Effective Student Teams for Collaborative Learning in an Introductory University Physics Course

Jason J.B. Harlow, David M. Harrison, and Andrew Meyertholen

Department of Physics, University of Toronto, Toronto, ON M5S 1A7

We have studied the types of student teams that are most effective for collaborative learning in a large 1st year university physics course. We compared teams in which the students were all of roughly equal ability to teams with a mix of student abilities, we compared teams with 3 members to teams with 4 members, and we examined teams with only one female student and the rest of the students male. We measured team effectiveness by the gains on the Force Concept Inventory. None of the factors that we examined had significant impacts on student learning. We also investigated student satisfaction as measured by responses to an anonymous evaluation at the end of the term, and found differences depending on how the 9 teams in the group were constructed.

PACS Number: 01.40.Fk

I. INTRODUCTION

Physics Education Research (PER) has led to an increased emphasis on collaborative learning in reformed-pedagogy physics courses around the world. Some of the best-known examples are Peer Instruction, McDermott’s Tutorials in Introductory Physics and Laws’ Studio Physics. Therefore, questions about how to structure teams of students for collaborative learning to achieve the best possible outcomes are increasingly important.

Psychologists investigate important questions about the types of collaborative learning that are most effective, and their research has a number of different approaches to these questions. Within the PER community, there is a long tradition of another approach to these questions: videotaping, transcribing, and analyzing student interactions. This methodology is echoed in more recent studies of teams using sociometric badges. These devices, about the size of a small TV remote or a classroom clicker, are worn by members of a team and measure the amount of face-to-face interaction, body language and orientation, dynamics of conversation often without actually recording the actual words being said, and similar properties. The devices and analysis of the “big data” that

1 Email: david.harrison@utoronto.ca
they produce has successfully been used to measure the characteristics that lead to effective teams in a commercial context, where “effective” is defined as getting a piece of research done, or a product marketed, or similar such tasks. Using these devices and methodology to study effective teams for learning would be an interesting project.

In this study we instead ask some comparatively simple questions about teams of students engaged in collaborative learning. First, should the teams of students be sorted by student ability, or instead should the teams contain a mix of strong, medium, and weak students? Second, is a team of 3 students better than a team of 4 students? Finally, the conventional wisdom is that a team of only 1 female student with the rest males should be avoided, because the male students will dominate the interactions in the team. Is that actually true?

We have compared team effectiveness in our 1000-student first year Physics course intended primarily for students in the life sciences (PHY131). PHY131 is the first of a two-semester sequence, is calculus based, and the textbook is Knight. Clickers, Peer Instruction, and Interactive Lecture Demonstrations are used extensively in the classes. The session that is studied here was held in the fall of 2014.

In addition to the classes, traditional tutorials and laboratories have been combined into a single active learning environment, which we call Practicals. In the Practicals students work in small teams on conceptually based activities using a guided discovery model of instruction, and whenever possible the activities use a physical apparatus or a simulation. The typical team has 4 students although due to logistic constraints about 15% of the teams have only 3 members. The students attend a two-hour Practical every week, and there are 10 Practicals in the term. It is the effectiveness of the teams in the Practicals that is studied here.

A third major component of the course is a weekly homework assignment. We use MasteringPhysics and the typical assignment takes the students about one hour to complete. Although we use some of the tutorials provided by the software to help student’s conceptual understanding, the principle focus of most homework assignments is traditional problem solving, both algebraic and numeric. We expect most students do these assignments as individuals, although we do not discourage the students from working on them together in a study team.

We give the Force Concept Inventory (FCI) to students in PHY131. The FCI was introduced by Hestenes, Wells and Swackhammer in 1992, and was updated in 1995. We use the common methodology of administering the instrument at the beginning of a course, the “Pre-Course”, and again at the end, the “Post-Course”. Our students were given one-half a point, 0.5%, towards their final grade in the course for answering all questions on the Pre-Course FCI, regardless of what they answered, and given another one-half point for answering all questions on the Post-Course FCI also regardless of what they answered. Below all FCI scores are in percent. The student’s score on the Pre-Course FCI was used to define whether he/she was strong, medium or weak, and the gains on the FCI from the Pre-Course to the Post-Course was used to measure the
effectiveness of instruction. We compared the gains on the FCI for different types of teams.

Although our study is of students in an introductory university physics course for life science majors, we expect that our results are relevant for many courses that do collaborative learning, both in other physics courses, in courses other than physics, and probably at the secondary as well as post-secondary level.

II. METHODS

We only studied Practical teams whose membership did not change from early in the term to the end of the term. Each Practical group contains up to 9 teams, and typically each team has four students, although due to logistic constraints 15% of the teams had three students and four teams out of 178 had five students. We do not allow teams of less than 3 or greater than 5 students. Each group of about 36 students has two Teaching Assistants present at all times.

This study caused us to make only two changes in the structure of the course for this year only. In past years the students were initially assigned to teams in the Practicals randomly, and halfway through the term the teams are scrambled, and the first meeting of the new teams begins with an activity on teamwork. This term we did not scramble the teams and, not entirely because of this study, we did not use the teamwork activity. Therefore, the composition of the teams typically only changed because dropouts required some re-distribution of students within a group, or sometimes because a team was felt by the TAs to be dysfunctional or even toxic. Thankfully these latter circumstances are very rare.

The second change in the structure of the course was that we assigned students to teams based on their Pre-Course FCI score. We had 30 groups, each typically consisting of about 36 students divided into 9 teams of 4 students. We used two methods for assigning students to teams, which we call “spread” and “sorted.” For the “spread” method, which we used for about half of the groups, we assigned team numbers 1, 2, 3, 4, 5, 6, 7, 8 and 9 to the top nine students based on FCI score, respectively. Then the next nine students were also assigned to teams 1 through 9, and so on. For the “sorted” method, which we used for the other half of the groups, we assigned all the students with the top four FCI scores to team number 1, then the next four students to team number 2, and so on; all four of the students with the lowest FCI scores were assigned to team number 9. In total, 16 groups were “spread” and 14 groups were “sorted”. At the end of the term we did anonymous student evaluations of the Practicals, and we looked at the student satisfaction as expressed on the evaluation for the spread and the sorted groups.
Classifying Students and Teams

We classified students as “strong,” “medium,” and “weak” by their score on the Pre-Course FCI. A strong student is one whose score was in the upper third of the class, a medium student in the middle third, and a weak student in the bottom third.

We defined a “weak” team as one for which all students were weak, and a “strong” team as one where all student Pre-Course scores were strong. A “mixed” team had at least one strong student, one medium student and one weak student. In an effort to reduce the statistical uncertainty in the values of gains by increasing the number of students in the sample, we also defined “strongish” teams as all the strong teams plus teams with one medium student and no weak students; similarly “weakish” teams were all the weak teams plus teams with one medium student and no strong students.

In addition to the standard 30 questions on the FCI, on the Pre-Course FCI we asked some further non-graded questions about the student’s background, motivation for taking the course, and their gender. The gender question, and the number and percentage of students in each category, was:

What is your gender?

A. male (405 students = 40%)
B. female (603 students = 59%)
C. neither of these are appropriate for me (9 students = 1%)

In our gender analysis, we ignored the nine students who chose C in the above question. Not all students answered this question, and therefore received no credit for taking the Pre-Course FCI.

The FCI

1045 students took the Pre-Course FCI, which was almost all students in the course. 910 students took the Post-Course FCI, again almost all students still enrolled in the course. The difference in these numbers is almost entirely because of students who dropped the course. In our analysis we only used FCI scores for “matched” students who took both the Pre-Course and the Post-Course FCI. This was 878 students. The 32 students who took the Post-Course FCI but not the Pre-Course FCI were late enrollees or missed the Pre-Course for some other reason.

Figure 1(a) shows the Pre-Course scores and Figure 1(b) shows the Post-Course scores for the matched students. The displayed uncertainties are the square root of the number of students in each bin of the histogram. Neither of these distributions are Gaussian, especially the Post-Course ones, so the mean is not an appropriate way of reporting the results. We will instead report on the median of the scores. The uncertainty in the median is taken to be $\pm 1.58 \times IQR / \sqrt{N}$ where $IQR$ is the interquartile range and $N$ is the number of students.$^{16}$ This uncertainty is taken to indicate very roughly a 95%
confidence interval, i.e. the equivalent of \(2 \times \sigma_m\) for normal distributions where \(\sigma_m = \sigma / \sqrt{N}\) is the “standard error of the mean.”\(^{17}\)

![Graphs showing FCI scores: Pre-Course (a) and Post-Course (b)]

Figure 1. FCI scores: (a) Pre-Course (b) Post-Course

We used the gain on the FCI to measure the effectiveness of different types of teams. The standard way of measuring student gains is from a seminal paper by Hake.\(^{18}\) It is defined as the gain normalised by the maximum possible gain:

\[
G = \frac{\text{PostCourse}\% - \text{PreCourse}\%}{100 - \text{PreCourse}\%}
\]

Clearly \(G\) cannot be calculated for Pre-Course scores = 100. This was 8 students in our course.

One hopes that the students’ performance on the FCI is higher at the end of a course than at the beginning. The standard way of measuring the gain in FCI scores for a class or subset of students in a class is called the average normalized gain, to which we will give the symbol \(<g>_{\text{mean}}\), and was also defined by Hake in Reference 18:

\[
< g >_{\text{mean}} = \frac{<\text{PostCourse}\%> - <\text{PreCourse}\%>}{100 - <\text{PreCourse}\%>}
\]

where the angle brackets indicate means. However, as discussed, since the distribution of FCI scores is not Gaussian, the mean is not the most appropriate way of characterizing FCI results. We will instead report \(< g >_{\text{median}}\), which is also defined by Eqn. 2 except that the angle brackets on the right hand side indicate the medians. In PER, the normalized gains are often taken to be a measure of the effectiveness of instruction.

The uncertainties in the median normalized gains reported below are the propagated uncertainties in the Pre-Course and Post-Course FCI scores. Since both of these are uncertainties of precision, they should be combined in quadrature, i.e. the square root of
the sum of the squares of the uncertainties in the Pre-Course and Post-Course scores. Therefore from Eqn. 2, for the median normalized gain:

\[
\Delta(<g>_{\text{median}}) = \sqrt{\left[ \frac{\partial(<g>)}{\partial(<\text{PreCourse}%>)} \Delta(<\text{PreCourse}%>) \right]^2 + \left[ \frac{\partial(<g>)}{\partial(<\text{PostCourse}%>)} \Delta(<\text{PostCourse}%>) \right]^2}
\]

\[
= \sqrt{\left[ \frac{<\text{PostCourse}%>-100}{(<\text{PreCourse}%>-100)^2} \Delta(<\text{PreCourse}%>) \right]^2 + \left[ \frac{\Delta(<\text{PostCourse}%>)}{100-<\text{PreCourse}%>} \right]^2}
\]

where \(\Delta(<\text{PreCourse}%>)\) and \(\Delta(<\text{PostCourse}%>)\) are the uncertainties in the medians of the Pre-Course and Post-Course FCI scores.

### III. RESULTS

In this section we will first discuss the normalised gains on the FCI, and then present the data from the student evaluations for the “sorted” and “spread” groups.

As discussed, we defined a “strong” student as one whose Pre-Course FCI score was in the upper third of the class, a “medium” student as one whose score was in the middle third, and a “weak” student a one whose score was in the bottom third. There were 273 strong students and for them \(<g>_{\text{median}} = 0.500 \pm 0.086\); there were 339 medium students with \(<g>_{\text{median}} = 0.467 \pm 0.036\); there were 266 weak students with \(<g>_{\text{median}} = 0.409 \pm 0.036\).

Table I summarises the median normalised gain for strong, strongish, mixed, weakish and weak teams. The different values of \(<g>_{\text{median}}\) and the median of \(G\) for different types of teams are all are roughly equal within uncertainties. Recall that the stated uncertainties correspond to a 95% confidence interval, i.e. they are equivalent to twice the uncertainty given by the standard deviation for data that are normally distributed.

Also shown in Table I are the results for all students. For \(<g>_{\text{median}}\) the value is for the 878 matched students, while for the median of \(G\) the value is for the 870 matched students who did not score 100% on the Pre-Course FCI.
Table I. Median normalised gains for different team types and for all students

<table>
<thead>
<tr>
<th>Team Type</th>
<th>Number of Teams</th>
<th>Number of Students</th>
<th>(&lt;g&gt;_{\text{median}})</th>
<th>Median of (G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weak</td>
<td>12</td>
<td>44</td>
<td>0.432 ± 0.088</td>
<td>0.438 ± 0.084</td>
</tr>
<tr>
<td>Weakish</td>
<td>17</td>
<td>62</td>
<td>0.405 ± 0.078</td>
<td>0.414 ± 0.077</td>
</tr>
<tr>
<td>Mixed</td>
<td>56</td>
<td>210</td>
<td>0.467 ± 0.072</td>
<td>0.439 ± 0.044</td>
</tr>
<tr>
<td>Strongish</td>
<td>37</td>
<td>127</td>
<td>0.571 ± 0.091</td>
<td>0.500 ± 0.059</td>
</tr>
<tr>
<td>Strong</td>
<td>22</td>
<td>80</td>
<td>0.50 ± 0.13</td>
<td>0.455 ± 0.082</td>
</tr>
<tr>
<td>All students</td>
<td>178</td>
<td>878 / 870</td>
<td>0.533 ± 0.034</td>
<td>0.452 ± 0.023</td>
</tr>
</tbody>
</table>

Examining the individual normalised gains \(G\) tells a similar story. Figure 2 is a boxplot for the different team types. The “waist” on the box plot is the median, the “shoulder” is the upper quartile, and the “hip” is the lower quartile. The vertical lines extend to the largest/smallest value less/greater than a heuristically defined outlier cutoff. The dots represent data that are considered to be outliers. The “notch” around the median value represents the statistical uncertainty in the value of the median; notched boxplots were first proposed in Reference 18. The vertical scale of the boxplot has been chosen so that the 10 values of \(G\) less than -0.95 are not displayed; these student outliers seem to have “blown” off the Post-Course FCI for some reason. These students were all in mixed, strongish, or strong teams: there were no outliers for the weak or weakish teams. The boxplot shows that there are no significant differences in the values and distributions of \(G\) for the different types of teams, except for a larger number of values less than the lower quartile for mixed, strongish, and strong teams.

For mixed teams, there were 64 strong students in the teams who had \(<g>_{\text{median}} = 0.50 ± 0.16\), the 74 medium students had \(<g>_{\text{median}} = 0.47 ± 0.05\), and the 72 weak students had \(<g>_{\text{median}} = 0.41 ± 0.06\). Figure 3 is the boxplot of the values.
of $G$. The vertical scale is chosen so that two strong students whose $G$ values were -3.0 and -3.6, and are therefore outliers, are not shown. Note that the strong students in this plot are completely different students than those in the strong teams of Fig. 2. For these strong students in mixed teams we see something similar what is seen for the strong students in strong teams, Fig. 2: there is a wider spread of values for $G$, particular for lower values.

![Boxplot of $G$ for different student strengths in mixed teams](image)

Figure 3. Boxplot of $G$ for different student strengths in mixed teams

We examined FCI gains for teams with 3 students and teams with 4 students. The values of $<g>_{\text{median}}$ were $0.57 \pm 0.14$ and $0.50 \pm 0.05$ respectively. The boxplot, which is not shown, also shows no significant differences in the distribution of values of $G$ for the two groups.

For the 21 teams with one female student and the rest males, the median normalised gain for the female students was $0.35 \pm 0.28$. The large uncertainty in this value is due to the small number of female students in the sample.

As mentioned, the initial team assignments were done two ways: in one-half the groups we assigned students so that all members of each team had roughly the same Pre-Course FCI scores, the “sorted” groups, and in the other half we distributed the students so that each team had a mixture of students with different Pre-Course FCI scores, the “spread” groups. At the end of the semester 912 students filled out an anonymous paper survey during the Practicals. Several questions asked about the TAs, but the first five questions asked specifically about student evaluations of the Practicals themselves; these questions are shown in the Appendix. Note that for all five questions, a response of 5 is the most favourable. For the 427 students in sorted groups, the mean of the five questions is $3.779 \pm 0.031$, and for the 485 students in spread groups it is $3.609 \pm 0.031$, where the uncertainties are the “standard error of the mean.” Figure 4 shows the distribution of the means of the five questions for the two types of groups. The displayed uncertainties are based on an uncertainty in the number of students in each bin equal to the square root of that number. The values of $<g>_{\text{median}}$ are $0.536 \pm 0.062$ for the sorted groups and $0.467 \pm 0.057$ for the spread groups; the difference in these values is $0.069 \pm 0.084$. 
IV. DISCUSSION

One of the core techniques of Physics Education Research is using the normalised gains on diagnostic instruments such as the Force Concept Inventory to determine what pedagogy is most effective. In the context of teaching classical mechanics, Hoellwarth and Moelter showed that for different sessions of the same course using the same research-based pedagogy but with different instructors, the normalised gains on the FCI were largely the same. Previously we have shown that, at least for this course, the normalised gain on the FCI is independent of whether or not the student was taking the course because it was required, and also independent of whether or not they had taken a senior-level high school physics course, although there was a correlation for these two factors with the scores on the FCI and with performance on the final examination. Here we have shown that the normalised gain on the FCI is also independent of the type of team that the student was in, on the size of the team, or for a single female student in a team of males. In fact, previously we have shown that for this course the only correlation in normalised gains that we know of is a small but statistically significant difference between the normal 12-week format of the course given in the fall term and the compressed 6-week format given in the summer.

The Practicals that are the subject of this study are only one component of this course. As mentioned, in addition to two hours a week of Practicals, the students also spend two hours a week in classes where Peer Instruction is used extensively, and one hour a week on homework assignments delivered with MasteringPhysics. The homework assignments are where the bulk of learning how to solve conventional problems occurs. MasteringPhysics in particular has been shown to have a large impact on student performance on tests and exams. The Practicals themselves spend very little time on problem solving. In addition to the classes, Practicals, and homework, the students also spend some time reading the textbook, although there are data that indicate that they
don’t use their textbook the way physics teachers think they do.\textsuperscript{24,25} Regardless of the type of team the students were in, they all have the same class, homework, and textbook, and we are examining the one component of many in the course that differs for different students.

Most teachers have probably noticed that the dynamics or “personality” of small groups of students can be very different for different groups, such as a small course in different years or different conventional tutorial groups for the same course. These variations seem to be greater than one might expect just from the statistics of small groups. Here we have found that student satisfaction with the Practicals was higher for sorted groups by $4\sigma$ over spread groups. In written comments on the evaluations, 9 students in the spread groups complained about the dynamics of their team, while no students in the sorted groups made a similar complaint. Of course, correlations are not necessarily indicators of causation. And also, of course, the normalised gains of the FCI were the same within uncertainties for the two types of groups.

Nonetheless, we can speculate on why there are differences in student satisfaction depending on the make-up of the Teams in the group. Perhaps in the spread groups the best students all ended up feeling a little alone at their pods, while the weaker students felt intimidated by the better students and less willing to participate. In the sorted groups perhaps there may be more of a feeling that each student “belonged” with his or her partners. So the four best students in Team 1 worked together to produce close to perfect work. Similarly, the four weakest students in Team 9 worked hard as well, and there was always a student at each of the weaker pods who rose to the challenge to become the leader, asking the TA for help and keeping the team focused. Also, in our rooms Teams 1 and 2 were directly across the aisle from Teams 8 and 9, which created some inter-team dialogue.

We attempted to examine the effectiveness of different types of teams by comparing student performance on the first term test to performance on the final examination. The attempt was not successful: the data give no meaningful information on team effectiveness, although they were not obviously inconsistent with the FCI data.

The students in the course studied here are mostly in the life sciences and, as reported in Ref. 21, only 16\% of the students say that their main reason for taking the course is for their own interest. As mentioned, there is a separate first-year course for students intending to be physics majors or specialists, and they are in general much more motivated to learn the material. It would be interesting to do a study similar to this one with that course.

\section*{V. CONCLUSIONS}

There are no statistically significant differences in the normalised student gains on the FCI for students in strong, medium, weak, or mixed teams. There are also no significant differences between students in a team of 3 compared to students in a team of 4. Also,
the conventional wisdom that a team with a single female student should be avoided is not supported by our data.

Therefore, we will go back to our former practice of initially assigning students to teams of 4 more-or-less randomly, and then scrambling the team memberships half way through the term. We will rely on our Teaching Assistants’ knowledge of their students, their ability in physics, and especially their ability to relate to each other to guide them in forming the new teams.

Finally, although we commonly tell our students that a null result of an experiment is often as important or even more important than a positive one, we will admit to some disappointment that our study has a null result.

ACKNOWLEDGEMENTS

We thank April Seeley of the Department of Physics for assistance in collecting the data of this study. We also thank undergraduate student Jody Chan and graduate student Dan Weaver of the Department of Physics for useful discussions on the design of this study. Dan Weaver also gave us useful comments on this paper.

APPENDIX

Here are the five questions about the effectiveness of the Practicals that were on the end-of-term anonymous student evaluation.

1. The heart of the learning in the Practicals is you working on the Activities with your Team. Was this an effective way of learning for you?
   1. Definitely not
   2. Probably not
   3. I’m not sure
   4. Probably
   5. Definitely

2. During the first part of Practicals you worked thorough problems with your team. Was this an effective way of learning for you?
   1. Definitely not
   2. Probably not
   3. I’m not sure
   4. Probably
   5. Definitely
3. Compared to other labs and tutorials at the same level (1st year U of T), the amount of help available to students during the Practicals is
   1. Well Below Average
   2. Below Average
   3. Average
   4. Above Average
   5. Well Above Average
4. Is the way we mark your work in the Practicals appropriate and fair?
   1. Definitely not
   2. Probably not
   3. I’m not sure
   4. Probably
   5. Definitely
5. The time available to complete the assigned activities (not including the “If you have time” activities) was, in general,
   1. Not Nearly Enough
   2. Not Enough
   3. Just Enough
   4. More than Enough
   5. Much More than Enough

REFERENCES

11 The U of T Practicals web site is: [http://www.upscale.utoronto.ca/Practicals/](http://www.upscale.utoronto.ca/Practicals/).
15 The student guide is at: [http://www.upscale.utoronto.ca/Practicals/Modules/Teamwork/Teamwork_Module.pdf](http://www.upscale.utoronto.ca/Practicals/Modules/Teamwork/Teamwork_Module.pdf)
17 http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.1978.10479236 (Retrieved November 15, 2014). Note that in this article the multiplier is 1.57, not 1.58: since the uncertainty itself is largely heuristic, the difference in these values is trivial.
18 Often the uncertainty in a value is called its “error”. Particularly for students, this word is misleading since it implies that some mistake has been made. This language issue is why, for example, it is Heisenberg’s uncertainty principle, not Heisenberg’s error principle. However, some standard phrases use the word “error,” such as the *standard error of the mean*. Here we enclose such phrases in quotation marks to remind us that the language of the phrase is poor.
20 There are various conventions for the cutoff definition. We use 1.5 times the interquartile range extending from the upper and lower quartiles, which was proposed in J.D. Emerson and J. Strenio, “Boxplots and Batch Comparison,” in D.C. Hoaglin, F. Mosteller, and J.W. Tukey eds., *Understanding Robust and Exploratory Data Analysis* (Wiley-Interscience, Toronto, 1983), p. 58. This cutoff definition is the usual one.
24 D.J. Palazzo, Y.-J. Lee, R. Warnaakulasooriya, and D.E. Pritchard, “Patterns, correlates, and reduction of homework copying,” *Phys. Rev. ST PER* 6, 010104 (2010). We have replicated these results with our students (unpublished).

Last revision: May 30, 2015