David Steer, David McConnell, Kyle Gray, Karen Kortz, Xin Liang

Analysis of Student Responses to Peer-Instruction Conceptual Questions Answered Using an Electronic Response System: Trends by Gender and Ethnicity

This descriptive study investigated students' answers to geoscience conceptual questions answered using electronic personal response systems. Answer patterns were examined to evaluate the peer-instruction pedagogical approach in a large general education classroom setting.

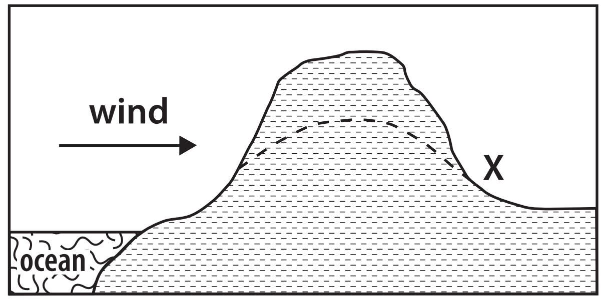
Over the past decade, it has become apparent that effective learning occurs in Science, Technology, Engineering and Mathematics (STEM) classrooms that use student-centered, active approaches that allow interactive exchange between and amongst students and instructors (American Geophysical Union, 1994; National Research Council, 1997, 2000; National Science Foundation, 1996). Such exchanges are facilitated when students use electronic personal response systems to answer conceptual multiple choice questions, called *conceptests* by Mazur (1997) and referred to as *think-pair-share* exercises in some disciplines (McTighe & Lyman, 1988). Conceptests are repetitive measures designed to explore student depth of understanding (both individual and group), and they often include answers with known preconceptions. Students consider the question and respond individually. Crouch and Mazur (2001) suggest that an initial correct response rate of 35% - 70% is optimal for these questions. Peer instruction is a practice in which students work together in pairs and small groups to discuss and defend their responses (Mazur, 1997), and this discussion may be followed by a second round of student responses. The use of conceptests is formative, because they provide timely feedback that the instructor and student can use to improve their performance. Much has been written about the ways in which this technique can be used by faculty (Cox & Junkin, 2002; Crouch & Mazur, 2001; Green, 2003; Hake, 1998; Mazur, 1997; McConnell, Steer, Owens & Knight, 2006; Pilzer, 2001; Rao & DiCarlo, 2000; Sokoloff & Thornton, 1997). The evidence is also compelling that this technique improves student learning from a course perspective (Crouch & Mazur, 2001; King & Joshi, 2008; Lasry, Maur, & Watkins, 2008; Smith et al., 2009) and that the technology is well received by students (MacGeorge et al., 2008a). Less is known about the impact this technique has on subpopulations of students based on gender and race.

Responses were analyzed for predictability, construct validity and gender reliability assuming a statistically normal response distribution.

The conceptests used in this study were taken from a large database of questions for the geosciences that were developed by more than 30 geoscience faculty members with multiple years of experience teaching introductory courses in a variety of settings (e.g. community college, small 4-year, and public universities). Those faculty members used their personal experiences and a review of the published literature to develop lists of geoscience concepts that are difficult for students to grasp and are discussed in most typical introductory geoscience courses for non-majors. Some of these concepts include plate tectonics, geologic time, the rock cycle, and the water cycle. The conceptests were generated according to good practices for writing multiple choice questions (Haladyna, Downing, & Rodriquez, 2002) by focusing on a single concept, using simple language or graphics, and including 3-4 short answers that require few or no calculations. The distracters (incorrect answers) also include alternative conceptions, misconceptions, or incorrect intuitive responses. The conceptests probe student understanding at various cognitive levels and emphasize the comprehension and application ("understanding" and "applying" levels in Anderson and Krathwohl [2001]) through analysis and evaluation levels of cognitive processing (Bloom, 1956).

This study focuses explicitly on conceptual questions at the *understanding*, *applying,* and *analyzing* cognitive levels (Anderson & Krathwohl, 2001), because, these are the most appropriate levels to assess using multiple-choice formats. The questions are posed as text-, diagram-, or graph-based problems, and they are similar to questions on the summative exams. At the *understanding* level, students demonstrate they are able to convert concepts learned as text to an illustration or vice versa. Students are also asked to compare and contrast objects or concepts, select reasons, compare solutions, or make predictions (see Figure 1). At the *applying* level, students apply rules or principles to new situations, use known procedures to solve problems, or demonstrate that they know how to do something. When working at the *analyzing* level, students select answers that explain how something works or distinguish fact from opinion. Questions that require students to scrutinize graphical data or images are interpreted as analysis questions, especially if the students have not previously seen the graph (see Figure 2).

Figure 1: Example of a diagram-based-based, understand-level conceptest related to the orographic lifting of air.



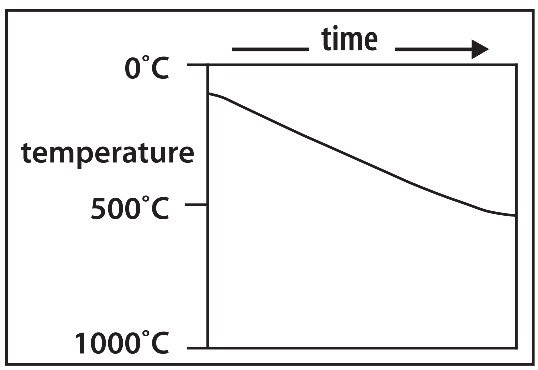
In the landscape pictured, how would the amount of rainfall change at location X if the mountain eroded down to the dashed line?

a) Rainfall would increase

b) Rainfall would decrease

c) Rainfall would stay the same

Figure 2: Example of a graph-based, analysis-level conceptest related to the rock cycle.



The graph illustrates how the temperature changed with time for part of the rock cycle. Which of the following is best represented by the graph?

a) Sand is lithified to form sandstone

b) Limestone is metamorphosed to form marble

c) Marble is uplifted to Earth's surface

d) Magma cools to form granite

e) Shale is heated and converted to magma

Methods

The data used for this study represents 4712 responses to conceptests collected from 242 students enrolled in four earth science classes for non-science majors and one physical geology class at a community college. These classes were taught by three instructors, each with over five years of teaching experience using active learning strategies. In addition to incorporating conceptests using peer instruction (Mazur, 1997; McConnell, Steer, Owens, & Knight, 2006), classes were taught using a variety of learner-centered activities including the use of student-manipulated physical models (Gilbert & Ireton, 2002), lecture tutorials (Kortz, Smay, & Murray, 2008), and predictive demonstrations (Sokoloff & Thornton, 1997). Students earned participation points for responding to conceptests, regardless of whether the answers were correct or incorrect. Three classes occurred in spring 2008, and two classes occurred in fall 2008.

This study reports conceptest response trends for paired answers from students who answered from 10-26 questions each over the course of the semester. The questions are assumed to be valid for content since they were developed by geoscience educators and have been reviewed for content validity by 12 experts across multiple institutions. Reliability and validity testing was completed for the questions using responses collected in spring 2006 from a large-format, general education introductory earth science class (155 students). Responses were analyzed for predictability, construct validity and gender reliability assuming a statistically normal response distribution. Correct response rates for the questions as a whole were not gender biased ($p>0.35$, $n=55$). Three individual questions appeared to show bias even after addition of response data for the same questions from fall 2005. As a set, the 52 remaining conceptest questions used in this study met predictive validity requirements. The percentage of students correctly responding to comprehension-, application- and analysis-level questions decreased with increasing question cognitive level ($p<0.0001$; 67%, 52% and 36% respectively).

Student responses from conceptests answered during lessons that used the peer instruction technique were scored using a rubric (Table 1). Those scores were used to evaluate the efficacy of this pedagogical technique for various populations (male, female, Causcasian, and minority). Students in selected courses completed a 15 question, Geoscience Concept Inventory (GCI) test (Libarkin & Anderson, 2005) as an independent assessment of geoscience conceptual understanding. The GCI is a valid and reliable assessment designed to assist geoscience faculty in evaluating teaching and learning (Libarkin & Anderson, 2005). Its purpose and design are similar to the Force Concept Inventory (Hestenes, Wells, & Swackhammer, 1992) that is widely used in physics education.

Table 1: Scoring rubric for studentresponses to conceptest questions.

Pre-Discussion Answer	Post-Discussion Answer	Score	% of Responses
Correct	Incorrect	**1**	5%
Incorrect	Incorrect	**2**	28%
Incorrect	Correct	**3**	26%
Correct	Correct	**4**	41%

Note that, as averaged over all questions, 33% of students recorded incorrect responses after peer instruction, and the remainder recorded a correct answer on the second attempt.

Note that, as averaged over all questions, 33% of students recorded incorrect responses after peer instruction, and the remainder recorded a correct answer on the second attempt.

Student engagement was determined by dividing the number of student answers to conceptests by the total number of questions posed. For example, a score of 70% on student engagement was recorded by a student who answered 70% of the conceptests analyzed in the study. These scores were a proxy indicator of attendance. Average conceptest scores were calculated by dividing the number of correct answers by the number of questions asked, and no deduction was made for unanswered questions. Individual student response rates for each question category (Table 1) were calculated by dividing the number of responses in a category by the total number of questions answered by that student. Final course grades and post-course GCI scores were also used as summative assessments of student success. Response data were grouped by gender and ethnicity for analyses. African American, Asian, Pacific Islander, and Hispanic were combined under the 'minority' classification.

All data fields were not available for all students (due to student absence during administration of the GCI, missing self-reported data, failure to complete the course, etc.). In all, five variables (pre-GCI, post-GCI, final grades, average proportion of correct answers on conceptests, and engagement) were analyzed for each of the four populations (minority male, minority female, Caucasian male, Caucasian female). Pearson's correlation coefficients (δ) were calculated for the 20x20 matrix with values of 0.1-0.3 considered of small significance, 0.3-0.5 moderate, and 0.5-1.0 large. Comparisons between larger populations (male-female, minority-Caucasian) were also completed using ANOVA or statistical T-tests using Cohen's d values for effect sizes, and values of $p<0.05$ were considered significant.

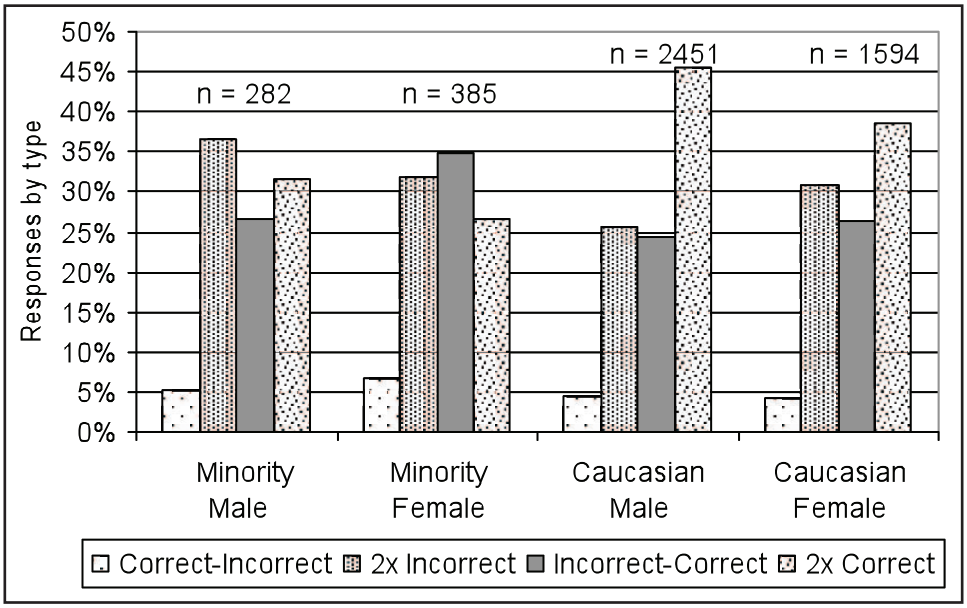
Data

Data were sorted by both gender and race (Figure 3) to show how student responses were distributed in the four paired-response categories (correct-incorrect, twice incorrect, incorrect-correct, twice correct; see Table 1). The total response database included 6% minority male ($n = 282$), 8% minority female ($n = 385$), 52% Caucasian male ($n = 2451$), and 34% Caucasian female ($n = 1594$) responses.

Correct-Incorrect: Overall, approximately 5% of responses showed students answered conceptest questions correctly the first time the question was posed and incorrectly on the second attempt (Table 1; Figure 3). There were not enough responses in this answer category for meaningful analyses between population groups.

Twice-Incorrect: About 28% of all responses were incorrect on both attempts (Table 1; Figure 3). As a percentage of their responses, male minority students were most likely to answer in this way (over 36% of their responses, Figure 3). Female students of both demographic groups answered in this fashion in approximately equal proportions (32%), and Caucasian males answered twice incorrect 26% of the time. Comparisons in twice incorrect response rates between minority populations and for Caucasian females compared to other populations showed small effect sizes (d = 0.0 to 0.3). There were moderate effect sizes when comparing male Caucasian response rates in the twice incorrect category to minority males and females (d = 0.5 and 0.6).

Figure 3: Percent of student conceptest responses by demographic group.



Twice Correct: The largest differences between populations were noted when analyzing the 41% of twice-correct responses (Table 1: score 4; Figure 3). Caucasian male students were most likely to answer correctly both times (45% of responses). Their female counterparts answered in this fashion about 39% of the time. Female minority students were least likely to answer twice correct (26% of responses), and minority males answered in this way 32% of the time. Effect sizes were small to moderate when comparing female Caucasian students to males (d = 0.3 for minority males; 0.4 for Caucasian males) and when comparing minority males and females (d = 0.4). Effect sizes were larger when analyzing Caucasian males to both minority populations (d = 0.6 for males; 1.3 for females) and when comparing female populations (d = 0.9).

Incorrect-Correct: Overall, 26% of the responses were incorrect on the first attempt, but correct after peer instruction (Table 1: score 3). At this level, minority females faired better than their minority male peers (35% versus 27% of responses) and slightly better than Caucasian students (26% for Caucasian females and 24% for males). Effect sizes were small when comparing minority males to both Caucasian populations (d = 0.0 for males; 0.2 for females). All other response rate comparisons in the incorrect-correct category displayed moderate effect sizes (d = 0.5 to 0.7).

Combined Responses: When average response rates for individual students by demographic group were compared to other course variables (pre- and post-GCI, final grades, conceptest average, and engagement), several trends appeared (Table 2).

Male minority students: Minority male conceptest averages were strongly correlated with post-course GCI scores (δ=0.9; Table 2: row D, column B) and moderately correlated with final grades (δ=0.5; Table 2: row D, column C). Pre-course GCI scores (Table 2: column A) were strongly correlated to final grades (Table 2: row C) and post GCI scores (Table 2: row B) for this population (δ=0.6 and 0.7). Engagement (Table 2: row E) displayed a moderately negative correlation with post-GCI scores (Table 2: column B) and moderately positive correlations to final grades and average conceptest scores (δ=0.4; Table 2: columns C and D).

Female minority students: Minority female average conceptest responses (Table 2, row I) displayed a strong negative correlation with pre-GCI scores (δ=-0.6; Table 2: column F). Final course grades (Table 2: row H)

Table 2: Pearson's correlation coefficients for studied populations

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A: Pre GCI Male Minority	1.0																			
B: Post GCI Male Minority	0.7	1.0																		
C: Final Grade Male Minority	0.6	0.5	1.0																	
D: CT Avg Male Minority	0.0	0.9	0.5	1.0																
E: Engage Male Minority	-0.3	-0.4	0.4	0.4	1.0															
F:Pre-GCI Female Minority	0.5					1.0														
G:Post-GCI Female Minority		-0.4				0.2	1.0													
H:Final_Grade Female Minority			0.0			0.5	0.2	1.0												
I: CT Avg Female Minority				-0.2		-0.6	0.3	0.1	1.0											
J: Engage Female Minority					-0.1	-0.1	0.2	0.5	0.3	1.0										
K: Pre-GCI Male Caucasian	0.0					0.4					1.0									
L: Post-GCI Male Caucasian		-0.3					0.1				0.4	1.0								
M: Final_Grade Male Caucasian			-0.3					0.0			0.0	0.1	1.0							
N: CT Avg Male Caucasian				0.2					-0.1		0.2	0.4	0.2	1.0						
O: Engage Male Caucasian					-0.2					-0.3	-0.2	-0.1	0.5	0.2	1.0					
P: Pre-GCI Female Caucasian	0.3					0.5					0.4					1.0				
Q: Post-GCI Female Caucasian		-0.4					-0.3					0.1				0.3	1.0			
R: Final_Grade Female Caucasian			-0.1					0.2					-0.1			0.1	0.1	1.0		
S: CT Avg Female Caucasian				-0.3					-0.3					0.0		0.3	0.0	0.3	1.0	
T: Engage Female Caucasian					0.3					-0.1					0.0	0.2	0.3	0.6	0.2	1.0

I Correlation Coefficient I			
Insignificant		δ=<	0.1
Minor	0.1	>δ=<	0.3
Moderate	0.3	>δ=<	0.5
Strong	0.5	>δ	

Within populations

Between populations (Note: Correlations for unrelated variables are removed from the table)

were moderately correlated (δ=0.5) to pre-GCI scores (Table 2: column F) and engagement (Table 2: row J, column H).

Male Caucasian students: Male Caucasian students recorded moderately correlated engagement and final course grades (δ=0.5; Table 2: row O, column M). Pre- and post course GCI scores (Table 2: row L, column K) were also moderately correlated to post GCI results (δ=0.4; Table 2: column L), as were average conceptest scores (Table 2: row N).

Female Caucasian students: Female Caucasian student data showed only one strong correlation (δ=0.6), and that was between engagement (Table 2: row T) and final grades (Table 2: column R). All other within-group correlations were small or insignificant.

Between Group Correlations: Moderately significant correlations were found when variables were compared between population groups. Male and female minority student pre-GCI scores were correlated (δ=0.5; Table 2: row F, column A), and male minority post-GCI scores (Table 2: column B) were negatively correlated with post-GCI scores of all other demographic groups. Pre-GCI scores for female minority students (Table 2: column F) were correlated with both Caucasian males and females (δ=0.4 and 0.5). Other correlations between groups were either between variables that had no practical relationship and were not shown (e.g. minority male pre-GCI scores and minority female post-GCI scores) or were of little or no significance.

Interpretation and Discussion

Correct/Incorrect Responses: We considered an initial correct response followed by an incorrect answer choice to be the least desirable response sequence. The 5% of responses for which students answered correctly initially but changed to an incorrect response following peer instruction was similar to the 6% rate reported by Crouch and Mazur (2001). These data suggest that such responses should be expected regardless of ethnicity or gender (Figure 3). The 5% rate closely matches a four-answer multiple choice question occurrence probability of 6% for random guessing on two identical questions (probability increases to about 10% for a three answer question). Since students were awarded credit for answering the questions (whether correct or incorrect), it is possible that some students were simply guessing or answering randomly to fulfill course requirements (King & Joshi, 2008). It is also possible some of these responses simply represent input error. Such an error was possible, because the electronic response software provided signals when student responses were received, but did not display individual responses. However, ineffectual peer instruction also can not be ruled out. If guessing and input error accounted for most correct-incorrect responses, those answers provided little information relevant to student assessment or teaching. Additional studies are necessary to determine if correct-incorrect responses are important indicators of student learning when using this technology.

Correct/Correct Responses: A twice-correct answer was considered a positive result, because such a response suggested students initially understood the concepts and then validated that understanding by answering correctly a second time. The overall twice-correct answer rates found here closely matched the 40% rate reported by Crouch and Mazur (2001) and played a major role in understanding similarities and differences between populations. Since Caucasian male and female students were more likely to answer twice correct, their other major answer categories had proportionally fewer responses than those of minority students (Figure 3). Such an observation supports the contention that differences within diverse populations can be more important than differences between populations (Harper, 2009). When student data were sorted into two groups (>40% and <40% of responses twice correct), there was a strong correlation between engagement and final grade for both the high- and low-performing groups (δ=0.6). Such a finding was not surprising, because engagement is a proxy for attendance, which has been previously correlated to course success (Newman-Ford, Fitzgibbon, Lloyd & Thomas, 2008; Scott, 2000).

Incorrect/Incorrect: As with the correct-incorrect answer, a twice incorrect response was considered a negative outcome, because it suggests the peer-learning technique was not effective for the students that answered in this fashion. The 28% overall response rate for twice-incorrect questions was higher than was reported for physics (22% in Crouch & Mazur, 2001). The finding that over one quarter of all responses were incorrect after peer discussion was particularly troubling in light of the fact that 40% of responses were twice correct, because this suggests that more correct responses result from other learning than from peer instruction. Students were randomly organized into four-person learning teams to encourage in-class discussion during the peer instruction phase of the class. The correct answer for most of the conceptests was also the most popular answer when students were polled on the first attempt. Armed with that information and group discussion support, such a high level of twice incorrect answers was considered problematic. ANOVA and correlation analyses showed that there were indistinguishable differences (p >0.05) and correlations (-0.2<= δ <=0.2) between students who frequently answered in this fashion (>25% of registered twice incorrect responses) as compared to those who did so less often (< 25%) for all analysis variables. If a significant number of students in these classes were not actually discussing answers, there may have been little propensity for students to change their answers. Perhaps students simply failed to change answers to questions if they did not understand the concepts and dialog was not effective enough to clarify understanding. Additional research that focuses on group interactions during peer discussion is necessary to determine the extent to which group communication affects twice incorrect response patterns.

This is the first study to examine the contrasts in student performance by both gender and ethnicity using electronic response systems in large classes.

Incorrect/Correct Responses: The type of response sought when using conceptests with peer instruction was that of changing from an incorrect to a correct response (Table 1: score 3). Approximately 26% of student responses in this study were of this type, which is lower than the 32% reported by Crouch and Mazur (2001). These data suggest that peer instruction was nearly equally effective for all populations, but perhaps slightly more so for female minority students (who were ~7% more likely than any other demographic group to answer this way). Since minority students were more likely to miss these questions on the first attempt than Caucasian students, they were in a better position to benefit from this approach.

Overall Responses: Combined analyses of all the response data suggested that there were similarities and differences in the ways that diverse populations respond when using this technology and pedagogical approach. When comparing males and females, all meaningful variable correlations were small or insignificant, which supports the suggestion made by King and Joshi (2008) that electronic response systems did not significantly hinder male or female student success in engineering. Within the male population, moderate correlations between pre- versus post-GCI scores, engagement versus final course grades, and conceptest averages versus post-GCI scores were again identified. Within the female population, engagement was strongly correlated to final grades (δ=0.6), and other variables correlated poorly. When the responses of all minority students were compared to the responses of all non-minority students, all meaningful variable correlations related to performance were insignificant or small, which suggests that this pedagogy provided an inclusive approach to formative assessment. Strong to medium correlations related to GCI scores and engagement suggest that prior knowledge and attendance played the most important role in minority students' course success. This finding supports the use of this technology with these populations if doing so encourages attendance, as has been noted in previous studies (MacGeorge et al., 2008b).

All populations could benefit if twice incorrect responses were minimized. This pedagogy relies on the positive group synergies known to be generated when learning with peers in a low-stakes environment (Mazur, 1997). Students placed in groups working toward a common goal, as is implicit in peer learning, provides a pseudo- organizational structure with social norms. Because of this, organizational learning theory (Argyris & Schön, 1996) may be an appropriate lens through which to view student response patterns. Central to such learning is the ability to detect errors (wrong answers) and take appropriate action (select correct answers) when responding to future opportunities (questions). This requires that members of the learning team work effectively and that the culture of the group be conducive to constructive dialog between all members of the team (Bensimon, 2005). An environment that is conducive to constructive dialogue is one in which all students are comfortable asking questions of their group members when they are not certain of the correct answer or when they consistently answer twice incorrect. The social dialog presumed to occur during peer instruction is known to result in successful performance among minority students (Quaye, Tambascia, & Talesh, 2009). However, the twice incorrect data presented here suggest that the optimal type of dialog was not occurring as often as desired for all populations. Clearly, all student groups have high and low performers. More detailed observations of student discussions are needed to better understand the dynamics and implications of dialog occurring in these groups and the impact of those peer discussions on response distributions.

This is the first study to examine the contrasts in student performance by both gender and ethnicity using electronic response systems in large classes. Given the ubiquity of this technology on college campuses, these data are available in electronic archives for a wide range of classes. We encourage others to analyze their data to determine if the trends reported here apply more widely.

Conclusions

The similarities and differences in conceptest response patterns found here illustrate how data from electronic response systems can be used to evaluate a pedagogical technique such as peer-instruction. The relatively small percentage of correct-to-incorrect responses may simply be a function of operator error or lack of interest in the class activity. As a percentage of all responses within populations, males' and females' answers show very similar distributions, which implies that the pedagogical technique is gender neutral. Furthermore, the distribution for answer changes from incorrect to correct suggests that all demographic groups benefit nearly equally from peer discussions. Perhaps as expected, students who answer conceptual questions correctly the most often tend to score highest in course grades, and correct response rates are a moderate function of prior knowledge and attendance. However, the consistently high rate of twice incorrect answers for all groups, and particularly among minority males, is cause for concern. Better dialog within groups appears to be necessary for diverse student populations to benefit most effectively from this intervention.

Acknowledgements

Partial support for this work was provided by the National Science Foundation's Geoscience Education and DUE CCLI programs under Award Nos. 0506518 and 0716397. The authors thank Julie Libarkin and Suttida Rakkapao for their evaluation of statistical techniques used in this study. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

American Geophysical Union. (1994). Scrutiny of Undergraduate Geoscience Education: Is the Viability of the Geosciences in Jeopardy? Washington, DC: American Geophysical Union.

Anderson, L.W., & Krathwohl (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. New York: Longman.

Argyis, C. & Schön, D.A. (1996). *Organizational Learning II: Theory, method and practice*. Reading, MA: Addison-Wesley.

Bensimon, E.M. (2005). Closing the achievement gap in higher education: An organizational learning perspective. In A.J. Kezar (Ed.), *Organizational Learning in Higher Education*. *New Directions for Higher Education:* No. 131. 99-111. San Francisco: Jossey-Bass.

Bloom, B.S. (Ed.) (1956). Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain. New York ; Toronto: Longmans, Green.

Cox, A.J., & Junkin, W.F., III. (2002). Enhanced student learning in the introductory physics laboratory. *Physics Education*, 37 (1), 37-44.

Crouch, C.H., & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics*, 69, 970-977.

Gilbert, S.W. & Ireton S.W. (2002). Understanding models in Earth and space science. 129. Arlington, VA: NSTA Press.

Green, P.J. (2003). Peer instruction for Astronomy. 178. New York, NY: Prentice Hall.

Hake, R.R. (1998). Interactive-engagement versus traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66, 64-74.

Haladyna, T.M., Downing, S.M. & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education,* 15(2), 309-334.

Harper, S.R. (2009). Institutional seriousness concerning black male student engagement: Necessary conditions and collaborative partnerships. In S.R. Harper & S.J Quaye (Ed.s), *Student engagement in higher education: Theoretical perspectives and practical approaches for diverse populations*: (137-156). New York, NY: Routledge Taylor and Francis Group.

Hestenes, D., Wells, M. & Swackhammer, G., (1992). Force concept inventory. *The Physics Teacher*, 30, 141-158.

King, D.B. & Joshi, S. (2008). Gender differences in the use and effectiveness of personal response devices. *Journal of Science Education and Technology*, 17(6), 544-552.

Kortz, K.M., Smay, J.J. & Murray, D.P. (2008). Increasing learning in introductory geoscience courses using lecture tutorials. *Journal of Geoscience Education*, 56(3), 280-290.

Lasry, N., Maur, E. & Watkins, J. (2008). Peer instruction: From Harvard to the two-year college. *American Journal of Physics*, 76(11), 1066-1069.

Libarkin, J.C. & Anderson, S.W. (2005). Assessment of learning in entry-level geoscience courses: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education*, 53(4), 394-401.

MacGeorge, E.L., Homan, S.R., Dunning, J.B., Elmore, D., Bodie, G.D., Evans, E., Khichadia, S., & Lichti, S.M. (2008a). The influence of learning characteristics on evaluation of audience response technology. *Journal of Computing in Higher Education*, 19(2), 25- 46.

MacGeorge, E.L., Homan, S.R., Dunning, J.B., Elmore, D., Bodie, G.D., Evans, E., Khichadia, S., Lichti, S.M., Feng, B. & Geddes, B. (2008b). Student evaluation of audience response technology in large lecture classes. *Educational Technology Research and Development*, 56, 125-145.

Mazur, E. (1997). Peer instruction: A user's manual. 253. Upper Saddle River, NJ: Prentice Hall.

McConnell, D.A., Steer, D.N., Owens, K.D. & Knight, C.C. (2006). How students think: Implications for learning in introductory geoscience courses. *Journal of Geoscience Education*, 53(4), 462-470.

McTighe, J. & Lyman, F. (1988). Cueing thinking in the classroom: The promise of theory-embedded tools. *Education Leadership*, 45(7), 18-24.

National Research Council. (1997). Science teaching reconsidered. 88. Washington, D.C.: National Academy Press.

National Research Council. (2000). How people learn: Brain, mind, experience and school. Washington, D.C.: National Academy Press.

National Science Foundation. (1996). Shaping the future: New expectations for undergraduate education in science, mathematics, engineering, and technology. 76. Arlington, VA: National Science Foundation.

Newman-Ford, L., Fitzgibbon, K., Lloyd, S. & Thomas, S. (2008). A large-scale investigation into the relationship between attendance and attainment: A study using an innovative, electronic attendance monitoring system. *Studies in Higher Education*, 33(6), 699-717.

Pilzer, S. (2001). Peer instruction in physics and mathematics. *Primus*, 11(2), 185- 192.

Quaye, S.J., Tambascia, T.P. & Talesh, R.A. (2009). Engaging racial/ethnic minority students in predominately white classroom environments. In S.R. Harper & S.J Quaye (Ed.s), *Student engagement in higher education: Theoretical perspectives and practical approaches for diverse populations*, 157-178. New York, NY: Routledge Taylor and Francis Group.

Rao, S.P., & DiCarlo, S.E. (2000). Peer instruction improves performance on quizzes. *Advances in Physiology Education*, 24(1), 51-55.

Scott, V. (2000). The significance of attendance in large college earth science classes. *GSA Abstracts with Program*, 32(7), 491.

Smith, M.K., Wood, W.B., Adams, W.K, Weiman, C., Knight, J.K., Guild, T.T. & Su, T.T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science*, 323, 122-123.

Sokoloff, D.R., & Thornton, R.K. (1997). Using interactive lecture demonstrations to create an active learning environment. *The Physics Teacher*, 35, 340-347.

David Steer is associate professor, Department of Geology and Environmental Science, The University of Akron, Akron, OH 44325-4101. Correspondence concerning this article can be sent to <steer@uakron.edu>.

David McConnell is professor, Department of Marine, Earth and Atmospheric Sciences, North Carolina State University, Raleigh, NC 27695-8208

Kyle Gray is assistant professor, Department of Earth Science, University of Northern Iowa, Cedar Rapids, IA 50614-0018

Karen Kortz is associate professor, Department of Physics, Community College of Rhode Island, Lincoln, RI 02865.

Xin Liang is associate professor, Department of Education Foundation & Leadership, The University of Akron, Akron, OH 44325-4208