A LIVELY ELECTRONIC COMPENDIUM OF RESEARCH, NEWS, RESOURCES, AND OPINION

Astronomy Education Review

Volume 3, Mar 2004 - Oct 2004
Issue 1

The Impact of Cooperative Quizzes in a Large Introductory Astronomy Course for Non–Science Majors

by **Michael Zeilik**
University of New Mexico
Vicky J. Morris
University of New Mexico
Received: 02/03/04, Revised: 07/22/04, Posted: 08/13/04

The Astronomy Education Review, Issue 1, Volume 3:51-61, 2004

© 2004, Michael Zeilik. Copyright assigned to the Association of Universities for Research in Astronomy, Inc.

Abstract

In Astronomy 101 at the University of New Mexico, we carried out a repeated- items experiment on quizzes and tests to investigate the impact of cooperative testing. This trial was the only change in a reformed course format that had been refined over previous semesters. Our research questions were:

- Did cooperative quizzes result in gains for the class overall?
- Did these gains "stick" within the semester?

In the spring and fall semesters of 2000, students took quizzes individually and in cooperative learning teams, and tests individually. Normalized gain, $\langle g \rangle$, on the quizzes averaged about 0.4, and effect size about 0.8 (approximately a 10% increase in class mean score). Repeating selected quiz items on a subsequent test demonstrated that the gain was sustained over a month in both semesters. In addition, we compared demographics of UNM students with those of the National Astronomy Diagnostic Test project. We found that UNM students are similar to the national sample, except in ethnicity (more Hispanic American, fewer White). Based on these results, we judge that our cooperative quiz strategy will likely succeed in other "Astro 101" classes.

1. Introducion

The typical Astronomy 101 course in the United States may well be one of the toughest introductory sciences to teach successfully. Such courses have a diversity of goals (AAS 2003), students (Deming & Hufnagel 2001), topics (Slater et al. 2001), and expectations on the part of the students (Lacey & Slater 1999). Often, classes are large, minimally staffed, and slim of resources. (At the University of New

Mexico, we typically have three sections of 300, 200, and 100 students at different times of the day; 300 is the average enrollment in the United States according to marketing surveys.) Many instructors do not have formal backgrounds in astronomy (Fraknoi 2001), and if they are newcomers, they do not yet know what's best for their class in the universe of astronomical resources (Zeilik 1997) that are available mostly on the Web.

National consensus documents (e.g., *How People Learn*, National Academy Press, 2000; and *Transforming Undergraduate Education in Science, Mathematics, Engineering, and Technology*, National Academy Press, 1999) have synthesized the research base and furnished policy statements for reform. For science, technology, engineering, and mathematics (STEM), these accounts boil down to these statements:

1. Higher education must provide diverse opportunities for all undergraduates.
2. Faculty must craft an active, sustained community of discourse on learning issues based on evidence, not conjecture.
3. STEM courses, especially introductory ones, must incorporate active learning strategies.
4. Continual assessment must be integrated into the instruction, focus on topics of value, and close the feedback loop to instructors to implement improvements.

In practice, although numerous ideas to reform "Astro 101" courses have been proposed, few renewed courses have been created, implemented, and assessed! Emerging from the physics and astronomy research, though, is the notion that "reformed courses" (those implementing active learning and a cognitive model of learning) result in greater achievement than "standard model" courses (those using didactic lecture and a transmission model of learning; see Crouch & Mazur 2001 for results from 10 years of reformed instruction in physics).

In this article, we present a learning strategy common in cooperative learning but not often used in Astro 101: *cooperative quizzes*. This approach logically arises in a class for which cooperative learning teams are a major part of the instructional strategy. In addition, our course specifically implements points 3 and 4 in a large-class environment (Zeilik, Schau, & Mattern 1998, 1999) taught with an integration of research-based strategies since fall 1994. The quizzes served as one assessment tool. They were not scheduled on the course syllabus; instead, we informed students one class beforehand that a quiz would be given. Essentially, they served as "reading quizzes" for the assigned textbook. *The addition of the cooperative quizzes was the only change made in the course design compared with prior semesters*.

Cooperative quizzes can vary in format (Byrd, Coleman, & Werneth 2004). Because of our large enrollments and nominal staffing, we employ a multiple-choice format. UNM computer services uses only 5- and 10-item Scantron sheets and a 20-year-old software package. We have nominal flexibility in administering multiple-choice exams because of these constraints. Our solution in this context: Use the five-item form in a two-step process. Students use Section 1 of the form to answer all of the quiz questions on their own. This part typically takes about 10 minutes for 15 items.

Students then form into their cooperative instructor-assigned learning teams of four to five students to discuss the quiz and debate answers. The class becomes very animated as students engage with their peers in negotiation over specific astronomical concepts. Once the discourse ends, individuals use Section 2 of the Scantron to answer all of the questions again. *Each student has the choice to use his or her original individual answer, or to change to a consensus response or any other response--students are not forced to do so!* The quiz score is an equally weighted average of Section 1 and Section 2 so that individuals are

held responsible for their own learning and active participation in their teams.

We monitored the teams to ensure that students were not going back and changing Section 1 answers after discussion. The team members applied peer pressure to enforce this rule socially. We also note anecdotally that students in a team do not blindly follow the "smartest student" when they have no consensus. Rather, they enter into sometimes heated debates and negotiations. We consider this interaction the core of the team learning process.

A common misconception about cooperative learning teams is that the "best" student dominates, while the others passively go along. This "freeloading" issue is well known, and effective strategies have been developed to cope with it (Johnson, Johnson, & Smith 1991). Larry Michaelsen and colleagues have developed, implemented, and refined a particularly effective course design for team learning (Michaelsen, Knight, & Fink 2004, and references therein) in which team scores on content and application assessments eventually are higher than those of the highest team member (see http://atlas.services.ou.edu/idp/teamlearning/collection.htm; click on "Individual vs. Team Cumulative RAT Scores" for a table of these results).

A few comments about the multiple-choice items. Since 1994, MZ has developed conceptually based test materials (Zeilik et al. 1997). The items fall into two broad categories: concept recognition and concept extension. The latter has two parts: near transfer and far transfer. These items probe student understanding of key concepts by testing their ability to apply concepts in situations similar to ones that they have seen before (near transfer), and novel situations (far transfer). The items are designed so that no amount of memorization will ensure a correct answer. In general, the class average percentage on near-transfer items is in the mid-70s, and on the far transfer, the mid-60s. The students call far-transfer items "tricky"; generally, they perceive the quizzes and tests as "hard." (They express these views on the end-of-semester standard UNM student course evaluation form.)

An example of a near transfer item:

> Imagine that you are viewing an evening scene in which the sun is just setting, Venus is near the Sun, Mars is at opposition, and Jupiter lies about halfway between Venus and Mars. Consider making a geocentric model and a heliocentric model of this situation. When you compare the two models, the angular relationships between the planets are explained:
>
> A. More simply in a heliocentric model.
> B. More simply in a geocentric model.
> C. Equally well in both models.
>
> Equally well, because all angular relationships are the same. Near transfer. Answer: C.

In our current project, we aimed at answering the following research questions:

- Did cooperative quizzes result in gains for the class overall?
- Did these gains "stick" within the semester?

2. Procedure and Results

We assessed two semesters of Astronomy 101 at UNM: spring 2000 and fall 2000. Both classes had similar sizes (120 in spring, 150 in fall) and demographics (we used the Astronomy Diagnostic Test version 2 to collect self-reported demographics on a volunteer basis). The classes met twice a week for 75 minutes. Class periods included about one third lectures focused on key concepts, and about two thirds learning team activities. The primary homework was reading the textbook (*Astronomy: The Evolving Universe*, 8th edition, John Wiley & Sons, 1997) *before* classwork on the assigned topics. We used the cooperative quizzes in part as an incentive to read the textbook in a timely manner (we have no formal data that this tactic worked).

- Did cooperative quizzes result in gains for the class overall?

In spring 2000, we gave five cooperative quizzes and three individual tests. The quizzes had 10 to 20 items. The tests had 50 items each and were taken by students with no peer interactions during the tests. For each, students were allowed to bring one 8.5" x 11" study sheet. None of the tests or quizzes was constrained by time; every student had ample time to finish within the class period. Note that the tests reinforce individual responsibility of students' learning generated in team interactions.

We used normalized gain, $\langle g \rangle$, as one metric of merit (Hake 1998; Hovland, Lumsdaine, & Sheffield 1949), where

$\langle g \rangle$ = (post% - pre%)/(100 - pre%)

A $\langle g \rangle$ of 0 means no gain, while a $\langle g \rangle$ of 1 indicates that all possible gain occurred.

Another figure of merit usually quoted in the realm of education research is the effect size:

Effect Size = ES = (posttest mean - pretest mean)/mean SD of the distributions

The effect size is the difference between the means of the pre- and postscores divided by their mean standard deviation (sometimes called "pooled standard deviations"). It is the difference between the means in standard deviation units. In essence, it measures the average superiority (if positive) or inferiority (if negative) of the final state compared with the initial state, while taking into account the variability of the population.

Effect size is a powerful indicator of the separation of the pre- and postcourse score distributions, and so of the gains across the semester. It permits a calibration of comparisons across different characteristics of a study by normalizing the results by standard deviations. In the educational research, effect sizes of 0.1 or less are considered small and of no practical import; 0.3 are considered medium and do have practical significance; and 0.5 or greater are considered large (and unusual; see Cohen 1988).

See Table 1 for the results for spring 2000. The prescores are the mean class results before the teams' interaction; postresults are the means after the teams' conversations. Figure 1 shows item-by-item results for Quiz 3 to give a sense of the gains.

Table 1. Astronomy 101 UNM Spring 2000 Cooperative Quiz Results

Quiz #; n = #students; N = # items	Premean (% ± SD)	Postmean (% ± SD)	<*g*>; Effect size
1; n = 116; N = 15	67.1 ± 17.2	76.1 ± 18.0	0.28; 0.51
2; n = 108; N = 10	72.4 ± 14.0	85.3 ± 10.8	0.47; 1.03
3; n = 112; N =12	73.8 ± 19.0	84.8 ± 18.5	0.42; 0.58
4; n = 112; N = 10	64.1 ± 20.4	70.9 ± 20.6	0.19; 0.33
5; n = 94; N =15	73.3 ± 11.2	82.1 ± 11.4	0.33; 0.78

Averaging all quizzes, we find that

Normalized gain = <*g*> = 0.32 ± 0.26 (SD)

and

Effect size = ES = 0.58 ± 0.27 (SD).

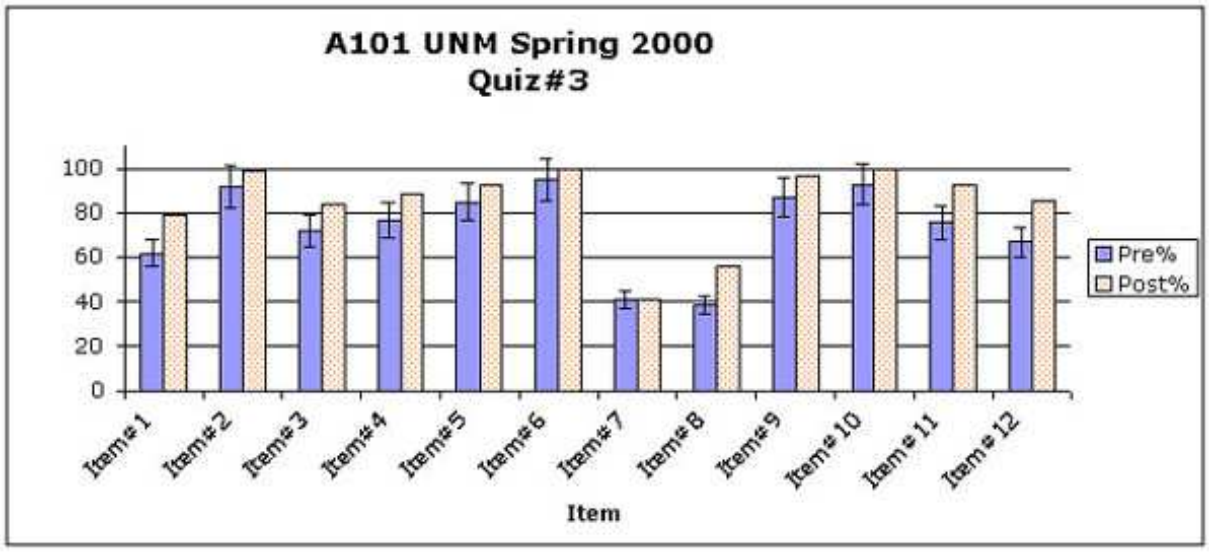

Figure 1. Premean (before team discussion) and postmean (after team discussion) results by item for Quiz 3. Error bars are standard deviations from the mean.

We performed a similar experiment in the fall 2000 semester. Figure 2 shows the results (mean class scores) for first three quizzes given, including <*g*> for each.

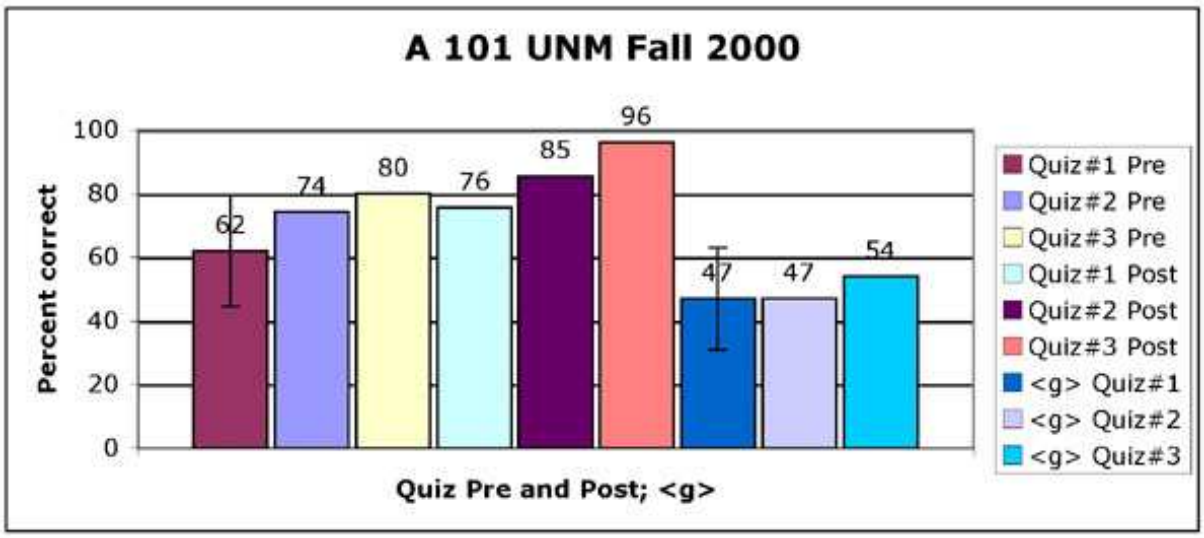

Figure 2. Pre- and postresults for cooperative quizzes. Quiz scores are the class means. Error bars are standard deviations (SD) from the mean and are essentially the same for each measure.

Here the average <*g*> is larger than spring 2000, in part due to our experience from that first implementation:

<*g*> = 0. 49 ± 0.04 (SD)

and

ES = 0.98 ± 0.42 (SD)

We conclude that cooperative quizzes resulted in consistent gains: about <*g*> ≈ 0.4, and ES ≈ 0.8. On grading, quizzes averaged a gain of about 10%--one letter grade by our grading rubric--for these two semesters, with improvement in the second semester. By educational standards, an effect size of 0.8 is considered large.

- Did these gains "stick" within the semester?

Within the context of a semester-long class, we wanted to see if the large cooperative gains persisted when students were held responsible individually for their learning, a key aspect of well-functioning cooperative learning teams. We used a selected test as a check (we gave three tests each semester, plus a final exam). Our hypothesis is that if learning persisted, scores on repeated items would remain high. If not, the mean scores would decline.

For each semester, we performed the following experiment on one test. Before the repeated-items test, the class took quizzes in the usual cooperative mode. We then selected items from the quizzes to be included on the next test, so the duration of "stickiness" checked was about one month. We picked questions that included important concepts and concepts related to known misconceptions. The answers to all items were posted on a physical class bulletin board after each quiz. Past experience has convinced us that few students actually review the quiz results; memorization of answers by many is not a confounding issue.

The repeated items were not reviewed in class (though other items were reviewed and did not show equivalent gains). On the test, the answers to the repeated items were scrambled so that the best choice on the quiz did not correspond to that on the test.

All quizzes and the tests were statistically reliable, with the reliability coefficient (Cronbach's alpha) ranging from 0.65 to 0.80 for the spring 2000 quizzes, and 0.71 for Test 3 in spring 2000. They ranged from 0.70 to 0.82 for the tests overall in both semesters. (Note that the reliability coefficient depends on the number of test items; more items result in greater reliability. Our quizzes were short, 10 to 15 items.) Acceptable reliability coefficients fall above 0.6.

Figure 3 depicts the results for spring 2000 for the repeated items, comparing the postdiscussion quiz means with subsequent test means, Test 3. (Note that the item numbers here are not the same as those in Figure 1.) Now to confirm our hypothesis, we need to show that the difference between the means of these two distributions is not statistically significant. That would indicate that the means remained the same and learning persisted. If the means were statistically significant and lower, that would indicate that learning declined. If the means were statistically significant and higher, that would indicate that learning increased.

One standard way to do this statistical check is through a *t* test. The *t* test is the most frequently used inferential statistics test to find out the statistical probability that the means from two samples come from populations with identical means. A statistically significant *t* value indicates that a mean difference of this size would have occurred because of sampling error (chance) at the probability level (*p* value) associated with the specific *t*-test value. When that probability is small (equal to or less than 0.05, or 5%, is the standard in a statistical analysis), we can conclude that the means likely differ. Using 5%, we have a 95% chance of being correct.

In our experiment, when comparing the postquiz distribution with that of the test for all repeated items, we find from a *t* test (two-tailed, using unequal variances) that $p < 0.62$ for the spring 2000 data in Figure 3. The score distributions do not differ statistically on this inferential test; our data do not give a 5% confidence level or lower. What does this signify? *That the means from two samples (quizzes and tests) come from populations with identical means*. A statistically significant *t* value indicates that a mean difference of this size would have occurred because of sampling error (chance) at the probability level *p* (62%). The small normalized gain, $\langle g \rangle = 0.089$, and effect size (ES = 0.14) in these data are just due to sampling error and are not statistically significant. This result supports our hypothesis that learning persisted over the time sampled by the tests, and essentially at the same level of performance averaged over the repeated items.

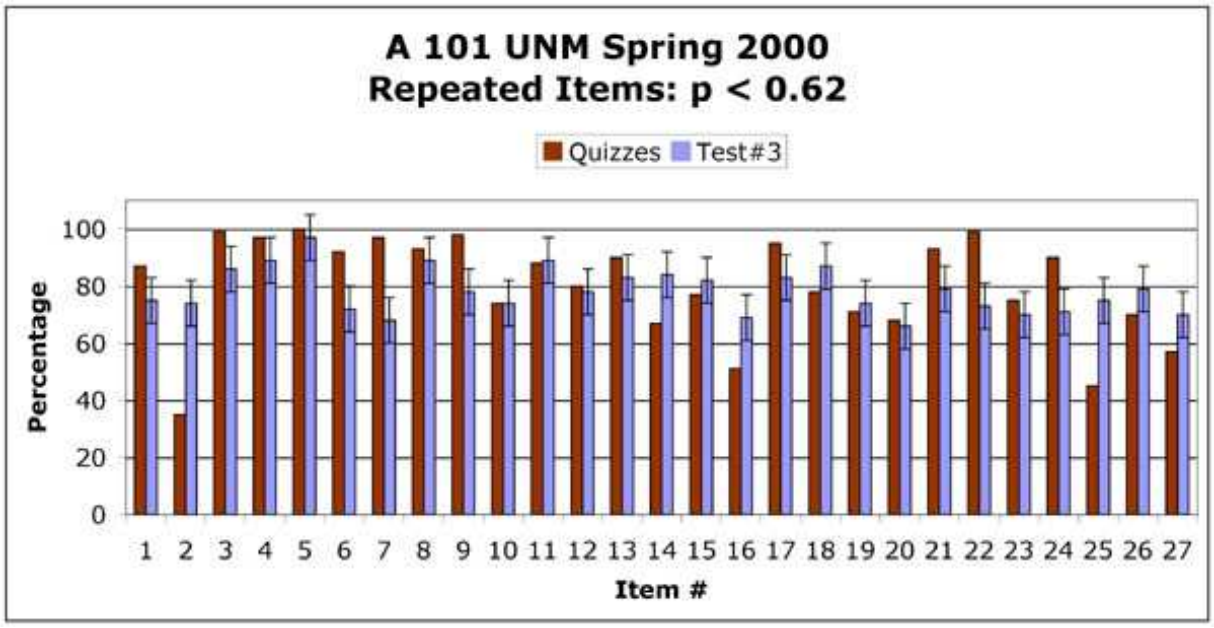

Figure 3. Results (mean class scores) of repeated items from quizzes given on Test 3 (n = 106). The quiz scores are the postdiscussion means for the class; Test 3 was taken individually with no team interactions. Error bars are standard deviation (SD) of the mean errors.

In the fall 2000 course, we carried out the experiment earlier in the semester, on the first test, with three quizzes preceding it. Figure 4 provides the results. We again performed a *t* test (two-tailed, unequal variances) comparing the postdiscussion class scores with those of identical items repeated on Test 1. Our hypothesis is that the two distributions would not differ if the students retained their understanding over a one-month period. A $p < 0.96$ indicates that the means of the postgroup and individuals' scores on these items were essentially the same within this time frame. We expect this result if the team discussions reinforced the individual learning of the concepts contained in these items.

Again, we checked the reliability of all measures and found a reliability coefficient of 0.84 for Test 1 and a range of 0.70 to 0.82 for the three quizzes. It is important that the reliabilities fall in this acceptable range; otherwise, the *t* test, $\langle g \rangle$, and effect size results are pointless.

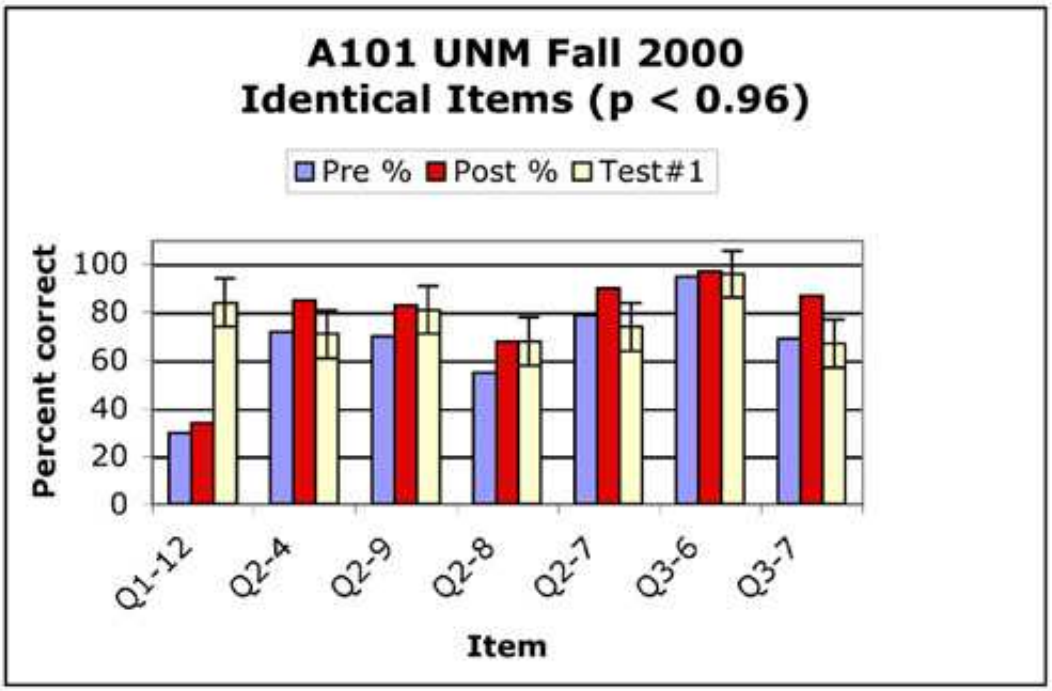

Figure 4. Items repeated on Quizzes 1 to 3 and Test 1. "Pre" is the class mean score on the quiz items, before team discussion. "Post" is the mean score after discussion. "Test #1" is the class average taking the exam individually, with no group discussion (n = 152). Error bars are standard deviations (SD) of the mean.

3. Conclusions and Limitations

We conducted the test-retest experiment one time per semester for only two semesters of UNM's Astronomy 101, but gave weekly cooperative quizzes. Both semesters used a naturalistic setting in a reformed class taught by the same instructor (MZ) using the same textbook. These two classes of our reformed course were the mature, primary implementations of the class after six years of development and assessment. We had no comparison group from a traditionally taught class of Astronomy 101 at UNM. With reasonably large numbers of participants and checks for reliability and statistical significance, we believe that our main results--large effect size gains on quizzes--will transfer to other Astro 101 classes:

- Cooperative quizzes result in a 10% gain overall in class average score and an effect size of about 0.8.
- The gain "sticks" for a time duration of at least one month.

Byrd et al. (2004) reported on a variation of cooperative quizzes in Astronomy 101 at the University of Alabama, Tuscaloosa. The found a gain in final exam performance from 57% (2001, no cooperative quizzes) to 80% (2002, cooperative quizzes), for a $\langle g \rangle = 0.53$. They also noted that the class GPA increased from 3.80 in 2001 to 4.33 in 2002. These independent results confirm the effectiveness of cooperative quizzing.

However, our results may not generalize to Astro 101 courses with very different demographics. In classes that participated in the Astronomy Diagnostic Test National Survey, Deming & Hufnagel (2001) found that 69% of students self-report an ethnic background of 69% White (non-Hispanic). Of the entire sample, 52% were women. At UNM, we used the ADT 2 as a pre/post-assessment with the same demographic items used in the national survey. Only two factors differ significantly: White, non-Hispanic (47% UNM, 69% national) and Hispanic American (17% UNM, 6 % national).

Aside from ethnic background, UNM students reflect the national outcomes fairly well. Unless ethnicity plays a key function in class gains, we deem that Astro 101 classes with national-like distributions should be able to attain the same magnitude of gains from cooperative quizzes as was achieved at UNM--essentially one letter grade in our class grading system.

Acknowledgments

This work was supported in part by NSF grant DUE-9981155. We thank Tim Slater, Gina Brissenden, and M. Jennifer Markus for a critical reading of early drafts.

References

Byrd, G. G., Coleman, S., & Werneth, C. 2004, Exploring the Universe together: Cooperative Quizzes with and without a Classroom Performance System in Astronomy 101, *Astronomy Education Review*, 3(1), http://aer.noao.edu/AERArticle.php?issue=5&section=2&article=3.

Cohen, J. 1988, *Statistical Power Analysis for the Behavioral Sciences*, Hillsdale, NJ: Erlbaum.

Crouch, C. H., & Mazur, E. 2001, Peer Instruction: Ten Years of Experience and Results, *American Journal of Physics*, 69(9), 970.

Deming, G., & Hufnagel, B. 2001, Who's Taking ASTRO 101?, *Phys. Teach.*, 39, 368.

Fraknoi, A. 2001, Enrollments in Astronomy 101 Courses: An Update, *Astronomy Education Review*, 1(1), 121, http://aer.noao.edu/AERArticle.php?issue=1&section=4&article=2.

Hake, R. R. 1998, Interactive-Engagement vs. Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses, *Am. J. Phys*, 66 (1), 64. http://www.physics.indiana.edu/~sdi/.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. 1949, A Baseline for Measurement of Percentage Change, In C. I. Hovland, A. A. Lumsdaine, & F. D. Sheffield (Editors), *Experiments on Mass Communication*, Wiley (first published in 1949). Reprinted as pages 77-82 in P. F. Lazarsfeld & M. Rosenberg (Editors), *The Language of Social Research: A Reader in the Methodology of Social Research*, New York: Free Press, 1955.

Johnson, D. W., Johnson, R. T., & Smith, K. A. 1991, *Active Learning: Cooperation in the College Classroom*, Edina, MN: Interaction Book Company.

Lacey, T., & Slater, T. F. 1999, First Contact: Expectations of Beginning Astronomy Students, BAAS, 31(2). Poster presented at the 194th meeting of the American Astronomical Society, Chicago, IL, 2 June 1999.

Michaelsen, L. K., Knight, A. B., & Fink, D. L. 2004, *Team-Based Learning: A Transformative Use of Small Groups in College Teaching*, Herndon, VA: Stylus Publishing.

Slater, T., Adams, J. P., Brissenden, G., & Duncan, D. 2001, What Topics Are Taught in Introductory Astronomy Courses?, *Phys. Teach.*, 39, 52.

Zeilik, M. 1997, To a Physicist New to Teaching Astronomy, *Phys. Teach.*, 35, 172.

Zeilik, M., Schau, C., & Mattern, N. 1998, Misconceptions and Their Change in University-Level Astronomy Courses, *Phys. Teach.*, 36, 12.

Zeilik, M., Schau, C., & Mattern, N. 1999, Conceptual Astronomy. II. Replicating Conceptual Gains, Probing Attitude Changes Across Three Semesters, *Am. J. Phys.*, 67, 923.

Zeilik, M., Schau, C., Mattern, N., Hall, S., Teague, K. W., & Bisard, W. 1997, Conceptual Astronomy: A Novel Model for Teaching Postsecondary Science Classes, *Am. J. Phys.* 65, 987, 65, 987.

ÆR
51 - 61