scores, one for the individual test (worth 80% of the assessment) and another for the collaborative test (worth 20% of the assessment).

Higher-order thinking. Five of the 30 questions (see Table 1 for examples) were designed around applied knowledge, as described by Bloom (Krathwohl, 2002). These questions represented more advanced knowledge compared to the other examination questions. As with measuring students' overall academic performance, an individual and collaborative test score for the five questions (highest possible score = 5; 1 mark for each question) were used as a measure of higher-order thinking performance.

<<Insert Table 1 here>>

Students' perceptions of testing. Four closed and one open question were used to garner students' perspectives about the collaborative test. The closed questions required students to

respond on a five-point scale (1 = *strongly disagree*; 5 = *strongly agree*) and asked their opinions about the opportunity to collaborate, the effects on learning, perceived levels of enjoyment, and task worth. The open question followed and sought students' insights about their experiences, both positive and negative, of the collaborative test.

Procedure

Following institutional ethical approval, students were presented with details about the study and consent was sought. When sitting the examination, students were first afforded 45 minutes to complete the 30 multiple-choice test questions individually. After the individual test, students were directed to form small groups. Group allocation was based on the groups students were already working in as part of the research report assessment task for the subject. In these groups, students were afforded 30 minutes to complete the same 30 multiple-choice questions as those presented in the individual test. In both the individual and collaborative test, students completed their own answer sheets, allowing for different answers between students. These answer sheets were collected following each part of the test to avoid students changing their individual test answers during the collaborative test. Following the examination, students were emailed a link to an online survey that included the questions related to students' perceptions of the collaborative test. Students' results were made available two weeks following the examination at which time the online survey was closed.

Data Analysis

Participants were divided into three roughly equal sized groups based on their individual test performances: upper performers (>90%; $n = 50$), middle performers (80%-89%; $n = 53$), and lower performers (<80%, $n = 64$). Additionally, normalised change scores were calculated to compare differences in overall individual and collaborative test scores, as well as performances on critical thinking questions while taking into consideration the potential for improvement. Similar equations as Gilley and Clarkston (2014) were used to calculate normalised change scores. Normalised change scores for students who performed better in the collaborative test compared to the individual test were calculated as follows:

$$c = 100 \times \left(\frac{Score_{coll} - Score_{ind}}{N_{questions} - Score_{ind}} \right)$$

For students who performed worse in the collaborative test compared to the individual test, the following equation was used:

$$c = 100 \times \left(\frac{Score_{coll} - Score_{ind}}{Score_{ind}} \right)$$

In both equations, c referred to normalised change, $score$ referred to the number of correct answers for the examination or the number of correct answers out of the five higher-order thinking questions; $coll$ and ind referred to the collaborative and individual tests respectively, and $N_{questions}$ referred to the number of questions pertinent to the calculation (that is, overall performance = 30; critical thinking = 5). A normalised change score of zero was awarded to students whose individual and collaborative test scores were the same, while students were excluded from the change score analyses if they scored full marks across the individual and collaborative tests overall ($n = 1$) or on the five higher-order thinking questions ($n = 12$) as no change was possible.

Manipulation Check

For the purpose of conducting a manipulation check, individual test results were converted to scores out of five so that they could be compared to the results for the five higher-order thinking questions.

Results

Analysis revealed that students performed worse on the higher-order thinking questions ($M = 3.28$, $SD = 1.03$) compared to the other questions ($M = 4.19$, $SD = 0.37$) in the individual test, indicating that the higher-order thinking questions were more challenging than the other test question, $t(167) = -13.92$, $p < 0.001$.

Table 2 shows the means, standard deviations, and mean differences for the student cohort, as well as the cohort separated into upper, middle, and lower performers.

<<Insert Table 2 here>>

Paired-samples t tests revealed that, as a cohort, students improved across the individual and collaborative tests and the higher-order thinking questions. A further series of paired-samples t -tests revealed that all groups performed better overall and on the higher-order thinking questions in the collaborative test compared to the individual test. The only exception to this was that the upper performers did not improve (nor worsen) when comparing their overall individual and collaborative test results. A one-way analysis of variance of the normalised change scores revealed a significant difference in overall performance between student groups, $F(2,164) = 24.12$, $p < 0.001$. Post-hoc tests using the Bonferroni correction indicated that the upper performance group's test scores ($c = 21.78$, $SD = 32.86$) changed less compared to the middle ($c = 47.83$, $SD = 24.05$), $p < 0.001$, and lower ($c = 54.85$, $SD = 21.17$), $p < 0.001$, performance groups. This finding suggests that upper performers did not benefit as much from the collaborative test as the middle and lower performers on overall performance. However, the normalised change scores for the upper ($c = 34.43$, $SD = 46.25$), middle ($c = 53.62$, $SD = 44.37$), and lower ($c = 43.25$, $SD = 39.25$) performers on the higher-order thinking questions were not significantly different, $F(2,153) = 2.33$, $p = 0.101$. This result indicates that students benefitted equally in the collaborative test when answering higher-order thinking questions, regardless of baseline performance.

Fifty-six participants completed the close question section of the online student perceptions survey. Responding *agreed* or *strongly agreed* to the closed questions was coded as positive experiences of the collaborative test. The majority (66%) of students indicated that they would like opportunities to sit collaborative tests in the future; 67% indicated that the test enhanced their learning; 83% found the test enjoyable; and 88% found the test worthwhile. Thirty students completed the open question and provided details about how the collaborative test facilitated learning, as well as suggestions for similar procedures in the future. Specifically, students identified that the collaborative test allowed answers from the individual test to be confirmed (for example, "it helped me confirm my answers"), encouraged students to justify their points of view (for example, "it was good to discuss questions and express why each answer was chosen for each group member"), and reduced examination stress (for example, "the relaxed environment was welcome and removed a lot of the stress usually associated with sitting an exam"). Very few suggestions were proposed for future collaborative tests, but two points were reported on several occasions: reducing the noise level (for example, "perhaps [sit the exam in] a bigger room so it would be less loud") and shortening the collaborative test (for example, "if doing a collaborative exam that is identical to that of the individual exam, no more than 30 mins is needed when doing the same exam").

Discussion

The aim of this study was to explore the effects of collaborative testing on student performance, with a particular focus on the effects of such protocols on higher-order thinking. The study hypothesis was partially supported. Collaborative testing resulted in improved student performances. However, when taking baseline performance into consideration, upper performing students did not appear to benefit from collaborative testing. This finding was further supported by the normalised change results that illustrated that middle and lower performing students benefitted most from the collaborative test. This said, collaborative testing resulted in improved performances on higher-order thinking questions, regardless of baseline performance. Further, the change in performance on higher-order thinking questions was similar regardless of baseline performance, suggesting that, regarding higher-order thinking, all students benefitted from the collaborative test equally. The benefits of collaborative testing also extended beyond test performance with many students reporting

positive experiences as a result of participating in the testing protocol. These findings are elucidated further in the following sections.

Overall Student Performance

The overall improvements in students' performances from the individual to the collaborative test are consistent with research (Bloom, 2009; Eaton, 2009; Gilley and Clarkston, 2014; Giuliadori et al., 2008; Rao et al., 2002). These findings have been explained as being due to the opportunity to share ideas, discuss points of view, debate opposing beliefs, and reach agreements (Bloom, 2009; Wiggs, 2011). Some have also suggested that collaborative protocols reduce test anxiety, which might otherwise hinder students' performances (Bloom, 2009; Kapitanoff, 2009). The students' perspectives described in this article resonate with these findings (Eaton, 2009; Kapitanoff, 2009; Wiggs, 2011). However, while the findings complement the existing body of literature, the lack of evidence in the study described in this article to suggest upper performing students benefit from collaborative testing is inconsistent with some work (Eaton, 2009; Gilley and Clarkston, 2014; Giuliadori et al., 2008). Researchers have demonstrated that, while not to the same extent as lower achieving student, upper achieving students still perform better overall under collaborative testing conditions (Giuliadori et al., 2008). This was not the case in the study described here. One possible explanation for this finding might be that upper performers succumbed to peer pressure and changed some of their correct answers from the individual tests to incorrect answers following group discussions (but also answered critical thinking questions correctly in the collaborative and incorrectly in the individual test, balancing their overall scores between the two tests; see below for further discussion). Elucidating this point further, upper performers were in the minority (that is, roughly one-third of the cohort) and, as such, may have been convinced by the majority that some of their correct answers were wrong. This possibility can also be explained in light of social constructivism. Specifically, social constructivism details that learning is predicated on by interactions with other, but makes no assumption that the knowledge of others is correct in the first place (Clark, 1998; Mishra, 2014).

Higher-Order Thinking

A finding to emerge from the study was the improved performances of students on higher-order thinking questions under collaborative testing conditions, regardless of baseline performances. This finding is consistent with the conceptual tenets of social constructivism (Clark, 1998; Mishra, 2014). That is, interacting with others calls upon higher levels of abstraction and promotes deeper understanding due to the active exchange of ideas and the need to argue for personal points of view. A surprising result of the study, especially when considering the different overall performances of students when accounting for baseline achievements, was the improvement in higher-order thinking demonstrated by the upper performers. These individuals not only answered more higher-order thinking questions correctly in the collaborative test compared to the individual test, but also improved to the same extent as middle and lower performers when comparing normalising change scores. One might assume – and the results seemed to indicate – that middle and lower performers are less likely to possess higher-order knowledge, especially when compared to upper performers. This notion raises the question, if middle and lower performers were unlikely to know the answers to the higher-order thinking questions, where did upper performers acquire this knowledge in the collaborative test? Alternatively, from where did this new knowledge emerge? In line with social constructivism, the argument could be made that this new knowledge emerged from meaningful interactions with others. That is, interactions that required all students, regardless of baseline performance and working on the assumption that no one had a firm understanding of the content, to deliberate, discuss, and debate higher-order concepts.

The findings from the study may provide further clarity concerning overall improvements in upper performing students in research (for example, Eaton, 2009; Gilley and Clarkston, 2014). Namely, the improvements in upper performing students under collaborative compared

to individual testing conditions in research may be due to more higher-order thinking questions being asked in such studies. Indeed, if a greater percentage of questions in the current study were designed to address higher-order thinking, it is possible that upper performing students would have demonstrated overall improvement following the collaborative test.

Students' Perceptions of Collaborative Test

Students' perceptions of the collaborative test indicate the acceptability of such protocols. These generally positive perceptions are consistent with findings (Bloom, 2009). The opportunity to collaborate appears to reduce test anxiety, but also promote the perception that students are active agents in their learning. In other words, collaborative testing protocols represent environments that empower students and engage them in their learning. While the majority of students appeared to appreciate the opportunity to complete collaborative testing, some students disliked the experience. This disliking may be due to pragmatic reasons identified by students in the study, such as the noise during and duration of the collaborative test, but may also be due to social factors. Freeriding, resentment towards peers, conflict, and distrust have been identified as negative aspects of engaging in collaborative tasks and may have been experienced by students in this study (Herrmann, 2013). In the future, researchers could explore why students dislike collaborative testing in more detail, as well as explore if students' perceptions of collaborative testing correlate with test performance.

A number of other avenues for future research arise when considering the limitations of this study. One limitation was that the unstandardized make-up of the collaborative test groups. It is possible that some groups consisted of all upper performing students, others of all lower performing students, and other still of all middle performing students, or a combination of all three performance levels. As an example of the consequences of this possibility, the improved performances on the higher-order thinking questions for the upper performing students in the collaborative test could have been due the presence of other, equally knowledgeable students in their group. While using existing groups was arguably a strength of the study, as students were more likely to have a knowledge of each other's communication styles and an awareness of each other's roles and abilities, researchers could initially allocate students to groups based on their academic track records from their previous studies. A second limitation of the current study was that information pertaining to the interactions between students in the collaborative test was not gathered. In line with social interdependency theory, it is assumed that students discussed and debated points of view and deliberated on their answers as a group (Clark, 1998; Mishra, 2014; Johnson and Johnson, 1989), but such assumptions are speculative. In the future, researchers could collect data related to student interactions by coding their communication as a way of clarifying the dialogue that occurs during collaborative tests. Another limitation of the current study was that only part of Bloom's taxonomy was considered when designing the test questions. In the future, researchers could design questions around all levels of Bloom's taxonomy to further explore the benefits of collaborative testing on student performance. Further limitations include that students were undergraduates only, only from one discipline, the numbers of students in their first and final years were very small, that they were from only one university and from one particular country. Future work is needed to explore collaborative testing with students from certain levels, within other disciplines and cultures/contexts given that there may be differences in light of these factors.

In conclusion, the study described in this article offers support, at least in part, to research that has shown that collaborative testing enhances students' overall performances. A substantive contribution of this study was how it advanced knowledge by demonstrating that collaborative testing improves students' performances on higher-order thinking questions. It was suggested that the tenets of social constructivism inform the underlying mechanisms that link collaborative testing and improved student performance. Specifically, it was posited that collaborative testing protocols improve student performance through active opportunities to deliberate, discuss, and debate ideas with others. These findings are meaningful for educators interested in designing assessment items that examine higher levels of abstraction and who also prescribe to an assessment-for-learning instead of, or as well as an

assessment-for-evaluation philosophy. That said, efforts are required to ensure that educators provide varied learning and assessment tasks. Collaborative testing is but one approach that educators might employ. If they do employ such approaches, they should do so in moderation, ensuring that the use of collaborative testing strategies align with both learning objectives and content. Employing collaborative testing strategies alongside other types of assessment reflects the idea that, in higher education, individuals, and not groups, are awarded grades. Collaboration is a valued activity of many higher education institutions because of how it reflects real-world experiences, but educators are principally responsible for measuring the knowledge and competence of individuals and assessment should primarily reflect this agenda. Nevertheless, taken together, the study demonstrates that through collaborative testing, as well as other pedagogical practices whether informed by social constructivism or other evidence-based research, it possible to promote deeper student learning.

References

- Bloom D. (2009) Collaborative test taking: Benefits for learning and retention. *College Teaching* 57(4): 216-220.
- Clark CM. (1998) Hello learners: Living social constructivism. *Teaching Education* 10(1): 89-110.
- Cortright RN, Collins HL and DiCarlo SE. (2005) Peer instruction enhanced meaningful learning: Ability to solve novel problems. *Advances in Physiology Education* 29(2): 107-111.
- Cortright RN, Collins HL, Rodenbaugh DW, et al. (2002) Student retention of course content is improved by collaborative-group testing. *Advances in Physiology Education* 27(3): 102-108.
- Driskell JE, Goodwin GF, Salas E, et al. (2006) What makes a good team player? Personality and team effectiveness. *Group Dynamics: Theory, Research, and Practice* 10(4): 249-271.
- Eaton TT. (2009) Engaging students and evaluating learning progress using collaborative exams in introductory courses. *Journal of Geoscience Education* 57(2): 113-120.
- French KA and Kottke JL. (2013) Teamwork satisfaction: Exploring the multi-level interaction of teamwork interest and group extraversion. *Active Learning in Higher Education* 14(3): 189-200.
- Gilley BH and 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 MJ, Lujan HL and DiCarlo SE. (2008) Collaborative group testing benefits high- and low-performing students. *Advances in Physiology Education* 32(4): 274-278.
- Herrmann KJ. (2013) The impact of cooperative learning on student engagement: Results from an intervention. *Active Learning in Higher Education* 14(3): 175-187.
- Humphrey SE, Hollenbeck JR, Meyer CJ, et al. (2011) Personality configurations in self-managed teams: A natural experiment on the effects of maximizing and minimizing variance in traits. *Journal of Social Psychology* 151(7): 1701-1732.
- Ilarda E and Findlay BM. (2006) Emotional intelligence and propensity to be a team player. *E-Journal of Applied Psychology* 2(2): 19-29.
- Johnson DW and Johnson RT. (1989) *Cooperation and Competition: Theory and Research*, Edina, Minnesota: Interaction Book Company.
- Johnson DW, Johnson RT and Smith KA. (2007) The state of cooperative learning in postsecondary and professional settings. *Educational Psychology Review* 19(1): 15-29.
- Kapitanoff SH. (2009) Collaborative testing: Cognitive and interpersonal processes related to enhanced test performance. *Active Learning in Higher Education* 10(1): 56-70.
- Kline TJB. (1999) The team player inventory: Reliability and validity of a measure of predisposition toward organizational team-working environments. *Journal for Specialists in Group Work* 24(1): 102-112.
- Krathwohl DR. (2002) A revision of Bloom's Taxonomy: An overview. *Theory into Practice* 41:212-218.
- Machemer PL and Crawford P. (2007) Student perceptions of active learning in a large cross-disciplinary classroom. *Active Learning in Higher Education* 8(1): 9-30.

- Micari M and Drane D. (2011) Intimidation in small learning groups: The roles of social-comparison concern, comfort, and individual characteristics in student academic outcomes. *Active Learning in Higher Education* 12(3): 175-187.
- Micari M and Pazos P. (2014) Worrying about what others think: A social comparison-concern intervention in small learning groups. *Active Learning in Higher Education* 15(3): 249-262.
- Millis BJ and Cattel PG. (1998) *Cooperative Learning for Higher Education Faculty*, Phoenix, Arizona: Oryx Press.
- Mishra RK. (2014) Social constructivism and teaching of social science. *Journal of Social Studies Education Research* 5(2): 1-13.
- Mohammed S and Angell LC. (2004) Surface- and deep-level diversity in workgroups: Examining the moderating effects of team orientation and team process on relationship conflict. *Journal of Organizational Behavior* 25(8): 1015-1039.
- Morgeson FP, Reider MH and Campion MA. (2005) Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge. *Personnel Psychology* 58(3): 583-611.
- Morrison S and Free KW. (2001) Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education* 40(1): 17-24.
- O'Neill TA and Kline TJB. (2008) Personality as a predictor of teamwork: A business simulator study. *North American Journal of Psychology* 10(1): 65-78.
- Peeters MG, Rutte CG, Van Tuijl HM, et al. (2006a) The big five personality traits and individual satisfaction with the team. *Small Group Research* 37(2): 187-211.
- Peeters MG, Van Tuijl HM, Rutte CG, et al. (2006b) Personality and team performance: A meta-analysis. *European Journal of Personality* 20(5): 377-396.
- Rao SP, Collins HL and DiCarlo SE. (2002) Collaborative testing enhances student learning. *Advances in Physiology Education* 26(1): 37-41.
- Robinson L, Harris A and Baurton R. (2015) Saving face: Managing rapport in a problem based learning group. *Active Learning in Higher Education* 16(1): 11-24.
- Slavin RE. (1996) Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology* 21(1): 46-69.
- Springer L, Donovan SS and Stanne ME. (1999) Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research* 69(1): 21-51.
- Vuopala E, Hyvönen P and Järvelä S. (2016) Interaction forms in successful collaborative learning in virtual learning environments. *Active Learning in Higher Education* 17(1): 25-38.
- Wiggs CM. (2011) Collaborative testing: Assessing teamwork and critical thinking behaviors in baccalaureate nursing students. *Nursing Education Today* 31(3): 279-282.
- Zhao CM and Kuh GD. (2004) Adding value: Learning communities and student engagement. *Research in Higher Education* 45(2): 115-138.

Table 1.

Examples of Two Test Items Designed to Assess Higher-Order Thinking

<p>1. A rugby team goes away during pre-season. What is most likely?</p> <p>A. They will be a more developed group by the end of the season compared to teams who didn't go away during pre-season</p> <p>B. They will be a more developed group by the start of the season compared to teams who didn't go away during pre-season</p> <p>C. They will be as developed at the start of the season as teams who didn't go away during pre-season</p> <p>D. They will be no different to other teams regardless of whether or not they went away at the start of the season</p> <p>(Answer: B)</p> <p>2. You're a trap shooter competing in your first ever Olympics. It's your first round. You take your position. Just before your first shot, a bird flies into your peripheral vision. What's likely to occur?</p> <p>A. You will become distracted and miss the shot</p> <p>B. You will become distracted, but still perform well</p> <p>C. You will become distracted, but actually perform better than expected</p> <p>D. You won't be distracted</p> <p>(Answer: D)</p>
--

Table 2.

Means, Standard Deviations, and Mean Difference for Overall Academic Performance and the Higher-Order Thinking Questions Across Tests

		Group							
		Cohort		Upper		Middle		Lower	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Test	Individual	25.14	2.20	27.65	0.87	25.51	0.50	22.83	1.18
	Collaborative	27.40	1.29	27.82	1.21	27.68	1.05	26.84	1.34
	p-value (paired t-test)	$t = -13.59, df = 167, p < 0.001$		$t = 0.92, df = 50, p = 0.36$		$t = -13.87, df = 52, p < 0.001$		$t = -17.99, df = 63, p < 0.001$	
HOTQ	Individual	3.28	1.03	4.00	0.77	3.32	0.78	2.67	1.01
	Collaborative	4.10	0.81	4.37	0.56	4.28	0.69	3.72	0.93
	p-value (paired t-test)	$t = -10.26, df = 167, p < 0.001$		$t = -3.33, df = 50, p = 0.002$		$t = -6.53, df = 52, p < 0.001$		$t = -7.90, df = 63, p < 0.001$	

Note. HOTQ = higher-order thinking questions

Figure 1. Normalised change in performance overall and on higher-order thinking questions (HOTQ) for the three achievement levels using individual scores as a baseline (bars represent standard error).