Protocol:
Collaborative Testing for Improving Student Learning Outcomes and Test-Taking Performance in Higher Education: A Systematic Review

E. Renée Cantwell, Jeanann Sousou, Yuri T. Jadotte, Jenny Pierce, and Leo E. Akioyamen

Submitted to the Coordinating Group of:

☐	Crime and Justice
☒	Education
	☐ Disability
☐	International Development
	☐ Nutrition
☐	Social Welfare
☐	Other:

Plans to co-register:

☒	No		
☐	Yes	☐ Cochrane	☐ Other
☐	Maybe		

Date Submitted: 10 February, 2014
Date Latest Revision Submitted: 10 September, 2016
Approval Date:
Publication Date:

Campbell Collaboration Systematic Review Protocol Template version date: 9 February 2014

BACKGROUND

The Problem

The primary purpose of assessment in an instructional setting is twofold: to determine whether learners have achieved the stated objectives and learner outcomes described in the curriculum and to determine whether educators meet those learning objectives in the classroom. Assessment, quite often, is administered in the form of a written or computerized test or exam. An exam is a measurement instrument designed to measure knowledge and understanding of defined content (Gaberson, 2008). Testing is an important activity for the learner as the learner needs to identify how they performed on the test, and whether test results allow them to progress in their program (Sainsbury & Walker, 2008). Testing is also important for the educator. Testing is a means to identify whether teaching is effective and how well the student comprehends the material.

Because testing is a "high-stakes" (Hoachlander, 1998) activity in which students must offer their best performance, it can be anxiety provoking and can negatively affect their performance. There are many reasons why students might underperform on tests, including poor performance related to test anxiety, poor study skills, improper test preparation, language difficulties, and the presence of cultural bias through poorly written questions (Lusk & Conklin, 2003). There is little empirical evidence available on the validity of various testing formats, yet educators still rely heavily on traditional individual testing as a method of evaluation or assessment (Halstead, 2007). An additional complicating factor is that all tests have error; without conducting a proper item analysis, poorly constructed or unvalidated tests can inaccurately reflect student knowledge (Gaberson, 2008). Because of the many disadvantages inherent in traditional testing, alternate testing methods have been explored. One of these is collaborative testing, a relatively recent modality that is becoming more frequently utilized as an alternative.

Collaborative Testing

The intervention that will be examined in this review is collaborative testing. Collaborative testing is a method in which students work together while taking a written evaluative exam. Collaborative, or cooperative, testing is an umbrella term used to describe a test method different from traditional or individualized testing. Collaborative testing has been utilized in undergraduate, graduate, and post graduate settings. Collaborative testing, which is a student-centered, active learning approach, is also referred to as group testing, double testing, paired testing, cooperative testing, and dyad testing (Centrella-Nigro, 2012).

Group testing is the term utilized when a test is administered to more than two students. Generally, in group testing the group consists of three to six students (Centrella-Nigro, 2012; Leight, Saunders, Calkins, & Withers, 2012; Wiggs, 2011). The terms dyad or paired testing are utilized when students are paired with a partner to take an exam (Centrella-Nigro, 2012;

Haberyan & Barnett, 2010; Mitchell & Melton, 2003). Double-testing is another procedural method for administering a group test. In double-testing, a student first takes an exam as an individual and then re-takes same exam either with a partner or in a group. Within each of these methods of collaborative test administration, partner and group selection, as well as time allowed to take the test and methods for grading all vary. The variation of methods and procedures for implementing collaborative testing makes comparisons of collaborative testing outcomes difficult. Collaborative testing is a fairly new concept with little research to clearly demonstrate that this testing method improves student learner outcomes and test taking performance in higher education. (Haberyan & Barnett, 2010; Mitchell & Melton, 2003; Sandahl, 2009; Sandahl, 2010).

Active Learning

Active learning is a teaching concept that promotes student engagement. It is often reported that a collaborative learning environment provides more effective learning. Collaborative testing can be a component of active learning where a small group of students actively engage in learning through testing. Cooperative and collaborative learning are subsets of active learning in which students work in groups, teams, or pairs to perform complex tasks. This collaboration encourages active engagement in the learning process as group members discuss test questions, debate and problem solve to determine the best answers and communicate supporting reasons for their particular answer. Because collaborative testing is accomplished by groups of students, the literature often refers to collaborative or cooperative testing as group testing (Jensen, Moore & Hatch, 2002; Sandahl, 2010; Vockell,1994; Wilder, Hamner & Ellison, 2007).

Rationale for Collaborative Testing

Collaborative testing may provide a mechanism that more accurately evaluates students' knowledge and understanding of course material. Testing, especially high stakes testing, often causes test anxiety for the student. Test anxiety tends to weaken students' test taking ability and ultimately their overall grade in a course. Anxiety interferes with the students' ability to recall knowledge about the material being tested consequently leading to poor test performance (Markman, Balik, Bercovitz & Ehrenfeld, 2010). Collaborative testing, where students' can look to their peers for answers or for verification of their own answers, can decrease pressure and anxiety allowing improved recall of material and improved test scores. (Mitchell & Melton, 2003).

Through active peer collaboration during testing, both high and low-performing students can complement each other's knowledge allowing the combined effort to improve the group's understanding and application of course material (Lusk & Conklin, 2003). Collaboration during testing also alters the concept of competition among the students, encouraging them to work as a team supporting each other's success. The idea that a higher performing student is depending on a lower performing student to contribute to each other's success, may positively impact the lower performing student's attitude about his or her own abilities

(Bransford, Brown & Cocking, 2000). Positive self-concepts and test scores ultimately lead a student to improved knowledge and desire to learn and be successful (Phillips, 1988).

Prior Reviews on Collaborative Testing

Analysis of collaborative testing research is difficult for many reasons. Collaborative testing research focuses on a variety of outcomes, including student satisfaction, anxiety, course and end of semester grades, retention of short-term and long-term knowledge, and effects on teamwork. Furthermore, studies reveal multiple methods and procedures for administering and scoring tests that are taken collaboratively. Examples of these variations include differences in group size and group makeup, timing of tests with regards to administration procedure and placement within course, and determination of grades. Collaborative testing research methods also vary widely in the number of students used in the studies, student grade level, type of exam administered (multiple choice, essay, etc.), and type of class or course in which the testing is used.

There are debates regarding the advantages and disadvantages of collaborative testing. It has been shown that students perceive that they learn better in collaborative testing and exhibit improved individual test scores (Leight et al., 2012; Sandahl, 2009). Other reported advantages of collaborative testing include, but are not limited to, better critical thinking skills, improved collaboration and team work among peers, enhanced learning and test taking skills, motivation to study material, reduced test anxiety, and improved test taking performance and confidence.

Conversely, researchers report that collaborative testing may artificially raise grades for students who otherwise would not pass a course (Jensen et al., 2002; Mitchell & Melton, 2003). Investigators have also demonstrated that collaborative testing does not result in increased learning and retention of course material. (Giuliordori, Lugan, & DiCarlo, 2008; Harton, Richardson, Barreras, Rockloff & Latane, 2002; Leight et al., 2012). Educators describe disadvantages of collaborative testing such as poor effects on student study time and true comprehension of material. Students may not spend the necessary time in individual study if they know that their group may help them to pass the exam. Other disadvantages include lack of student preparation for the group, thus making the overall group score decrease as well as student reports of knowledge insecurity. This insecurity arises from student lack of preparation for the exam, yet passing the exam without truly knowing the material. Internal group conflict is another disadvantage that may negatively impact exam scores as students may be pressured to change originally correct answers. Lastly, long-term retention of material was identified as a disadvantage in that students who tested collaboratively may not ultimately retain the material for a comprehensive individual exam (Lusk & Conklin, 2003; Sandahl, 2009).

Enhanced Learning Opportunities

Several studies have demonstrated that collaborative testing enhanced student learning, validated misconceptions of testing material among peers, and improved test scores. Mitchell and Melton (2003) used collaborative testing for a unit exam on fluid and electrolytes in which students in an Associate Degree nursing program were randomly assigned to groups. The students who volunteered to participate took the exam individually and were allotted 50 minutes for a 50 –item multiple-choice exam. After submitting the individual test, a second test form was given to student groups who were allowed 10 minutes to discuss the questions on the exam and permitted to change their original answers. The second test form permitted an assessment on whether collaboration resulted in a change in the student's original exam grade. Averages of both the individual and group grade were used to determine the final test score. Study authors found through student surveys that collaborative testing helped to validate knowledge and clarify misconceptions. However, they found that collaboration gave unprepared students the opportunity to receive a higher grade than the one they received individually.

Haberyan and Barnett (2010) examined the efficacy of collaborative testing in an Educational Psychology class. Students completed the first and final exam individually and had the option of completing the third exam collaboratively. The researchers found that those testing with a partner scored significantly higher (M=9.63, SD 2.20) than those testing alone (M=8.15, SD=2.60). A second part of the study was used to determine effects of exam performance conditioned on whether students studied with a partner. Study authors found that exam performance was improved with collaboration, regardless of whether the student studied alone or with a partner.

Sainsbury and Walker (2008) studied group testing with students in a Bachelor of Pharmacy program in which thirty-five percent of the course grade was based on quizzes. In the study, quizzes were administered individually, then to a group. After discussion with their peer partners, students were permitted to either keep their initial answers or change them. Immediately after the quiz was submitted, course faculty allowed students to continue to engage in discussion about the quiz to enhance learning of material. Giuliodori, Lugan and DiCarlo (2008), studied students in a Veterinary Physiology course who took individual tests then answered the same questions in teams of two. Students who tested in a group answered questions correctly 70.2% of the time while students who tested individually responded correctly 58.7% of the time [t = 11.4, df = 137, p < .001, 95% CI: 1.88-2.67].

In a study by Zimbardo, Butler and Wolfe (2003) students in a psychology course were given the option of taking their midterm or final examinations with a partner of their choice. The final grade obtained by the team was considered the common grade. The researchers found that students' performance on the team exams was better than those who took individual exams yielding means of 73.1 versus 64.6 (t =7.25, df = 202.74, p < .0001). When repeated with students in a comparable course the following term, authors found nearly identical

results. Students also reported that they experienced reduced test anxiety, that the collaboration gave them practice with negotiating differences with answers, and that the experience enhanced learning and comprehension of material.

Several studies have demonstrated that collaborative testing helps to enhance all student test scores within a course. Wink (2004) studied collaborative testing for two exams in a Health Care Policy course in which students were placed in groups of six to ten. Students took the exam individually, and then again as a group. Scores were computed using an average of the individual and the group results, and demonstrated that student scores on both exams increased along with their final course grades.

Potential Barriers to the Use of Collaborative Testing

Researchers discuss grade inflation as a potential barrier to the use of collaborative testing (Wink, 2003; Wilder et al, 2007). To prevent grade inflation of poor performing students, different approaches have been reported by various researchers. Wilder et al. (2007) used an average of the individual and group grade to obtain the students final grade. Students who did not receive a passing grade on the individual exam were permitted to participate in the group test activity, but could not elevate their grade based on the group grade. Jensen et al., (2002), for a second quiz in an Anatomy and Physiology course, had students meet in their group where they were given one copy of the questions and one answer sheet to turn in. The researcher recommended a scoring rubric with points assigned for ranges of correct answers. As an example, a group grade of 8 to 10 out of 10 possible points resulted in 5 points added to their final test score. Their overall mean scores for students were all higher when compared to those taking individual quizzes (t = 20.3, df = 405, p < .0001). Overall, both studies found that cooperative testing produced better performance on exams than individualized exams, but did not greatly impact students' final grades.

Another barrier to the use of collaborative testing is the concern that performance on group exams does not translate to similar performance on comprehensive exams and retention of material. Woody, Woody and Bromley (2008) studied students who participated in an individual retest approximately three weeks after their initial group exam. The students formed groups of three or four in advance of the group examination. They found that students scored higher in a group format test (M=84.1%, SD=8.22) than in the individual format (M=75.5%, SD=12.56); however, collaborative testing did not result in later retention of course material. Leight et al., (2012) determined that the vast majority of students scored higher on group exams than on individual exams. In this study, students were semi-randomly divided into two groups based on their performances on the first exam. The second exam, which included class content from the first exam, was taken individually and then again in small groups of two to four. The research revealed that the positive effect of collaborative testing did not result in increased retention of content. Lastly, Sandahl (2010) randomly assigned students to groups of three or four, in which two exams were administered to a group, and two to individuals. Students in the group exam were not

required to reach consensus on the answers. The researcher found that students who took the exams collaboratively scored significantly higher than those who took the exam individually ($p < .05$). However, there was no significant effect found on retention of material (Lamda (12, 262) = .921, p=.526).

Contributions of This Review

This review intends to determine whether collaborative testing in the classroom setting produces positive learner outcomes in the higher education setting. Not only will this review identify whether learner performance will improve students' class grades when measured against their individual grades, but it will also assess student's perception of anxiety during test taking in either individual or collaborative format. The review will also evaluate the students' evaluations of the collaborative process, their perceptions of confidence with test taking skills, and their overall perceptions of comprehension and retention of material.

It is important that classroom teaching and learning continue to evolve. A greater emphasis has been placed on teamwork in the workplace. Collaborative learning may be one mechanism that can enhance the concept of teamwork. This review intends to contribute to the growing literature on collaborative testing by demonstrating its impact on teamwork. As the classroom expands and the learners change, new and creative formats of teaching and learning are also required. Although collaborative testing is a fairly new method of learning, this review will contribute to the growing body of literature and provide relevant information for both faculty practice and student learning.

OBJECTIVES OF THE REVIEW

It is the aim of this review to assess the effect of collaborative, group, or double testing on learning outcomes for students in higher education settings. Specifically, this systematic review asks the following research question:

What is the effect of collaborative testing on learning outcomes for higher education students?

METHODS

Inclusion Criteria

Intervention Characteristics

This review will look at the evidence on collaborative testing, defined as instances where two or more students work together to complete an evaluative exam.

Participant Characteristics

This review will only include adult higher education students, who are defined as students aged 18 years or older attending post-secondary institutions. We will not exclude studies on the basis of their participants' socioeconomic status, gender, race or any other demographic variable.

Outcome Characteristics

This review will have the following primary outcomes:

- learner class performance as measured via examination grade
- learner class performance as measured through final class grade
- long-term retention as measured through subsequent individual examination on course material
- student's perception of anxiety during test taking in any evaluative format

This review will have the following secondary outcomes:

- individual learners' test and course scores and individual learner's GPA
- grade inflation as measured by a student passing a course who otherwise would not have passed the course
- other described measures of grade inflation
- student evaluation of collaborative process
- student perception of confidence with test taking skills
- student perception of comprehension of material

Study Design

All comparative quantitative study designs with at least two time points for measurement of the outcomes—including but not limited to randomized control trials and quasi-experimental studies, pre-post and time series designs with control groups, as well as observational studies with control groups, such as longitudinal prospective or retrospective cohort studies and case control studies—will be included in this review. The use of these study designs in systematic reviews of comparative effectiveness is an acceptable strategy per the IOM (2011) and Campbell (2013) guidelines, particularly in the context of the GRADE (Guyatt, Oxman, Schünemann, Tugwell, & Knottnerus, 2011) approach for evaluating the quality of the evidence.

Quantitative study designs that do not include a comparator/control group will be excluded, such as correlational designs, descriptive studies, or single cohort prospective or retrospective studies. Before-after studies in which participants serve as their own control will not be excluded from the analyses. The manner in which the baseline equivalence of the comparator group was established, or whether baseline equivalence was established at all, will not be a consideration in the selection of studies for inclusion.

Search Strategy

Strategy Development

The search strategy will be developed in three stages. First, keywords derived from the study's inclusion criteria will be combined using Boolean operators and used to identify potentially relevant studies using the PsycINFO database (see Appendix I for this list). Citation chaining will also be used to identify additional relevant studies as well as those citing studies comprising the initial subset. Second, indexing of the identified citations will be examined for subject headings and relevant keywords capable of being added to terms of the search. Finally, the revised search strategy will be executed in the individual databases using controlled vocabulary including database-specific subject headings derived from database thesauruses and the revised list of keywords.

The assistance of a librarian with experience supporting systematic reviews will be sought in searching published, peer-reviewed literature using validated search strategies for the following electronic databases: MEDLINE, Embase, Academic Search Premier, CINAHL, PsycINFO, ERIC and VET-Bib. In addition, the ProQuest Dissertation & Theses database and a general internet search engine (i.e., Google) will be used to identify literature not formally published in sources such as books or journal articles (i.e., grey literature) using a combination of keywords and outcomes of interest. Pubmed will also be searched for publications ahead of print. Databases will be searched from inception to the present date. Search results will be merged using appropriate systematic review software and duplicate records of the same report will be removed. Our search will not employ hand searching of individual journals. We will however contact investigators to retrieve papers for which only abstracts are readily available. We will also contact experts within the field with queries regarding other potentially relevant studies that they may be aware of. Finally, the search employed in the review will be updated prior to submission for publication to ensure that the most recent data is captured.

Information Retrieval

Title and Abstract Screening

Two reviewers (RC and JS) will examine the titles and abstracts of the studies that are identified from implementation of the search strategy, and will determine whether these studies meet the inclusion criteria of this review. Disagreements between these reviewers

will be resolved first via consensus among them with by consultation of a third reviewer (JP) employed in irreconcilable cases.

Full-text Screening

The full text of studies that pass the initial screening by title and abstract will then be examined by two reviewers (RC and JS) in order to ensure that they meet the inclusion criteria. Disagreements between these two authors in this process will be resolved first by discussion, or if necessary, in consultation with a third reviewer (JP).

Assessment of Methodological Quality

Studies that meet the inclusion criteria will be independently critically appraised by two reviewers (RC and YTJ) in order to evaluate their methodological quality. We will determine the methodological quality of the studies by performing a risk of bias assessment, as prescribed by the Cochrane Handbook for Interventions Reviews (Higgins, 2011) for experimental studies. Observational studies will be critically appraised using the CASP critical appraisal tools for observational studies (http://www.casp-uk.net/#!casp-tools-checklists/c18f8). Risk of bias assessment entails a determination by the reviewers of the extent to which a study was performed such that the risk of systematic bias is reduced. This includes:

1. Randomization (including allocation concealment and random sequence generation) to address the risk of selection bias

2. Blinding (of the participants, investigators or providers, and outcomes assessors) to address the risk of performance bias and outcome assessment bias

3. Use of an intention-to-treat analysis to address the risk of attrition bias

The selected critical appraisal tools provide means of examining whether or not the study investigators attempted to deal with these inherent methodological flaws in observational designs, such as the use of matching as a sampling strategy to account for known confounders, or statistical testing on demographic characteristics to identify potentially significant differences between the comparator groups. The presence of these strategies will be identified and examined as part of the methodological appraisal of the studies that meet the inclusion criteria. We will use the GRADE (Guyatt et al., 2011) approach to downgrade or upgrade the evidence from experimental studies or observational studies, respectively. We will not exclude any studies based on ratings on individual risk of bias items or global study ratings of risk of bias.

Any disagreement between the two reviewers will be resolved first through discussion, then if necessary, in consultation with a third reviewer (JS).

Data Extraction

Data will be extracted using a de novo researcher-developed tool in Microsoft Excel (Appendix II.), which will be pilot tested on 10% of the included studies between the two reviewers who will be performing the data extraction (RC and JS). Disagreements between reviewers will be resolved via consensus or in consultation with a third reviewer (YTJ). The data that will extracted will include information on:

Study characteristics: full citation details, study design

Population characteristics: post-secondary category of student, age, male-to-female proportion, race/ethnicity, discipline

Interventions: type of collaborative testing (for example, test-retest: individual, then group testing- immediate; individual, then group testing delayed timing or group testing without an individual test), nature of the comparator

Outcomes: type of outcome, description of how the outcome is measured (i.e. what type of instrument was used to measure the outcome, if any).

Results: information on the effect size for each outcome (ratio of risks/rates/proportions, i.e., relative risk and odds ratio for dichotomous outcomes; mean difference for continuous outcomes), sample size in each group, measure of variability provided in the study (p-value, confidence interval, standard error or standard deviation)

Data Synthesis

Measures of Treatment Effect

Observational and experimental studies will be pooled separately for meta-analysis. We will express continuous outcomes (e.g., mean examination scores, differences in mean grade point averages; changes in anxiety scores measured using validated scales) as means, or standardized mean differences (SMD) to allow comparability of studies using different outcome scales. Categorical outcomes such as grade inflation (when evaluated as the proportion of students who pass the course who otherwise would not have) will be expressed as relative risks or odds ratios. The variability of all outcomes will be reported using the 95% confidence interval. We do not intend to meta-analyze students' ratings of confidence, the collaborative process, or perceptions of comprehension on scales used within individual studies. If these outcomes are dichotomized, we will combine results to generate a pooled odds ratio.

Calculating Effect Size

Studies that are included in this review will be combined into a meta-analysis for each outcome of interest in this review if the following criteria are met:

- If the statistical heterogeneity between studies is small, as measured by an I square value of 50% or less, and
- If there is clinical homogeneity between the studies, i.e. the interventions and comparators are similar enough across studies to justify pooling them.

We will also combine studies employing disparate methods of testing (e.g., double testing, dyad testing, and group testing) separately in the meta-analysis. Finally, narrative synthesis will be employed where between-study heterogeneity precludes meta-analysis and for outcome variables that are not amenable to meta-analysis (e.g., student perceptions). Where narrative synthesis is indicated, thematic analysis will be employed to identify salient insights prevalent among multiple studies. Studies evaluating the outcomes of interest using disparate methods (unable to be transformed into equivalencies) will either be combined separately for meta-analyses or summarized narratively using the aforementioned methods. Further, meta-analysis will not be conducted in the event that the majority of studies are deemed to present a high risk of bias on study characteristics that would limit their ecological validity.

Analyses will be conducted in Review Manager using the random effects model described by Dersimonian & Laird (1986). We undertake random effects analysis for multiple reasons. To start, the literature on collaborative testing shows a range of effects on student performance; we assume that the intervention effect described by these studies is therefore not identical but rather falls along a distribution. Second, we expect some between-study variance (i.e., individual study estimates of the effect size) to vary by chance; random effects modelling accommodates for this. Third, we expect some heterogeneity to be attributed to study design characteristics – such as participant selection methods, method of collaborative testing employed, group size, subject area. These factors suggest that a more conservative meta-analytic approach is warranted (i.e., a random effects analysis). Finally, in the absence of true between-study heterogeneity, the random effects model is reduced to the fixed-effects model.

Sensitivity Analysis

We will perform a sensitivity analysis based on the risk of bias criteria for each outcome across studies. For each risk of bias criterion, we will exclude studies that are found to be at high risk of having a systematic bias, and we will repeat the meta-analysis without these studies in order to determine whether that type of risk of bias has an impact on the interpretation of the meta-analysis. Further sensitivity analyses will be conducted to account for missing data (see Missing Data).

Publication Bias

Publication bias will be investigated if there are at least 10 studies available to pool in the meta-analysis. We will assess publication bias visually using funnel plots generated in

Review Manager using current best practices (Higgins, 2011). We will not formally assess publication bias (e.g., using Egger's test).

Missing Data

The reviewers will make a reasonable attempt to contact study authors in order to retrieve missing data. We will examine whether the data that is missing appears to be missing at random, or whether it is not missing at random but instead is likely due to attrition bias and selective outcome reporting. No studies will be excluded from the meta-analysis on the basis of missing data unrelated to the outcomes of interest. Sensitivity analyses will be concluded in which studies with non-random missing data are subsequently excluded to examine the effects on summed estimates of effect size.

Subgroup and Moderator Analyses

Post hoc subgroup analyses will not be conducted. We will employ meta-regression to examine the potential impact of the following variables, which we are identifying a priori, on the effect size of the meta-analyses: mean age of the participants, gender, study discipline (e.g., health sciences, humanities, mathematics), collaborative testing group size, and methods of testing (e.g., multiple choice, short answers, presentations, objectively structured clinical exams). We will also use meta-regression to examine the impact of each type of systematic bias identified in the studies on the pooled effect size for each outcome across studies. Meta-regression models will be generated in Stata 13.1; moderator variables will be examined individually as covariates.

It has previously been suggested that maturational effects may contribute to collaborative testing performance as learner test-taking skills improve (Sandahl, 2009). It follows that older learners may have had more opportunity to develop adaptive learning strategies than novices in higher education settings; the impact of learner age on effect sizes may therefore warrant exploration. Breedlove, Burkett and Winfilied (2007) reported that despite finding no correlations between gender and preference for collaborative testing over individual work, male students were more likely to perceive collaborating with peers as reducing test-associated stress. If lower anxiety and stress leads to higher test scores, then male students may, on average, be more likely to receive the benefit of collaboration. Thus, exploring the impact of gender on effect sizes way also be warranted. Certain types of examination questions may find themselves more amenable to student collaboration than others and consequently impact both learning and student performance. Yet, individual studies typically examine the impact of collaborative testing on only a single type of examination. Meta-regression analysis would allow for comparison of different examination characteristics on effect sizes. Along, the same logic, it seems reasonable to assume that course content may be an important moderator variable and as such will also be examined.

REFERENCES

Bransford, J.D., Brown, A.L., & Cocking, R.R. (eds.) (1999). *How people learn: brain, mind, experience, and school.* Washington DC: National Academy Press.

Breedlove, W., Burkett, T., & Winfield, I. (2007). Collaborative testing, gender, learning styles, and test performance. *International Journal for the Scholarship of Teaching and Learnin,g 4*(2), 33-42.

Centrella-Nigro, A. (2012). Collaborative testing as a post-test review. *Nursing Education Perspectives,* 33(5), 340-341. http://dx.doi.org/10.5480/1536-5026-33.5.340

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177-188.

Gaberson, K. (November 11, 2008). *Evaluation and testing*. [Videotape]. (Available from Duquesne University School of Nursing. Retrieved from: http://mslweb.cr.duq.edu/nursing/Viewer/?peid=f504677a-6ec9-45e9-85f6-c80a3a66228e).

Giuliodori, M., Lujan, H. & DiCarlo, S. (2008). Collaborative group testing benfits high and low- performing students. *Advances in Physiology Education,* 32, 274-278. doi: 10.1152/advan.00101.2007.

Guyatt, G.H., Oxman, A.D., Schünemann, H.J., Tugwell, P, Knottnerus, A. Grade guidelines: A new series of articles in the Journal of Clinical Epidemiology. *Journal of Clinical Epidemiology* 2011;64(4), 380-2. doi: 10.1016/j.jclinepi.2010.09.011

Haberyan, A. & Barnett, J. (2010). Collaborative testing and achievement: Are two heads really better than one? *Journal of Instructional Psychology*, 37(1), 32-41.

Halstead, J. (2007). *Nurse Educator Competencies: Creaeting an Evidence-Based Practice for Nurse Educators*. New York: National League for Nursing.

Higgins, J.P. & Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [March 2011]. Cochrane Collaboration. 2011: Available from www.cochrane-handbook.org.

Hoachlander, E. G. (1998). Is there a “best” way to test? Assessing assessment. *Techniques: Making Education And Career Connections*, *73*(3), 14-16.

IOM (2011). Finding what works in healthcare: Standards for systematic reviews. http://www.iom.edu/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews/Standards.aspx. Accessed on February 28, 2013. In. 2011.

Jensen, M., Moore, R., & Hatch, J. (2002). Cooperative Learning- Part I: Cooperative Quizzes. *The American Biology Teacher, 64*(1), 4. http://dx.doi.org/10.1662/0002-7685(2002)064[0029:CLPICQ]2.0.CO;2

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 Scieneces Education,* 11, 392-401. Doi: 10.1187/cbe.12-04-0048

Lusk, M., & Conklin, L. (2003). Collaborative testing to promote learning. *Journal of Nursing Education, 42*(3), 121-124.

Markman U, Balik C, Braunstein-Bercovitz H Ehrenfeld M. (2010). The effects of nursing students' health beliefs on their willingness to seek treatment for test anxiety. *Journal of Nursing Education, 50,* 248-251.

Mitchell, N., & Melton, S. (2003). Collaborative testing: An innovative approach to test taking. *Nurse Educator, 28*(2), 95-97.

Sainsbury, E., & Walker, R. (2008). Assessment as a vehicle for learning: extending collaboration into testing. *Assessment & Evaluation in Higher Education, 33*(2), 15. doi: 10.1080/02602930601127844

Sandahl, S. (2009). Collaborative testing as a learning strategy in nursing education: A review of the literature. *Nursing Education Perspectives,* 30(3), 171-175.

Sandahl, S. (2010). Collaborative Testing. *Nursing Education Perspectives, 31*(3), 142-147.

The Campbell Collaboration. Campbell Collaboration Systematic Reviews: Policies and Guidelines Campbell Systematic Reviews 2013: Supplement X. doi: 10.4073/csr.200x.x

Vockell, E. L. (1994). The group retest: A route to effective cooperative learning. *Contemporary Education, 66*(1), 25. Retrieved from: http://search.proquest.com/docview/1291708905?accountid=13626

Wilder, B., Hamner, J., & Ellison, K. (2007). Student perceptions of the impact of double testing. *Nurse Educator, 32*(1), 6-7.

Wink, D. (2004). Effects of Double Testing on course grades in an undergraduate nursing course. [Education measurement]. *Journal of Nursing Education, 43*(3), 138-143.

Woody,W., Woody, L. & Bromley, S. (2008). Anticipated group versus individual examinations: A classroom comparison. *Teaching of Psychology,* 35(1), 13-17. doi: DOI: 10.1080/00986280701818540.

Zimbardo, P., Butler, L., & Wolfe, V. (2003). Cooperative college examinations: More gain less pain when students share information and grades. *Journal of Experimental Education*, 71(2), 101-125.

SOURCES OF SUPPORT

None.

DECLARATIONS OF INTEREST

None to declare

EXPECTED TIMEFRAME

Work on this review will begin upon approval of the protocol by the Education Coordinating Group of the Campbell Collaboration and will be completed by June 2015.

APPENDIX I: INITIAL SEARCH TERMS

adult OR “higher education” OR “post-secondary” OR college* OR university*

and

“collaborative test*” OR “group test*” OR “double test*” OR “paired test*” OR “cooperative test*” OR “dyad test*”

and

(“Class performance” ***or*** “Class grade” ***or*** Anxiety ***or*** **“**Test score*” ***or*** **“**Grade point average” ***or*** GPA ***or*** **“**Grade inflation” ***or*** Confidence ***or*** Comprehension OR satisfaction OR perception

Class performance ***or*** Class grade ***or*** Anxiety ***or*** Test scores ***or*** Grade point average ***or*** GPA ***or***

Grade inflation ***or*** Confidence ***or*** Comprehension

APPENDIX II: DATA COLLECTION FORM

Part I – Information on the study					
Full citation of the study	**Geographic location study took place**	**Study design**	**Population characteristics**	**Sample size**	**Type of collaborative Testing, nature of comparator**

Part II – Results of the study				
Outcome measure	**Instrument**	**Measure of variability**	**Effect size**	**Misc.**