Learning during a collaborative final exam

Örjan Dahlström

Linköping University Post Print

N.B.: When citing this work, cite the original article.

This is an electronic version of an article published in:
Educational Research and Evaluation is available online at informaworld™:
http://dx.doi.org/10.1080/13803611.2012.674689
Copyright: Taylor & Francis (Routledge)
http://www.routledge.com/

Postprint available at: Linköping University Electronic Press
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-77640
Collaborative testing has been suggested to serve as a good learning activity, for example compared to individual testing. The aim of the present study was to measure learning at different levels of knowledge during a collaborative final exam in a course in basic methods and statistical procedures. Results on pre- and post-tests taken individually ($N=30$) before and after the collaborative part of the final exam confirmed learning effects at the uni- and multi-structural levels as well as on the relational level of Bigg’s structure of the observed learning outcome (SOLO) taxonomy. Low performers at pre-test raised their test scores more than high performers. No differences could be generalized at the extended level of knowledge. Results also suggest that it might be preferable to collaborate without first deciding on questions individually. The experimental design can be applied when evaluating learning, at different levels of knowledge, during a variety of learning activities.
Learning during a collaborative final exam

This study will examine learning *during* a collaborative final exam in a program in Speech and Language Pathology and will serve as an example of how teachers in higher education can evaluate learning effects of a given activity (in this case the collaborative final exam) in their course. The background of this study is the positive response from students to the final exam in a course in basic quantitative research methods and statistical procedures. For several consecutive years the students have mentioned in course evaluations that the very moment when they really learned the content of the course was during the collaborative final exam. This raised questions of whether this reported learning effect during the collaborative final exam was real or only perceived.

Collaborative testing

Learning through collaboration, i.e. why cooperation can improve learning has been explained by several factors, such as experimental learning, ‘social interdependence’, motivation and involvement, and engagement in ‘elaboration’ (Zipp, 2007). Several studies have examined the effects of collaborative exams (Stearns, 1996; Shen, Hiltz and Bieber, 2008; Eaton 2009; Meseke, Nafziger and Meseke, 2008; Zipp, 2007). These have however mainly focused on the effects of using collaborative exams regularly during courses as a preparation for a final exam (who was most often taken individually). Eaton (2009) used such a strategy during a semester in two geo-science courses. The strategy was shown to be successful as evaluated by results from the midterm and final exams taken individually. Stearns (1996) used three exams as learning tools. Each exam was first taken individually and thereafter as a group, followed by an explanation of any wrong answers to the group. A final exam was then taken individually and results were compared to results from a previous class where no such procedure was used. It is not clear whether the two classes had identical tests or not, but the mean score was
significantly higher for the class which took collaborative exams during the semester. Zipp (2007) reports learning effects in an introductory sociology class by a ‘pyramid exam’ where students first take an exam individually and thereafter the same exam together with their learning group, consisting of six persons stratified by gender. By discussing and debating exam answers in small groups the exam can become an active learning experience. Four such tests were given during the course followed by a final exam consisting of questions from the preceding four tests. The final exam was however optional and that might have meant that highly skilled students satisfied with their initial test scores did not attend. Across the four tests using the ‘pyramid exam’ there were only 3.8% of the students that scored better individually than their group. Meseke, Nafziger and Meseke (2008) used continuous tests (six weekly quizzes and three unit examinations) prior to a final exam during a course on a chiropractic program. Students were divided into two skilled-homogeneous groups where the treatment group was randomly divided into smaller subgroups that were able to discuss among each other during the quizzes while the controls had the quizzes individually. The unit exams were all taken individually. The treatment group showed significantly higher scores on all quizzes and higher scores on two out of three unit examinations. At the final examination there was, somewhat surprisingly, a better performance for the controls \((p = .056)\). Shen, Hiltz and Bieber (2008) involved students in online collaborative exams. Students worked together in groups designing questions and grades (answering was made individually) and the perceived learning increased. Cortright, Lollins, Rodenbaugh and DiCarlo (2003) used a randomized cross-over design to show that individual testing followed by collaborative testing, compared to individual testing only, increased performance on individual testing a few weeks later. In the individual follow-up tests a subset of the questions from the test a few weeks earlier was repeated and scores were compared.
There are several potential reasons for why cooperative testing works, such as reduction of test anxiety and stress, increased motivation to learn, and help (by discussions) to think at higher levels (Zipp, 2007).

**Group sizes**

Optimal sizes of groups for collaborative learning were examined by Oakley, Felder, Brent and Elhajj (2004) and team-sizes of three to five were recommended, although ‘there are obvious exceptions to these rules, such as laboratories with two-person work stations’ (Oakley et al., 2004, p. 12). In the studies using collaborative exams different group sizes have been used, although it is not always clearly specified, for example five or six persons (Stearns, 1996; Zipp, 2007), ‘small groups’ (Meseke, Nafziger and Meseke, 2008), two or three persons (Rao, Collins and DiCarlo, 2002) or unclear (Eaton, 2009).

**Taxonomy of learning outcome**

Learning is not always clearly defined in different studies, but is often explained in terms of measurement, such as by multiple choice questions (Cortright et al., 2003; Rao, Collins and DiCarlo, 2002; Meseke, Nafziger and Meseke, 2008), ‘fill in the blanks’ (Cortright et al., 2003; Rao, Collins and DiCarlo, 2003), ‘short-answer essay questions’ (Cortright et al., 2003; Rao, Collins and DiCarlo, 2003), ‘true or false’ (Rao, Collins and DiCarlo, 2003). Eaton (2009), however, refer to the six-graded Bloom’s taxonomy scale (Bloom, 1956) by stating that collaborative testing appears to provide learning at higher levels; raising measurement from the basic knowledge (1st grade) category to higher levels characterized by comprehension (2nd grade) of interpretation, application (3rd grade) to novel situations, and analysis (4th grade) of logical reasoning. The remaining grades described by Bloom are synthesis (5th grade) and evaluation (6th grade). Revere, Elden and Bartsch (2008, p.14) state that while it is ‘relatively easy to test information at lower levels of Bloom’s taxonomy (e.g.,
Knowledge); it is more difficult with this type of exam to test higher levels (e.g., synthesis)’. Collaborative testing has also been reported to improve performance in lower-level but not in higher-level exams (Breedlove, Burkett and Winfield, 2004a; 2005).

The relation of test performance to learning at different levels of knowledge will in this study rely upon Bigg’s structure of the observed learning outcome (SOLO) taxonomy (Biggs & Collins, 1982). It is described by five sequential levels of performance. The pre-structural level is when the content or point is not understood and reasoning is about getting around the question. The uni-structural is when one and the multi-structural is when several aspects have been understood and can be used separately. Explanations at these levels of performance are mainly quantitative. The relational level is when different explanations can be related to each other, such as by comparing profits and drawbacks with different approaches. This is what we normally mean by understanding of a topic. The extended level is when the previous understanding can be used in higher levels of abstraction and generalized to new areas or to new research topics. In comparison to the aforementioned Bloom’s taxonomy the SOLO taxonomy is more easily applied when scoring answers. While the use of the Bloom’s taxonomy will often make it difficult to classify answers from students to the correct grade, use of the SOLO taxonomy will make the classification more evident. The SOLO taxonomy is also organized in a learning progression manner, where the lower levels are assumed to precede the higher levels. However, this is not obvious in Bloom’s taxonomy.

Methodology and collaborative exams

Measuring performance in groups and comparing this to individual performance is not straight-forward. For example, Giuliodori, Lujan and DiCarlo (2008) compared individual scores with group scores to measure learning. There are studies showing that performance in groups fails to reach the full potential as given by the individuals’ individual performance (Dahlström, Danielsson, Emilsson and Andersson 2011; Weldon and Bellinger, 1997), which
in this case would mean that their group scores probably are lower than the sum of the groups’ non-redundant individual scores would be. Also, comparisons of individual scores with collaborative scores to compare learning, rely upon the assumption that all individuals fully agree and understand the answers given collaboratively. In other words, measures of individuals' learning based on performance in groups are often associated with problems of measurements.

Another difficulty is how to design studies without introducing learning effects, for example by repeating the very same question again. If different sets of questions are used in two tests and scores compared one has to validate that the questions of the two tests were of comparable difficulty. In the otherwise beautifully designed study by Cortright et al. (2003) questions were repeated and it cannot be excluded that repeating the questions introduced a considerable learning effect.

In studies where exams are used regularly during the semester the increased knowledge for students taking the exams collaboratively can sometimes also be attributed to feedback given by the teacher (e.g. Stearns, 1996; Zipp, 2007). To elicit the effect of the collaborative exam one would also need a group taking collaborative exams but without the following feedback from the teacher, or alternatively a group taking individual exams followed by the same feedback from the teacher as given to students taking collaborative exams. Nevertheless, those studies indicate that collaborative exams taken regularly during a semester together with feedback given by the teacher have a positive effect on learning.

Most studies report how they operationalize the measurement of knowledge, but there are rarely reports of how different levels of knowledge are affected by collaborative exams. Eaton (2009) does, however, acknowledge Bloom’s taxonomy, but does not offer specific measures of knowledge at different levels.
Even though many studies have examined learning during collaborative exams none appear to have examined learning during a collaborative final exam. The main aim of this study was to measure the amount of learning at different levels of knowledge during a collaborative final exam. A secondary aim was to provide an example of how teachers in higher education can evaluate the amount of learning during a learning activity (in this case a collaborative final exam) by applying a rather straight-forward experimental design.

**Methods**

*Participants*

Thirty students (24 females) in their third year of a Program of Speech and Language Pathology participated in the study (mean age = 23.7, SD = 1.5). They were all part of a course in basic quantitative methods and statistical procedures.

*Material*

The collaborative final exam consisted of two sets of questions constructed to elicit answers relating to different levels of the SOLO taxonomy. The first part was designed to simulate a part of a real research process, the analysis and interpretation part. It was accompanied by an electronic dataset and had to be answered in pairs using IBM SPSS 19. The second part had to be answered individually but students were allowed to discuss the questions with each other. Students were allowed to use all kinds of references and to be at any place at or outside the University Campus. The time limit for the first and second parts of the main exam was three and four hours respectively.

In the experimental design applied to measure learning at different levels of the SOLO taxonomy pre- and post-tests (see Appendix for examples of questions) were performed immediately before and immediately after the collaborative exam. The pre- and post-tests
were based on a base set of questions constructed to measure knowledge at three different levels based on the SOLO taxonomy; (1) uni- or multi-structural, (2) relational and (3) extended. The base set of questions were constructed by identifying different parts of the course content and based on them formulating a cluster of two, three or four questions for each such part. The ambition was to have questions of as similar difficulties as possible within each such cluster. The questions were carefully formulated to not coincide with questions from the collaborative exam. (After the exam, students were provided feedback to incorrect answers and to pass the course they had to leave additional answers, both for the collaborative exam and for pre- and post-tests.)

**Design**

The experimental design was a pre-test post-test design. The questions were uniquely organized for every individual test by a pseudo-randomization procedure organizing each test to be of similar difficulty with equal number of questions from each cluster within the base set of questions. The post-test contained more questions, i.e. some questions were repeated while others were new (Table 1). This procedure made it possible to design pre- and post-tests of similar difficulty while also testing for ‘repeating-questions-effects’. The scoring of the answers was made question-wise following a pre-defined set of requirements for what should be treated as correct answers. The score relative to maximum score was calculated for each of the three levels of knowledge and for each individual test.

Low and high performers were identified twice, first for uni- and multi-structural questions and thereafter for relational questions. A low performer was defined as a student having at most 50% correct answers at pre-test for the respective knowledge level.

To examine if classification of low and high performers at pre-test could be dependent on the specific set of questions on their individual tests the ratings of all questions were calculated (i.e. the mean score for that question for all answers in the study). Thereafter, for
each specific pre-test, the mean of the ratings was calculated, resulting in a measure of difficulty for each individual pre-test. Differences in the difficulty of questions between low and high performers were compared.

Table 1.

Number of questions constructed in total (Questions), and number of questions used at pre-test and at post-test (repeated questions in parenthesis) at different levels in the SOLO-taxonomy.

<table>
<thead>
<tr>
<th>Level</th>
<th>Questions</th>
<th>Pre-test</th>
<th>Post-test (repeated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uni- and Multi</td>
<td>18</td>
<td>9</td>
<td>12 (3)</td>
</tr>
<tr>
<td>Relational</td>
<td>3</td>
<td>2</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Extended</td>
<td>4</td>
<td>2</td>
<td>2 (0)</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>13</td>
<td>16 (4)</td>
</tr>
</tbody>
</table>

See appendix for example of questions.

Statistics

Descriptive statistics were given by median values, but for the sake of completeness also by mean values and standard deviations. Comparisons were made by non-parametric statistics. Comparisons between different groups were made by Mann-Whitney U test (reported by U-statistic and the corresponding z-value to show departure from null-hypothesis of ‘no difference’) and comparisons between repeated measures were made by Wilcoxon signed rank test (reported by the corresponding z-value to show departure from null-hypothesis of ‘no difference’). Each comparison was attached by p-value and also by r-value as a measure
of effect size (rule of thumb: \( r > .1 \) is small, \( r > .3 \) is medium, and \( r > .5 \) is large). All statistics were performed using the IBM SPSS Statistics 19 for Windows.

**Results**

Table 2.

Performance at pre- and post-test at the uni- and multi-structural, at the relational and at the extended level of the SOLO-taxonomy. (Mean and SD are given for the sake of completion).

<table>
<thead>
<tr>
<th>Level</th>
<th>Pretest</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median</td>
<td>Mean (SD)</td>
</tr>
<tr>
<td>Uni- and Multi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>0.62</td>
<td>0.60 (0.18)</td>
</tr>
<tr>
<td>Repeated</td>
<td>0.67</td>
<td>0.59 (0.26)</td>
</tr>
<tr>
<td>New</td>
<td>0.63</td>
<td>0.61 (0.20)</td>
</tr>
<tr>
<td>Relational</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>0.63</td>
<td>0.61 (0.22)</td>
</tr>
<tr>
<td>Repeated</td>
<td>0.63</td>
<td>0.58 (0.32)</td>
</tr>
<tr>
<td>New</td>
<td>0.75</td>
<td>0.63 (0.30)</td>
</tr>
<tr>
<td>Extended</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>0.50</td>
<td>0.43 (0.29)</td>
</tr>
</tbody>
</table>

**Pre- and post-tests**

Scores for pre- and post-test are shown in Table 2. Scores for uni- and multi-structural questions were increased during the collaborative exam, \( z = 2.96, p = .003, r = .54 \) (median increased from 0.62 to 0.70). Looking at new and repeated questions separately showed the same pattern but more evident for new questions, \( z = 2.40, p = .02, r = .44 \) (median increased from 0.63 to 0.72), than for repeated questions where the increase was not significant, \( z = \)
1.87, \( p = .06, r = .34 \) (median remained the same, but mean values increased).

Scores for relational questions also increased, \( z = 2.83, p = .005, r = .52 \) (median increased from 0.63 to 0.75). For questions given only once scores also increased significantly, \( z = 3.02, p = .002, r = .55 \) (median increased from 0.75 to 1.00), but for repeated questions there were no significant change, \( z = 0.34, p = .74 \) (median in fact decreased from 0.63 to 0.50).

Scores for extended questions showed no significant differences between pre- and post-test, \( z = 0.33, p = .74 \).

**High and low performers**

Comparing low and high performers on their performance on pre- and post-tests for uni- and multi-structural questions (Figure 1) showed that low performers increased their performance on new questions, \( z = 2.31, p = .02, r = .77 \) (median increased from 0.35 to 0.50) as well as on repeated questions, \( z = 2.08, p = .04, r = .69 \) (median increased from 0.33 to 0.67). For high performers on pre-test there were no significant differences in performance for repeated questions (\( p = .57 \)), nor for new questions (\( p = .14 \)). Interestingly, while the new questions showed a minor increase in performance (median increased from 0.75 to 0.78) the repeated questions showed a decrease in performance (median decreased from 0.83 to 0.67). This difference was, however, not significant. In fact, at post-test the differences between high- and low-performers shown at pre-test were almost equalized.

Comparing performance on pre- and post-tests for relational questions for low and high performers separately (Figure 2) showed that low performers increased their performance on new questions, \( z = 2.50, p = .01, r = .69 \) (Median increased from 0.50 to 1.00), as well as on repeated questions, although that increase was not significant, \( z = 1.34, p = .18 \) (median increased from 0.25 to 0.50). The new questions were more accurately answered than the repeated questions at post-test, \( z = 2.01, p = .01, r = .56 \). For high
performers on pre-test there were no significant differences between pre- and post-test, neither for new questions, $z = 1.66, p = .10$, nor for repeated questions, $z = 1.67, p = .10$. New questions had a median of 1.00 for both tests, but scores on repeated questions decreased on post-test. In fact, at post-test the differences between high- and low-performers shown at pre-test were almost equalized.

![Graph showing performance on uni- and multi-structural questions at pre- and post-test. Significant differences are indicated by their $p$-values in figure. Low performers on pre-test ($N = 9$, 50% or less correct answers).](image-url)
Figure 1b.

Performance on uni- and multi-structural questions at pre- and post-test. Significant differences are indicated by their $p$-values in figure. High performers on pre-test ($N = 21$, more than 50% correct answers).

Figure 2a.

Performance on relational questions at pre- and post-test. Significant differences are indicated by their $p$-values in figure. Low performers on pre-test ($N = 13$, 50% or less correct answers).
Performance on relational questions at pre- and post-test. Significant differences are indicated by their $p$-values in figure. High performers on pre-test ($N = 17$, more than 50% correct answers).

**Impact of easy and difficult questions**

There were no significant differences in magnitude of individual pre-tests’ difficulty between low and high performers, neither for uni- and multi-structural questions, $t(28) = 1.55, p = .13$ (mean values $= 0.59$ and $= 0.61$ respectively), nor for relational questions, $t(28) = 0.03, p = .98$ (mean values $= 0.62$ and $= 0.62$ respectively). In fact, the differences were almost zero.

**Discussion**

The main aim of was to measure the amount of learning during a collaborative final exam by introducing a pre- and a post-test. Performance at the uni- and multi-structural level, as well as on the relational level, of the SOLO taxonomy was increased by the collaborative exam for
students performing low on pre-tests. Somewhat surprising was that the increased performance was more obvious for new questions than for repeated questions (indicated by larger effect sizes). One possible reason for this is that once students have taken a stand they are more inclined to maintain it even though there might be other arguments against it. It is possible that there are mechanisms other than the validity of different answers that decide what is treated as a correct answer. For example, a more dominant student with an initially wrong answer can sometimes persuade a less dominant student with an initially correct answer of what is correct. It is possible that those kinds of discussions took part during the collaborative exam.

There were no measurable increments in knowledge at the extended level. It is possible that measures at the extended level have to be more exhaustive and that learning at the extended level needs more time than a one-day exam to evolve. The test was due to time-limitations however not as exhaustive on the extended level, compared to the other levels. This might also have put limitations on the sensitivity of measuring knowledge at the extended level.

The low performers gained more from the collaborative exam than did high performers. This was not surprising since if collaboration means learning from each other, it is reasonable that low performers have more to learn from high performers than vice versa. An interesting pattern in this sample is that while both low and high performers did better on new questions, high performers did worse on repeated questions. The pattern for uni- and multi-structural, as well as for the relational level, showed that for repeated questions the differences at pre-test almost converged to a similar performance at post-test. Statistically, one could argue that the convergence of performance, for uni- and multi-structural as well as for the relational levels, at post-test is a statistical artefact (regression to the mean). That is, however, not the probable reason for the convergence between the low and high performers’
post-test results on repeated questions since there is no such tendency on the new questions. One possible explanation to this discrepancy is that students during the collaborative exam also discussed the questions at the pre-test and compared their answers without that much reflection. While, on the other hand, topics discussed more in general during the collaborative exam put things into a context without presumptions given by answers (correct or incorrect) to questions given on a pre-test. Having a false perception of something might in other words be more inhibiting than not having any opinion at all. In general, this could support the opinion that learning by pre-defined questions and answers (i.e. discussing different answers given during a pre-test) will not make students reflect upon the content as much as other stimulating tasks (i.e. a collaborative exam).

In terms of the SOLO taxonomy the collaborative exam increased learning at lower (uni- and multi-structural) level, but also at higher levels (relational). This differs from the results of Breedlove, Burkett and Winfield (2004a; 2005) who found no learning at higher levels (defined as “theory questions”). Results in the present study suggest that collaboration included discussions of details, but also discussion of how to relate different answers to each other. In the current form extended knowledge could not be concluded to increase. It is possible that generalizations at the extended level also were discussed during the collaborative exam but the pre- and post-test need to have a higher sensitivity to be able to measure that. The time of the collaborative exam might also need to be extended to more than seven hours (perhaps a few days) since extended thinking is way more complex and demanding than is thinking at the lower levels.

The benefit of this design is that every individual’s performance at pre- and post-test could be directly compared and thereby individual measures of learning could be estimated. The design also made it possible to compare the amount of learning for repeated as well as new questions at post-test. The design also made it possible to control for differences in
difficulty of the questions as a possible confounder. The amount of learning during the collaborative final exam was not affected by teacher feedback and therefore provides less confounded measure of collaborative learning compared to some earlier studies.

A limitation of the design is that there is no control for what the collaboration was like, i.e. no video recordings of who acted, what was said, how much different topics were discussed, how large were the groups collaborating, etc. The aim of the study was, however, not to examine different underlying mechanisms of collaboration but to examine the amount of learning at different levels of knowledge during a collaborative final exam. Another limitation is that it was not possible to include a control group taking the collaborative exam individually, since it would be unfair to give different students very different settings for the exam. Yet another limitation is that the retained knowledge in a longer perspective is not considered here. The within-group design and the possibility to compare differences between repeated and new questions were however strengths of the study and in contrast to some previous studies the amount of learning in this study was not confounded with feedback from the teacher.

One question that can be raised based on the results is: If the form of the final exam really contributes to learning, why not incorporate the form earlier in the course as a preparation for a final exam? The answer is that it could be a great idea. If it works, use it! What differs, however, is that it is not obvious that motivation during minor exams during the semester is the same as during a final exam. In this study it was examined if learning was achieved during a collaborative final exam, given whatever preconditions are present during final exams. A difference might be that the collaborative part reduced the test anxiety, although both pre- and post-tests were taken individually. Additionally, there are results indicating negative effects of regularly taken collaborative exams during the semester (Meseke, Nafziger and Meseke, 2008).
Finally, the second aim was to provide an example of how teachers in higher education can evaluate the amount of learning during a learning activity (in this case a collaborative final exam) by applying a rather straight-forward experimental design. The sample size was 30 but the design was nevertheless powerful enough to generate interesting results. In further studies one should aim for larger sample sizes, but the results of this study can hopefully encourage others to evaluate students’ learning even though the number of students are only as many as in this study. Hopefully also the design, analysis and discussion of the results will encourage others to measure learning, at different levels of knowledge, at different activities in higher education, not least in collaborative settings.

In further studies the collaborative approach to the final exam can be examined in other courses, with higher sensitivity of knowledge at the extended level and with more time of the collaborative exam and questions stimulating discussions about generalizations and extensions.

**Conclusion**

Learning at different levels of the SOLO taxonomy during a collaborative final exam showed a positive amount of learning at the uni- and multi-structural as well as at the relational levels of knowledge. Low performers at pre-test are the beneficiaries in increased learning. Interestingly, for repeated questions the initial differences between low and high performers at pre-test are almost equalized at post-test, but that was, however, not the case with new questions suggesting that it might be preferable to collaborate without first having made up one’s own mind. No differences could be generalized at the extended level of knowledge. The experimental design can be applied when evaluating learning, at different levels of knowledge, during a variety of learning activities such as for example collaborative final exams in other settings.
Acknowledgements

The author gratefully acknowledges Dr. Gisela Eckert for valuable input and support.

Notes on contributor

Örjan Dahlström is Senior Lecturer at the Department of Behavioural Sciences and Learning at Linköping University, Sweden. He holds a Master’s degree in Electrical Engineering and Applied Physics and a PhD in Disability research. He has a general interest in teaching in higher education. His main research interest is within the area of Cognitive Hearing Science.

References


