Assessment of Learning in Entry-Level Geoscience Courses: Results from the Geoscience Concept Inventory

Julie C. Libarkin — Department of Geological Sciences, 316 Clippinger Labs, Ohio University, Athens, OH 45701 libarkin@ohio.edu

Steven W. Anderson — Science Department, Black Hills State University, Spearfish, SD 57799-9102 steveanderson@bhsu.edu

ABSTRACT

Assessment of learning in entry-level college science courses is of interest to a wide variety of faculty, administrators, and policy-makers. The question of student preparedness for college instruction, as well as the effect of instruction on student ideas, has prompted a wide range of qualitative and quantitative studies across disciplines. In the geosciences, faculty are just beginning to become aware of the importance of conceptual change in instruction. The development of the Geoscience Concept Inventory (GCI) and application to the study of learning in entry-level geoscience courses provides a common framework from which faculty can evaluate learning and teaching effectiveness. In a study of 43 courses and 2500 students, we find that students are entering geoscience courses with alternative conceptions (sometimes called "misconceptions"), and in many cases are leaving the classroom with these alternative ideas intact. Comparison of pre- and post-test results show that students with the lowest pre-test scores show the most improvement, whereas those with higher pre-test scores show little, if any, improvement. We also find no relationship between self-reported teaching style and learning as measured by the GCI, suggesting significant research needs to be done to evaluate teaching effectiveness in geoscience classrooms.

INTRODUCTION

Learning is the goal of all instruction. Accurate assessment of learning is an important first step in determining the links between learning and teaching, and ultimately in developing instructional approaches that are effective and transferable to other classrooms and institutions. Some disciplines, primarily physics and math, have made significant headway into unraveling the complex relationships between learning and teaching, often through the application of learning research pioneered by people like Piaget and Driver (e.g., Redish, 1994). Ultimately these efforts strive to determine how people learn, factors that can influence learning, and innovations to the teaching environment that can improve learning for all participants. Significant effort has been made to disseminate effective teaching methods for use in college level geosciences courses (e.g., Digital Library for Earth System Education), although quantitative assessment research documenting this effectiveness has been slower to evolve.

Quantitative assessment instruments for college classrooms have been used in a variety of scientific disciplines, particularly for the evaluation of attitudes or conceptual understanding (e.g., Hestenes et al., 1992; Zeilik et al., 1999; Libarkin, 2001; Yeo and Zadnick, 2001; Anderson et al., 2002). The development of the Geoscience Concept Inventory (GCI) is a first step in determining how entry-level college courses are affecting our students and in identifying factors that influence learning. The GCI is a set of conceptually based questions geared towards fundamental concepts in the Earth Sciences, including foundational concepts in physics and chemistry. We developed the GCI over a two-year period; to date, 73 questions have been evaluated and validated using item analysis techniques from both classical test theory and item response theory, particularly Rasch analysis (Libarkin and Anderson, in preparation). We report here on the analyses of pre- and post-test results from 29 GCI questions administered to ~2500 students enrolled in 43 courses across the United States. These GCI questions covered concepts related to geologic time, plate tectonics, and the Earth's interior.

Previous Research - Assessment of learning in the geosciences has traditionally focused on K-12 students, with studies of college students or other adults only recently emerging (DeLaughter et al., 1998; Trend, 2000; Libarkin, 2001; Libarkin et al., 2005; Dahl et al., 2005). Qualitative studies are concentrated outside of the U.S. (e.g., Happs, 1984; Marques and Thompson, 1997; Trend, 2000; Dodick and Orion, 2003), with those of American students focusing primarily on pre-college populations (Schoon, 1992; Gobert and Clement, 1999; Gobert, 2000). Existing quantitative studies have dealt with attitudes (Libarkin, 2001), visualization (e.g. Hall-Wallace and McAuliffe, 2002), and logical thinking skills (McConnell et al., 2003). Quantitative study of student conceptual understanding in the geosciences lags far behind other disciplines.

The development of the Force Concept Inventory (FCI; Hestenes et al., 1992) in the early 1990's dramatically changed the way physicists viewed teaching and learning in college level physics courses. A sharp increase in studies related to conceptual change in college-level physics (see Kurdziel and Libarkin, 2001 for a discussion) has led to significant changes in physics instruction, as well as a new perspective on the importance of physics education research in academic physics (e.g., Gonzales-Espada, 2003). Subsequent development of quantitative instruments in other disciplines followed, including development in biology (Anderson, 2002), physics (Yeo and Zadnick, 2001), astronomy (Zeilik et al., 1999), and now, the geosciences (Libarkin and Anderson, in preparation).

METHODS

Design and Procedure - We developed the GCI over several years, with question generation and validation based upon a variety of qualitative and quantitative data (Libarkin et al., 2005; Libarkin and Anderson, in preparation). Determination of reliability and validity of test questions (also called 'items' by test developers) evolved through qualitative means, such as validation by experts in geosciences and education, and through

Figure 1. Map of the continental United States. Numbers indicate number of institutions in each state participating in this study.

quantitative evaluation of student test data. Test data were analyzed using classical test theory approaches, particularly item analysis, and through item response theory using simple Rasch models. The Rasch analysis resulted in development of a test scale that allowed scaling of raw test scores to more meaningful scaled scores, and also provided information on item discrimination. One test question (of 29 original questions) was removed from the analysis based upon gender discrimination in both the pre- and post-test data. Although several questions were modified between the pre- and post-test administration based upon analytical results and expert feedback, the ordering of these questions on the Rasch scale did not change significantly, and we concluded that item revision did not dramatically impact our ability to compare pre- and post-test results (Libarkin and Anderson, in preparation).

The 29 GCI questions were distributed as two test versions of 20 questions each, with eleven common questions and nine version-specific questions. Tests were randomly distributed to courses, with each version administered to roughly half of the students. One institution administered the test via computer and used only one version; one course from this institution also post-tested with the same version. Analysis of test data from all institutions indicated that the two versions were of similar difficulty, producing nearly identical Rasch scaling functions for conversion of raw to scaled scores. GCI data were collected in Fall 2002 from 43 courses at 32 institutions located in 22 states across the U.S. (Figure 1). Tested courses were introductory level, and included physical geology, oceanography, environmental science, historical geology, and specialty topic courses. Faculty from 21 public and six private four-year institutions, four community colleges or two-year institutions, and one tribal college participated (Table 1). Individual classes ranged from nine to 210 students, with most courses falling between 35 and 75 students. The GCI was administered during normal course hours during the first two weeks of the academic semester or quarter, and again during the last week or final exam period. Students were informed that collection of GCI data was part of a research study and were informed that participation was both voluntary and anonymous. 2500 students were pre-tested at the beginning of the Fall 2002 semester, and a subset of 1295 students from 30 courses were post-tested at the end of the semester. In addition, matched pre- and post-test results from 930 individual students were obtained and compared.

Institutional Type	Number of Schools	Number of Courses	Course Size (n students)
Four-year public	21	31	11 to 190
Four-year private	6	6	13 to 91
Two-year community college	4	6	15 to 82
Two-year tribal college	1	1	9

Table 1. Sample size, recruitment, and institutional setting.

Instructors of post-tested courses used a variety of teaching methods including lecture, demonstration, whole class discussions, small group activities, laboratory exercises, and technology. Instructors in the study provided their estimated breakdown of the time spent on each of these instructional strategies, and we have made an initial comparison that relates teaching style to changes in pre- to post-test results on the GCI. Teaching approaches varied greatly, such that the reported percentage of class time devoted to lecture ranged from 0-100%, demonstration ranged from 0-30%, small group work ranged from 0-50%, lab exercises ranged from 0-60%, and use of technology ranged from 0-100%. Faculty self-reporting of teaching approaches is probably less accurate than direct classroom observation (e.g., Johnson and Roellke, 1999), although our large data set prohibited direct observation of all studied courses.

Data Analysis - Developers of multiple-choice instruments for higher education generally perform classical item analysis on test results (e.g., Hestenes et al., 1992; Anderson et al., 2002). Item analysis is primarily used to observe the statistical characteristics of particular questions and determine which items are appropriate for inclusion on a final instrument. Classical Test Theory generally drives most item analysis, with focus on item difficulty and item discrimination, and thus item characteristics are tied closely to the population sampled. Item Response Theory (IRT), an alternative item analysis technique, assumes that the characteristics of a specific item are independent of the ability of the test subjects. IRT at its foundations is the study of test and item scores based upon assumed relationships between the trait being studied (i.e. conceptual understanding of geosciences) and item responses. Most researchers would agree that items on any test are generally not of equal difficulty, and in fact most published concept tests report "item difficulty", defined by the percentage of participants answering a specific item correctly. For example, Anderson et al. (2002) present a 20-item test on natural selection, with item difficulties ranging from 13-77%. In addition, discriminability reported for these items suggests a strong correlation between the difficulty of items and the overall score achieved by a student. This suggests, then, that some items are easier to answer than others. Because difficulty ranges so widely on this and most concept tests, the question of linearity must be addressed. Linearity implies that conceptual understanding is linearly correlated with raw test scores;

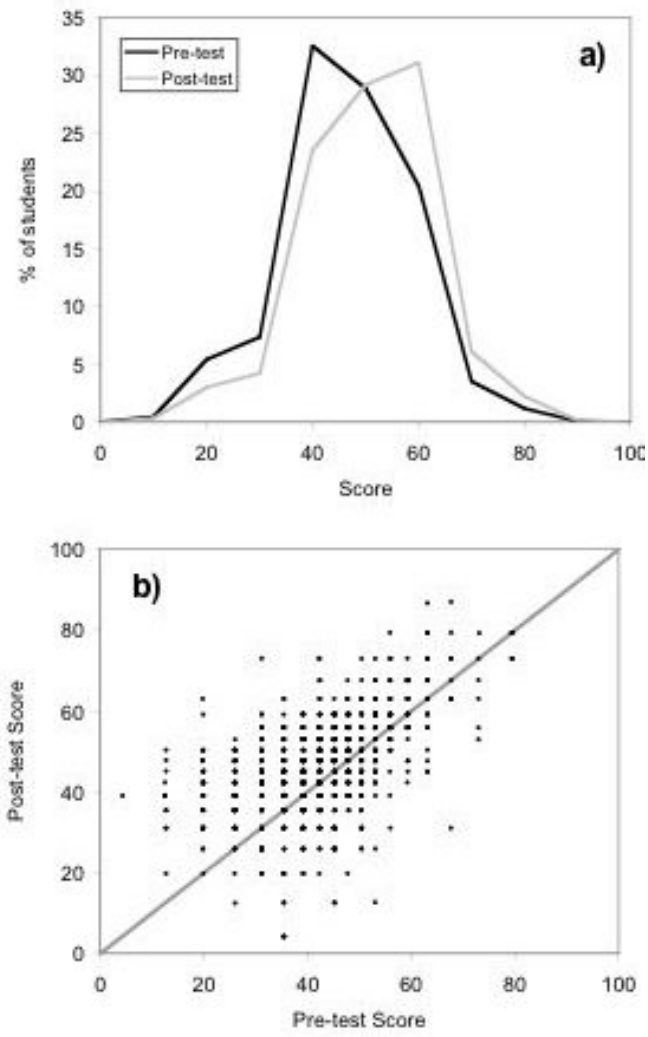


Figure 2. a) Distribution of scaled scores for all pre- (n = 2493 students) and post-tests (n = 1295 students). The lowest individual score was 0; the highest was 100. b) Matched pre and post-tests for individuals. The gray line represents the zone of no change; points falling along this line represent identical pre- and post-test scores. Points falling above the line indicate an increase in score from pre- to post-test and points falling below the line indicate a decrease from pre- to post-test.

a student answering 1/3 of items correctly has exactly half the understanding of a student answering 2/3 correctly.

Equivalent changes in raw score for multiple students may not translate to equivalent changes in conceptual understanding. IRT implies that not all test items are created equal, and some items will be more difficult than others. Rather than calculate a raw test score that simply reflects the number of "correct" responses, IRT allows for score scaling that more accurately reflects the difficulty of a given set of test items. Using a statistically calculated IRT scale to offset the assumption of scale linearity allows the determination of test scores that more accurately reflect "understanding". In addition, one of the test items included on both administered versions of the GCI exhibited gender bias as measured by Mantel-Haenszel differential item functioning. The removed item related to the distribution of earthquakes worldwide, and although it is unclear why this item behaved more favorably towards men it was removed prior to score scaling. All raw GCI test results were scaled on a 0-100% scale based upon a simple IRT approach (Rasch analysis), following the methodology presented by Libarkin and Anderson (in preparation). The relationship between raw score and scaled Rasch score, as fit by the statistical package JMP, is approximately:

$$S = 3.9 + 9R - 0.71R^2 + 0.025R^3 \quad (1)$$

where S is the scaled score on a 0-100% scale and R is the raw score on a 19-item GCI.

Pre- and post-test results were then compared using simple t-tests; this comparison was conducted for the entire population of students as well as sub-groups categorized from demographic or course information. All t-tests were two-tailed and based upon $p<0.05$, with some courses passing at the $p<0.001$ level.

RESULTS AND DISCUSSION

These data provide a unique opportunity to evaluate the pre-course conceptual frameworks of students enrolled in geoscience courses nationwide. In addition, evaluation of test data relative to course factors such as class size, institutional type, and faculty instructional approaches provides insight into the effectiveness of entry-level geoscience courses nationwide. Finally, preliminary evaluation of these data indicates that some ideas are stable across instruction, suggesting an until now unknown entrenchment of ideas (Anderson and Libarkin, 2003).

Overall, students found the test difficult, with nearly identical pre-test means of 41.5±12 (all 43 courses; n=2493 students) and 42.2±12 (for only the 29 courses post-testing, and where course #41 could not be included; n=1498 students). The post-test results suggest that the population of post-testing students experienced minimal learning over the course of the semester, with a mean of 45.8±13 (n=1295 students; Figure 2a). Results are most illuminating when pre- and post-test scores are matched for individual students; in this case, 930 pre- and post-tests were matched. The pre-test mean for students with matched post-tests (n=930 students; 43±11) is similar to all pre-tests. The matched post-test mean is also similar to overall results (47±12), with a small effect size of 0.17. With the exception of one course containing nine students, students on average were familiar with only half of the concepts covered by this test after completing these courses. We should point out that the 29-item GCI used in this portion of the study represented fixed content. Instructors may not have covered all of the represented content in their courses, although the limited difference in test scores across courses suggests that this factor is probably unimportant. Development of over 70 GCI questions since the completion of this study and the use of statistically similar sub-tests in GCI assessment removes any concerns about the impact of fixed content testing on results (Libarkin and Anderson, in preparation; see http://newton.bhsu.edu/eps/gci.html)

Comparison of matched pre- and post-tests (Figure 2b) indicates that statistically significant improvement occurred on the post-tests, as shown on a paired, two-tailed t-test (tstat = 1.96 < tcrit =12.1). Analysis of

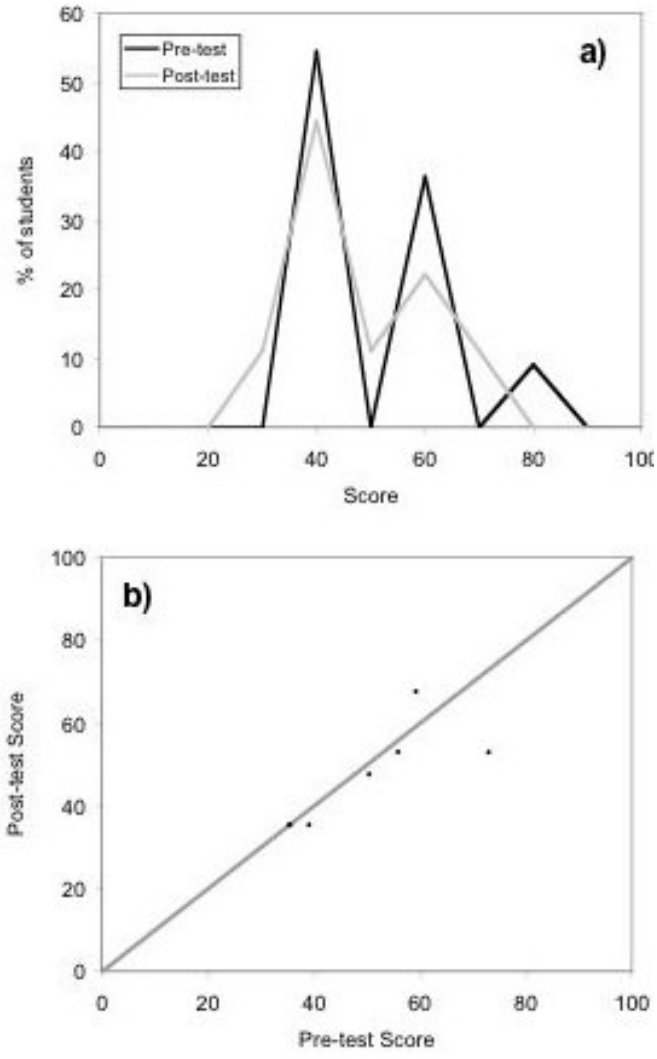


Figure 3. Course 19. a) Pre (n = 11) and post (n = 9) course distributions. Notice that the post-course distribution has shifted to the left, suggesting either 1) a decrease in conceptual understanding for some students; or 2) the two students who did not post-test were high scorers on the pre-test. b) 8 students had pre- and post-tests that could be matched. Notice that test scores do not change significantly for most individuals.

sub-populations indicates that students with low pre-tests (<40%, n = 388) dominate this effect, with a pre-test mean for this group of 32±7 and a post-test mean of 41±10, extreme significance on a t-test (tstat = 1.96 < tcrit =15), and an effect size of 0.46. Students with intermediate scores (40-60%, n = 489 students) exhibited a minimal change in GCI score, with a pre-test mean of 48±5, a post-test of 50±10, and minimal significance on a t-test (tstat = 1.96 < tcrit =3.5). Students pre-testing >60% (n = 52 students) exhibited no change in pre- to post-test scores (average score on both tests was 67%).

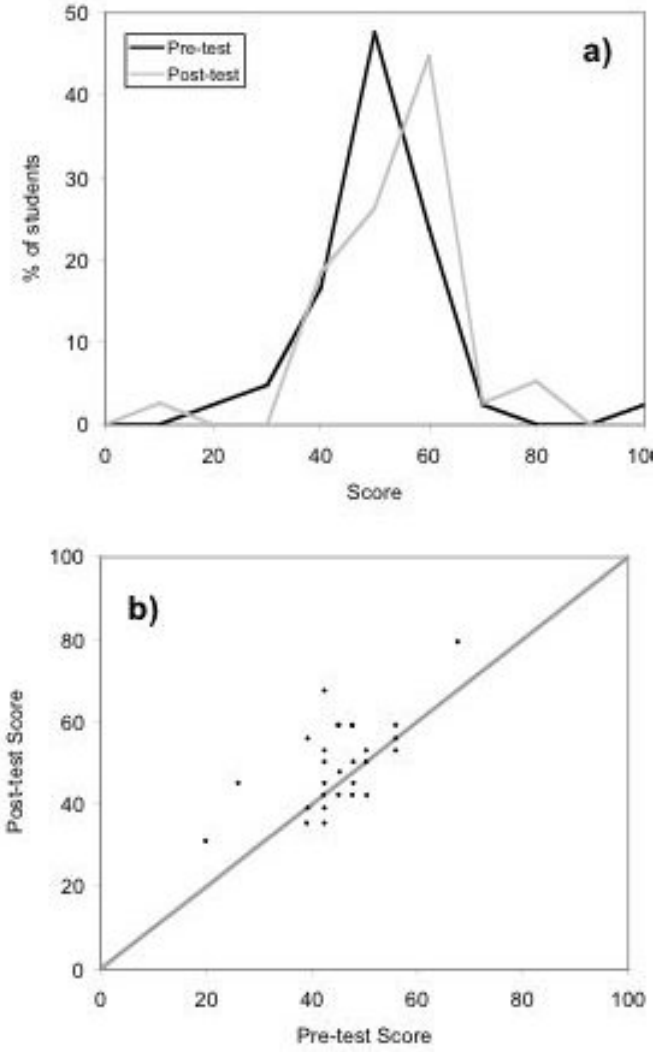
A brief discussion of the statistical phenomenon of regression towards the mean (e.g., Mee and Tin, 1994) is warranted here. Given a random population of test-takers, an individual has the highest probability of receiving an average score on a test. Scores that are lower or higher than average represent either true scores or bad or good luck on the part of the test-taker, respectively. Luck should not be constant across test administrations, and therefore scores are expected to move towards the average over multiple administrations. In our GCI study, the improvement of low scorers towards the mean and movement of some high scorers towards lower scores could be a result of this phenomenon. A statistical test of expected regression is possible, based upon the correlation between the results of two test administrations. The Pearson correlation coefficient, r, provides a measure of expected regression to the mean. If scores are standardized to z-scores, then the expected post-test score is equal to r times the pre-test score. For the matched tests in our data set, r = 0.56. This suggests that the average score of 32% for low achievers should increase to 37% simply as a result of regression. However, low achievers in this study showed an average increase in score to 41%, indicating that improvement is real.

Considering the phenomenon of regression toward the mean and the expected change in GCI scores from pre- to post-test provides valuable insight into our GCI results. The statistically significant change in score for the entire population as well as the improvement of low achievers from pre- to post-test indicates that: a) Low pre-test scores are a real phenomenon representative of students with low conceptual understanding as opposed to simply bad luck; b) The post-test increase past the expected regression level indicates that conceptual change is occurring for low achievers; and c) Low achievers are leaving courses with many alternative conceptions still intact. Similarly, the identical pre- and post-test averages for high achievers means that high pre-test scores are representative of conceptual understanding for this sub-population.

Overall, these data suggest that students with minimal knowledge at the beginning of an entry-level geology course are leaving with increased conceptual understanding, while students with intermediate and advanced understanding are leaving, as a population, with mixed effects. Those students with pre-test scores that are higher than their post-test scores may be using instruction to reinforce non-scientific conceptions. Interview data supports this hypothesis, suggesting that some students apply instruction in one area of geosciences to other areas. For example, the notion of Pangea is used by students to describe the Earth's surface at many different times in the past (Libarkin et al., 2005); students reinforce this idea by explaining that plate tectonics causes the continents to move. Further evaluation of the underlying causes of decreasing or increasing GCI scores, as well as changes in understanding of specific concepts, is needed.

Example Courses - Three courses have been chosen as representative samples of the courses tested. These include courses from different types of institutions, as well as courses of differing size that reflect the instructional strategies reported by participating faculty. Of the thirty post-tested courses, only 8 showed significance on a t-test. Overall raw scores indicate that the average student gained one question on the post-test, moving from 8 correct questions to 9 out of 19.

Course 19 is a representative small course taught at a public school in the south. The instructor of this course reported using traditional lecture and laboratory pedagogical approaches, with some alternative methods. The pre-course GCI average was 47±13 (n=11), with a



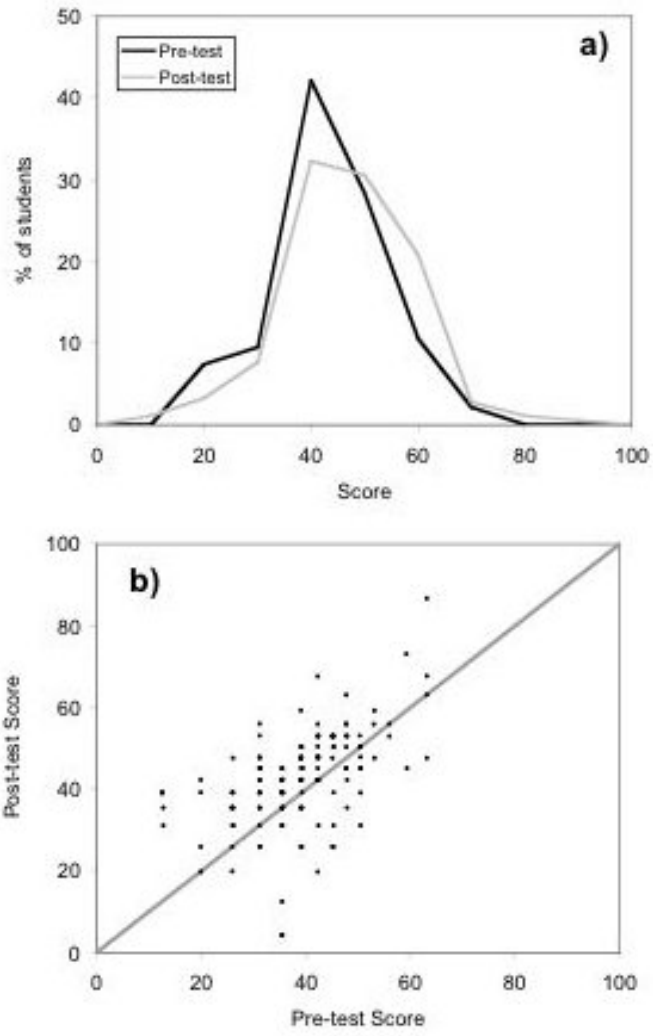


Figure 4. Course 3. a) Pre (n = 42) and post (n =38) course distributions. Notice that the post-course distribution has shifted to the right, suggesting an increase in conceptual understanding for some students. As with most courses, students with the poorest performance on the pre-test experienced learning as measured by this test. b) 28 students had pre- and post-tests that could be matched. Notice that the majority of students experienced an increase in test score at the end of the semester.

Figure 5. Course 12. a) Sample population is large enough to ascertain that distribution of pre (n = 190) and post (n = 183) course scores are both normal. Notice that the post-course distribution has shifted to the right, suggesting an increase in conceptual understanding for some students. As with most courses, students with the poorest performance on the pre-test experienced learning as measured by this test. b) 135 students had pre- and post-tests that could be matched. Notice that the effect of this course on individual students is mixed, although almost all low-performing students experienced significant gains.

post-test score of 43±13 (n=9; Figure 3a). Eight students were matched on pre- and post-tests; analysis of these matched tests indicates static GCI scores (Figure 3b). This suggests that student conceptions of the content covered by the GCI questions used in this study did not change as a result of instruction.

Course 3 is a representative intermediate course taught at a public school in the mid-west. The instructor of this course reported a predominantly lecture and in-class discussion approach to teaching. The pre-course GCI average was 46±12 (n=42), with a post-test score of 49±13 (n=38; Figure 4a). Matched pre- and post-tests for 28 students indicates that nearly all students experienced conceptual gain after instruction (Figure 4b). Gains were between one and two questions per student.

Course 12 is a representative large course taught at a public institution in the north-central U.S. The instructor of this course reported lecturing 100% of the time, using a traditional approach. The pre-course GCI average was 38±11 (n=190), with a post-test score of 42±12 (n=183; Figure 5a); the pre-course average for this class was much lower than the small and intermediate courses shown here. Analysis of matched tests (N=135) indicates that course effects varied, although the majority of students experienced static or positive gains (Figure 5b). As with the overall study population (Figure 2), those

Institution and Type	Code	Pre-Test n	Post-Test n	Course Type	Instrucitonal Methods %					
					L	D	IC	A	T	LAB
A, Public	1	69	25	Physical Geology	70	30				X
A, Public	2	38	26	Historical Geology	100					
B, Public	3	42	38	Geology	50	<10	30	<10	<10	
C, Public	4	24	23	Physical Science	60	30	<10			X
D, Public	5	81	67	Marine Science	50		50			
E, Public	6	29	25	Physical Geology	50	<10	<10	50	<10	
F, Public	7-1	57*	36	Geology	80	<10	<10			X
G, Public	7-2	-	16	Geology-online					100	X
H, Public, 2-yr	8	28	25	Physical Geology	60	<10	<10		<10	X
H, Public, 2-yr	9	21	13	Physical Geology	60	<10	<10		<10	X
I, Public	10	86	39	Unkown						
J, Public	11	108	85	Physical Geology	80		20			
K, Public	12	190	183	Earth Science	100					
L, Public	13	129	107	Physical Geology	60	<10	<10			X
M, Private	14	13	-	-	-	-	-	-	-	-
N, Public	15	40	-	-	-	-	-	-	-	-
O, Private	16	58	-	-	-	-	-	-	-	-
P, Public	17	15	16	Physical Geology	60		<10			X
P, Public	18	120	75	Unkown						
P, Public	19	1	9	Historical Geology	60	<10	<10		<10	X
Q, Navajo, 2-yr	20	9	6	Historical Geology	50	30	<10		<10	X
R, Public	21	67	57	Earth Systems	80		<10	<10		X
S, Public	22	50	54	Geology for Engineers	60	<10	<20		15	X
T, Public	23	18	17	Geology of National Parks	90	<10	<10		<10	X
U, Public, 2-yr	24	82	56	Physical Geology	65	15	10		10	X
V, Private	25	91	-	-	-	-	-	-	-	-
W, Private	26	54	-	-	-	-	-	-	-	-
X, Public	27	59	-	-	-	-	-	-	-	-
Y, Private	28-1	37*	19	Earth History	70	10	5		15	
Y, Private	28-2	-	14	Oceanography	70	5	5		20	X
Z, Private	29	24	22	Geology and Environment	45	10	40		5	X
AA, Public	30	69	-	-	-	-	-		-	-
BB, Public	31	31	22	Oceanography	45	5	15	25	10	X
CC, Public, 2-yr	32	24	-	-	-	-	-	-	-	-
DD, Public, 2-yr	33	39	-	-	-	-	-	-	-	-
EE, Public	34	18	-	-	-	-	-	-	-	-
FF, Public	35	75	55	Unkown	-	-	-	-	-	-
FF, Public	36	41	32	Unknown	-	-	-	-	-	X
GG, Public, 2-yr	37	15	13	Hyrogeology	90	10				X
HH, Public	38	21	-	-	-	-	-	-	-	-
II, Public	39	97	-	-	-	-	-	-	-	-
JJ, Public	40	128	-	-	-	-	-	-	-	-
KK, Public	41	269	120	Physical Geology	-	-	-	-	-	-

Table 2. Courses participating in GCI testing in Fall 2002. #41 combined several courses.

Topic	Conception*	Prior to Instruction	After Instruction
Techniques for Calculating Earth's Age	Analyses of fossils, rock layers, or carbon are the most accurate means for calculating the Earth's age	78% (n=1377)	72% (n=669)
Location of Tectonic Plates	The Earth's surface is not the top of the tectonic plates; tectonic plates are located beneath the Earth's surface.	56% (n=2483)	46% (n=1287)
Earth's surface when humans appeared	A single continent existed when humans first appeared on Earth	52% (n=2483)	47% (n=1287)
Life at Earth's formation	Simple, one-celled organisms existed when the Earth first formed	47% (n=2481)	43% (n=1286)
Appearance of dinosaurs	Dinosaurs came into existence about halfway through geologic time.	37% (n=1089)	40% (n=604)

Table 3. Prevalent ideas and persistence after instruction. *Students in the study population who did not exhibit these conceptions often held other alternative conceptions.

students who entered the course with pre-tests less than 40% experienced statistically significant positive gains.

Entrenchment of Ideas - The student population retained several alternative ideas over the course of the semester, with remarkably consistent results across institutions (Anderson and Libarkin, 2003; Table 3). The extreme persistence of some ideas suggests that current approaches to instruction, either traditional or alternative, may not be adequate for engendering conceptual change. In particular, students have a poor idea of the scale of geologic time, the occurrence of events in geologic history, and the specifics of absolute age dating. Not surprisingly, students also ascribe a Pangea-like supercontinent to many different times in the past, including the time of Earth's formation, and as noted here, at the appearance of humans. Although entry-level geoscience textbooks universally discuss the Theory of Plate Tectonics and most faculty spend significant time on this topic, most students are exiting courses with a poor understanding of the location of tectonic plates. Previous research utilizing qualitative approaches is in agreement with these data (Libarkin et al., 2005).

CONCLUSION AND IMPLICATIONS

The diverse data set collected in this study allows for a unique glimpse into entry-level geoscience courses being taught nationwide. Most notably, students are entering these courses with prior experiences in Earth Science and alternative conceptions about geologic phenomena. Post-instructional gains in understanding at the college level are generally small, with most students exiting courses with conceptions similar to those held prior to instruction. As an exception, those students entering college geoscience courses with little familiarity or significant misconceptions (and, thus, low GCI test scores) experience significant gain across all courses and institutions, regardless of instructional approaches. This universal improvement of students with very low pre-test scores is similar to findings in physics as tested with the FCI (Pollock, 2004) and with student attitudes (Libarkin, 2001). Most likely, these students are simply "catching up" with their peers and acquiring some of the simpler scientific concepts tested by the GCI.

Although the geoscience community has spent significant time and energy developing and disseminating alternative instructional strategies for use in college-level classrooms, the limited conceptual gain experienced by students suggests that a different curriculum-development approach is warranted. In particular, the effects of curriculum and pedagogy on student conceptual understanding, as well as the mechanisms for conceptual change in college-level geosciences, need to be studied in detail. Qualitative and quantitative research approaches have the potential to unravel the complex relationships between teaching and learning, and implementation of research approaches into the curriculum development-testing-dissemination cycle may result in significant modification in the way faculty view entry-level instruction. Certainly, further research in all realms of conceptual change in the geosciences is needed, with potential benefits to students and faculty alike.

ACKNOWLEDGMENTS

We thank all students and faculty who graciously participated in this research. Michaela Brewster, James Vinton, and Saunia Withers are thanked for their assistance with data entry. We thank Eric Flodin and Erin Frew for thoughtful suggestions regarding statistical implications of our study. This study was funded by the National Science Foundation through an Assessment of Student Achievement in Undergraduate Education grant to Libarkin and Anderson (DUE-0127765; DUE-0350395) and a Postdoctoral Fellowship in Science, Mathematics, Engineering, and Technology Education for Libarkin (DGE-9906478).

REFERENCES

Anderson, D.L., Fisher, K.M., and Norman, G.J., 2002, Development and validation of the conceptual inventory of natural selection, Journal of Research in Science Teaching, v. 39, p. 952-978.

Anderson, S.W., and Libarkin, J., 2003, The retention of geologic misconceptions: Alternative ideas that persist after instruction, EOS, v. 84, Abstract ED22E-07.

Dahl, J., Anderson, S.W., and Libarkin, J.C., 2005, Digging into Earth Science: Alternative conceptions held by K-12 teachers, Journal of Science Education, v. 12, p. 65-68.

DeLaughter. J.E., Stein, S., and Stein, C.A., 1998, Preconceptions abound among students in an Introductory Earth Science Course, EOS, v. 79, p. 429-432.

Dodick, J., and Orion, N., 2003, Cognitive factors affecting student understanding of geologic time, Journal of Research in Science Teaching, v. 40, p. 415-442.

Gobert, J.D., 2000, A typology of causal models for plate tectonics: Inferential power and barriers to understanding, International Journal of Science Education, v. 22, p. 937-977.

Gobert, J.D., and Clement, J.J., 1999, Effects of student-generated diagrams versus student-generated summaries on conceptual understanding of causal and dynamic knowledge in plate tectonics, Journal of Research in Science Teaching, v. 36, p. 39-53.

Gonzales-Espada, W.J., 2003, Physics education research in the United States: A summary of its rationale and main findings, Revista de Educacion en Ciencias, v. 4, p. 5-7.

Hall-Wallace, M.K., and McAuliffe, C.M, 2002, Design, implementation, and evaluation of GIS -based learning materials in an introductory geoscience course, Journal of Geoscience Education, v. 50, p. 5-14.

Happs, J.C., 1984, Soil generation and development: view held by New Zealand students, Journal of Geography, v. 83, p. 177-180.

Hestenes, D., Wells, M., and Swackhamer, G., 1992, Force Concept Inventory, The Physics Teacher, v. 30, p. 141-158.

Johnson, S.D., and Roellke, C.F., 1999, Secondary teachers' and undergraduate Education faculty members' perceptions of teaching effectiveness criteria, A national survey: Communication Education, v. 48, p. 127-38.

Kurdziel, J.P., and Libarkin, J.C., 2001, Research Methodologies in Science Education: Assessing Students' Alternative Conceptions, Journal of Geoscience Education, v. 49, p. 378-383.

Libarkin, J.C., 2001, Development of an assessment of student conception of the nature of science, Journal of Geoscience Education, v. 49, p. 435-442.

Libarkin, J.C., and Anderson, S.W., Science concept inventory development in higher education: A mixed-methods approach in the geosciences, Journal of Research in Science Teaching, in preparation.

Libarkin, J.C., Anderson, S.W., Dahl, J., S., Beilfuss, M., Boone, W., and Kurdziel, J.P., 2005, Qualitative analysis of college students' ideas about the Earth, Interviews and open-ended questionnaires: Journal of Geoscience Education, v. 53, p. 17-26.

Marques, L., and, Thompson, D., 1997, Misconceptions and conceptual changes concerning continental drift and plate tectonics among Portuguese students aged 16-17, Research in Science and Technological Education, v. 15, p. 195-222.

McConnell, D.A., Steer, D.N., and Owens, K.D., 2003, Assessment and active learning strategies for introductory geology courses, Journal of Geoscience Education, v. 51, p. 205-216.

Mee, R.W., and Tin, C.C., 1991, Regression towards the mean and the paired sample t-test, American Statistician, v. 45, p. 39-42.

Pollock, S.J., 2004, No single cause: Learning gains, student attitudes, and the impacts of multiple effective reforms, Paper presented at Physics Education Research Conference, summer 2004.

Redish, E. F., 1994, Implications of Cognitive Studies for Teaching Physics, American Journal of Physics, v. 62, p. 796.

Schoon, K.J., 1992, Students' alternative conceptions of Earth and space, Journal of Geological Education, v. 40, p. 209-214.

Trend, R., 2000, Conceptions of geological time among primary teacher trainees, with reference to their engagement with geoscience, history, and science, International Journal of Science Education, v. 22, p. 539-555.

Yeo, S., and Zadnick, M., 2001, Introductory thermal concept evaluation: Assessing Students' Understanding, The Physics Teacher, v. 39, p. 496-503.

Zeilik, M., Schau, C., and Mattern, N., 1999, Conceptual astronomy. II. Replicating conceptual gain, probing attitude changes across three semesters, American Journal of Physics, v. 67, p. 923-927.

NAGT

National Association of Geoscience Teachers

Membership Application or Renewal Form

Name: _____ Phone: _____

Mailing Address: _____ Fax: _____

_____ Email: _____

City: _____ State: _____ Zip: _____

___College/University Professor @ _____
___Precollege Teacher @ _____
___Other @ _____

Checks, MasterCard, or VISA (US funds only) are payable to:National Association of Geoscience Teachers. Mail to: NAGT, PO Box 5443, Bellingham, WA 98227-5443

Membership	**Rates (US funds)**
Regular USA	$35 _____
Outside USA	$47 _____
Student–USA	$20 _____
Student–outside USA	$32 _____
Retired NAGT member	$30 _____
Library Subscriptions	
Regular USA	$55 _____
Outside USA	$67 _____
_____New	_____Renewal

- ❑ Check
- ❑ Credit card: MC/VISA (circle one) Number: _____

Signature: _____ Exp. Date _____

The *Journal* and membership year runs from January to December. Subscriptions received after June 1 will begin receiving the *Journal* in January of the following year. Back issues are available for $15 (foreign $18) each.

*To qualify for student rate, indicate and obtain verification from a NAGT member:

___Undergraduate ___Graduate _____

Signature of NAGT member School