Diagnostic Testing of Introductory Geology Students

Cinzia Cervato
Department of Geological and Atmospheric Sciences, Iowa State University, 253 Science I, Ames, IA 50011-3212, cinzia@iastate.edu

James A. Rudd, II
Department of Chemistry and Biochemistry, California State University at Los Angeles, 5151 State University Drive, Los Angeles, CA 90032, jrudd@calstatela.edu

Vivian Z. Wang
Department of Chemistry and Biochemistry, California State University at Los Angeles, 5151 State University Drive, Los Angeles, CA 90032

ABSTRACT

A diagnostic test for assessing the general and Earth science knowledge of entry-level college students was administered to 451 students in 2002 and 401 students in 2003 enrolled in an introductory geology course at Iowa State University. The study shows that male students, seniors, and science-technology-math majors score higher than female students, freshmen, and non-science-technology-math majors and that the differences are statistically significant. Also, students who scored higher on the diagnostic test were more likely to pass the course. The results support the feasibility of a standardized diagnostic test as a tool for geoscience instructors for curriculum planning, student advising, and curriculum assessment, similar to standardized diagnostic testing and pre-post testing used in chemistry and physics courses. Standardized national tests would enhance college geoscience education.

INTRODUCTION

In the last decade there has been a dramatic increase in research aimed at studying and improving geoscience education. The ever-increasing number of manuscripts submitted to the Journal of Geoscience Education (JGE) for publication unequivocally signifies this change (Drummond, 2003). Most of the articles in JGE describe innovative techniques devised to improve student learning or to engage students in the study of Earth sciences. Compared to geoscience education, the production of education literature in general is monumental in scope and size: the Education Research Information Center, or ERIC, the leading educational database, contains more than one million citations and abstracts from over 700 educational journals and thousands of reports. The overwhelming volume of literature published in this field has one common goal: to enhance student learning.

When developing a new teaching technique, revising a syllabus to incorporate innovative activities, or designing a new curriculum, the question that each instructor naturally would ask is, "Will it improve learning of the subject matter?" In other words, how much more or better will students learn with the new approach? Because curricular innovations require time and effort, instructors must find them to be worthwhile. But how do we measure learning? Assessment is a critical part of planning educational research, and a very strong emphasis is currently placed on the development of successful assessment techniques. The recommended method is to give "before" and "after" exams to both an experimental and a control group.

The problems with this approach are multiple and well known: is the grading scale the same? Is the new technique the only part that changed in the course, or has the instructor revised other aspects of the course as well? How much time did the students spend on the activity? Is the demographic make-up of the courses the same? Are the tests comparable in length, type, and difficulty? On a larger scale, instructors may wonder what and how much their students are learning compared to students in other schools or other States. One way, and possibly the only way, to find an answer to all of these questions would be to create standardized national tests for the geosciences. Chemistry and physics instructors have actively used such tests for decades. For introductory college physics and chemistry courses, diagnostic tests have been developed in the last 10 years, and they are beginning to be used as statewide examinations (e.g., Krishnan and Howe, 1994; Russell, 1994; Steinberg and Sabella, 1997; McFate and Olmsted, 1999; Legg et al., 2001). Recently, the California Chemistry Diagnostic Test (CCDT) (Russell, 1994) was used to analyze the probability of success of students in general chemistry with a logistic regression analysis (Legg et al., 2001), so the CCDT can be used to predict student success and to advise students about their readiness when they start the course. In an effort to establish national standards in the understanding of chemistry, the Division of Chemical Education of the American Chemical Society has been providing K-16 instructors with standardized tests since 1934 (http://www.uwm.edu/Dept/chemexams/INTRO/index.html). Similar tests are available for physics instructors (http://www.psrc-online.org/ under "Evaluation instruments"). These tests allow instructors to assess student knowledge to conduct pre-post testing of curriculum effectiveness, to compare local results to the national level, and to compare against the national science standards. By giving a standardized test at the beginning and end of the course, the instructor can assess individual student learning and how the students in the course compare to students at peer institutions, across the State, or nationwide.

With a similar goal, the American Geological Institute initiated the Earth Science Curriculum Project in the 1960's, which attempted to establish standardized tests for high school students. These tests, however, are no longer used, and there is no national exam on Earth science knowledge. The National Science Education Standards (National Research Council, 1996) and AAAS Benchmarks for Science Literacy have emphasized again the critical role of the Earth sciences in science education and the need for content standards in Earth sciences. After this study was conducted, Libarkin and Anderson (2005, 2006) published their Geoscience Concept Inventory (GCI), a pool of 73 multiple-choice questions covering various aspects of physical geology and fundamental physics and chemistry concepts (http://newton.bhsu.edu/eps/gci.html). Similarly, McConnell et al. (2006) have developed a large database of geoscience ConcepTests available online (http://serc.carleton.edu/introgeo/interactive/ctestexm.html).

Establishing a national level of Earth science education would require the creation of standardized tests similar to the ones that exist for chemistry. The Geoscience Concept Inventory and the geoscience ConcepTests are excellent starting points for this community initiative. This paper presents the results from our attempt to develop and implement a diagnostic test for an introductory geology course for the purposes of (1) measuring incoming student knowledge of geology and science from high school or previous science courses, and (2) testing the feasibility of a standardized diagnostic test for introductory geology courses.

METHODS

Diagnostic Test - As a starting point to test the feasibility of a standardized, multiple-choice diagnostic test similar to the CCDT, in 2002 we selected questions from the New York State Regents Earth Science Exams (http://www.emsc.nysed.gov/ciai/testing/scire/regentearth.html) as well as four questions from the Group Assessment of Logical Thinking (GALT) test (Roadrangka et al., 1983). The New York State Regents Earth Science Exams have been administered by the New York State Education Department to high-school seniors for many years and old exams are available on line. Some challenge the validity of these statewide standardized tests (e.g., Olson, 2006) and we could not find any information on how the questions are selected and if they go through a process of validation as rigorous as the Geoscience Concept Inventory (Libarkin and Anderson, 2005). However, these were the only independently developed tests of high-school-level Earth science knowledge based on the National Science Education Standards that were available at the time of the study. Because these tests were designed for high school students, they were deemed appropriate and valid for measuring Earth science knowledge at a high school level.

Questions were selected by two colleagues of the senior author to avoid potential bias. The 41 questions used on the test in 2002 and the 40 questions used in 2003 were grouped into four types: general science, geology, mathematics, and logic questions (see Electronic Appendix for the complete text of the exams). Some questions may be considered to be more than one type, so we asked a science colleague not involved in the study to assign each question to one of the four types to avoid researcher bias. The questions were mainly on geology and general science topics (2002: 17 geology, 15 general science, 4 math, 5 logic; 2003: 13 geology, 19 general science, 5 math, 3 logic). Nineteen questions from the 2002 exam were re-used on the 2003 exam (9 general science, 3 geology, 4 math, 3 logic). Since the goal of the study was to test the students' incoming knowledge of geology and general science, and most of them acquired this in high-school, we chose to use a high-school level test even if the questions are mainly based on memorization of facts instead of critical thinking. In fact, most of the questions can be categorized as testing skills in the lower half of Bloom's Taxonomy, i.e., knowledge, comprehension, and application, rather than the upper half of analysis, synthesis, and evaluation. For the same reason, we chose to not administer the same test at the end of the semester and instead evaluate the students' progress during the semester using the combination of assessment tools (homework, in-class assignments, tests) based on deeper conceptual understanding used in the class.

Study Sample - To give context to student performance on the diagnostic test, Iowa State University (ISU) is a large, Midwestern, land-grant research institution with approximately 21,000 undergraduate and 5,000 graduate students. For most Iowa high school graduates, if they rank above the 49th percentile in their graduating class and have completed the required courses, they are automatically admitted to ISU. Students with lower rankings (20th-49th percentiles) must additionally achieve minimum ACT and SAT I scores to receive probationary admission. The high school science preparation required for ISU admission is a total of three years distributed across at least two subjects from among biology, chemistry, and physics. Earth sciences are not required, and as a result, many high school students do not take Earth science or, rarely, take it in 9th grade. These requirements are typical of many States, but various organizations, including the American Geological Institute, are attempting to change the requirements to include the Earth sciences (Robert Ridky, personal communication, 2003).

The diagnostic test was administered during the second class-meeting of Geology 100 (a three-credit, lecture-only class) in Fall 2002 and Fall 2003, and the test counted as extra credit towards the final course grade. Geology 100 - The Earth - is a traditional physical geology course designed for mainly non-science majors fulfilling a science requirement. The 2002 test results come from 451 students enrolled in two sections (Section A, n=228; Section B, n=223) taught by the same instructor (Cervato). Similarly, the 2003 test results come from 401 students (Section A, n=229; Section B, n=172). The 451 students in 2002 consisted of 202 females and 249 males, and the 401 students in 2003 consisted of 195 females and 206 males. Students were categorized as SMT (science, math and technology) majors or non-SMT majors, and most students were non-SMT majors (2002: 82 SMT, 369 non-SMT; 2003: 71 SMT, 299 non-SMT). Major data for 31 students in 2003 were not available. Freshmen represented the largest class of students (2002: 180/451; 2003, 167/401). Class standing data for 21 students in 2003 were not available. Students were categorized as passing students by earning a course grade of C or higher, and failing students had grades of C-, D, F, and W. Student data are shown in Tables 1-5.

RESULTS

Diagnostic Test and Passing of Course - The overall diagnostic test average for 2002 (70.3%) was higher than the 2003 average (66.4%) (Table 1), however, the 2002 test had one more question than the 2003 test. Because the diagnostic test averages and standard deviations were practically the same in both sections in a given year, only the combined data are shown for each year in Table 1.

Table 1 also shows the 2002 and 2003 data by gender. A t-test comparing the mean scores by gender yielded statistical evidence that male students score higher on the diagnostic test than female students (t=7.086, df=850, p=0.000). The results from a Pearson chi-square test comparing the pass-fail proportions by gender were not statistically significant (2=0.777, df=1, p=0.378).

	2002	2003	Males			Females		
			2002	2003	All	2002	2003	All
N Students	451	401	249	206	455	202	195	397
DTest AVE%	70.3	66.4	73.6	69.6	71.8	66.2	62.9	64.6
DTest SD	14.1	15.6	13.9	15.7	14.7	14.3	15.9	15
N fail	105	78	59	44	103	46	34	80
N pass	36	323	190	162	352	156	161	317
% passing	76.7	80.5	79.6	78.6	77.4	77.2	82.6	79.8

Table 1. Student numbers, diagnostic test averages and standard deviations, and pass-fail numbers and percentage by year and gender.

	SMT Majors			Non-SMT Majors			Freshmen			Seniors		
	2002	2003	All	2002	2003	All	2002	2003	All	2002	2003	All
N Students	82	71	153	369	299	668	180	167	347	48	33	81
DTest AVE%	77.4	71.5	74.6	68.7	64.3	66.7	69.5	64.6	67.2	75.0	66.7	71.6
DTest SD %	13.5	14.4	13.9	14.3	15.3	14.8	14.5	15.2	14.8	16.3	15.4	15.9
N fail	12	8	20	93	64	157	48	30	78	7	7	14
N Pass	70	63	133	276	235	511	132	137	269	41	26	67
% passing	85.4	88.7	86.9	74.8	78.6	76.5	73.3	82.0	77.5	85.4	78.8	82.7

Table 2. 2002 and 2003 diagnostic test averages and standard deviations and pass-fail rates by major and for freshmen and seniors.

	All	Passing Students			Failing Students		
		2002	2003	All	2002	2003	All
N students	852	346	323	669	105	78	183
DTest AVE%	68.4	71.7	67.0	69.4	65.8	63.9	65.0
DTest SD%	15.1	13.9	15.4	14.6	15.7	18.3	16.8
% passing	78.5	76.7	80.5				

Table 3. 2002 and 2003 diagnostic test averages and standard deviations and pass-fail rates for passing and failing students.

Table 2 shows the combined 2002 and 2003 data by major and for freshmen and seniors. A t-test comparing the mean scores by major yielded statistical evidence that SMT majors score higher on the diagnostic test than non-SMT majors (t=5.908, df=819, p=0.000). The results from a Pearson chi-square test comparing the pass-fail proportions by major were statistically significant (2=8.010, df=1, p=0.005), meaning that SMT majors pass the geology course more frequently than non-SMT majors. A t-test comparing the mean scores of freshmen vs. seniors yielded statistical evidence that seniors score higher on the diagnostic test than freshmen (t=2.423, df=426, p=0.016). The results from a Pearson chi-square test comparing the pass-fail proportions of freshmen vs. seniors were not statistically significant (2=1.050, df=1, p=0.305).

Table 3 shows the 2002 and 2003 data for passing and failing students, and a t-test comparing the mean scores of passing students vs. failing students yielded statistical evidence that passing students score higher on the diagnostic test than failing students (t=3.394, df=850, p=0.001), evidence for a relationship between diagnostic test score and probability of passing the course.

Student Performance on 19 Common Questions - The 2002 and 2003 test averages in Tables 1 and 2 showed in each year the same patterns of males scoring higher than females, SMT majors scoring higher than non-SMT majors, and seniors scoring higher than freshmen. Student results from the 19 questions that appeared on both tests were analyzed to examine more closely the year-to-year consistency of student performance. Similar to the patterns in performance on the entire diagnostic test, males scored higher than females, SMT majors scored higher than non-SMT majors, and seniors scored higher than freshmen on the 19 questions, however, as a group, the 2002 students performed almost identically to the 2003 students (Tables 4 and 5). A t-test comparing the 2002 and 2003 mean performances on the 19 questions for all students was not statistically significant (t=0.312, df=850, p=0.755). Also, students in particular subgroups performed nearly the same as students in the same subgroup from a different year, e.g., the mean score for 2002 males was almost the same as the mean score for 2003 males (Tables 4 and 5). T-test results from comparisons of 2002 and 2003 mean performances for subgroups were not statistically significant (2002 males vs. 2003 males: t=0.472, df=453, p=0.637; females: t=1.331, df=395, p=0.184; SMT majors: t=0.791, df=151, p=0.430; non-SMT majors: t=0.169, df=666, p=0.866; freshmen: t=0.066, df=345, p=0.948; and seniors: t=1.786, df=79, p=0.078). There was no statistical evidence to support that students in different years perform significantly different on the 19 common questions. Therefore, the 2002 and 2003 data were combined for the remaining analyses presented in this study.

	All Students			Males			Females		
	2002	2003	All	2002	2003	All	2002	2003	All
N students	451	401	852	249	206	455	202	195	397
AVE %	76.8	77.1	76.9	80.6	79.9	80.3	72.1	74.2	73.1
SD %	15.1	15.1	15.1	14.7	15.1	14.8	15.5	15.2	15.4

Table 4. 2002 and 2003 averages and standard deviations for the 19 common questions for all students, male students, and female students.

	SMT Majors			Non-SMT Majors			Freshmen			Seniors		
	2002	2003	All	2002	2003	All	2002	2003	All	2002	2003	All
N students	82	71	153	369	299	668	180	167	347	48	33	81
AVE %	83.1	81.2	82.2	75.4	75.6	75.5	75.9	75.8	75.9	83.2	76.9	80.6
SD %	15.2	14.1	14.7	15.4	15.1	15.2	15.6	14.4	15.0	16.5	14.5	15.9

Table 5. 2002 and 2003 averages and standard deviations for the 19 common questions by major and for freshmen and seniors.

Question	Type	Level	Diff	Disc
24	math	1	94.2	0.15
29	general science	1	92.7	0.16
28	math	1	90.5	0.21
25	general science	1	88.6	0.25
20	math	1	87.7	0.29
34	general science	1	86.3	0.33
22	general science	1	85.6	0.21
9	general science	2	83.1	0.31
8	geology	2	82.7	0.35
3	general science	1	80.9	0.29
17	geology	1	80.5	0.34
35	general science	1	79.0	0.70
41	logic	4	67.5	0.70
40	logic	3	66.7	0.63
36	math	1	66.4	0.36
15	logic	4	65.1	0.42
2	geology	3	63.3	0.47
30	general science	1	63.1	0.50
37	general science	1	37.4	0.53

Table 6. Type, Bloom's Taxonomy Level, and Item analysis results for 19 common questions ranked by DIFF values (DIFF = Item Difficulty value, DISC = Discrimination Index value). Levels: 1 = Knowledge, 2 = Comprehension, 3 = Application, 4 = Analysis.

Areas of Student Strengths and Weaknesses - Item analysis of student answers from 852 students for the 19 common questions produced Item Difficulty values and Discrimination Index values for the 19 questions (Table 6). For a particular question, the Item Difficulty (DIFF) equals the percent of students choosing the correct answer, and a high DIFF indicates a higher percent of correct answers, i.e., Question 24 in Table 6 was the "easiest" question for students to answer. The Item Discrimination Index (DISC) is a measure of how effectively a particular question distinguishes between high and low performing students on the test as a whole. Positive values for DISC mean that students who scored above the test average had more success answering the question than students who scored below the test average. DISC values close to zero indicate that above-average and below-average students had approximately equal success on the question. All 19 questions had DISC values indicating that the questions successfully differentiated student performance, although the easiest questions have values approaching zero, which is common for questions that nearly all students answer correctly.

The test questions with a DIFF values above the test average by one standard deviation were considered areas of student strength. This cut-off value was 76.9+15.1% = 92.0% (Table 4), identifying Questions 24 and 29 as specific strengths of the students (Table 6). Question 24 was a math question on converting from scientific notation into standard notation, and Question 29 was a general science question about a brief definition of evolution.

The test questions with DIFF values below the test average by one standard deviation were considered areas of student weakness. This cut-off value was 76.9-15.1% = 61.8% (Table 4), identifying Question 37 as a specific weakness of the students (Table 6). Question 37 was a general science question about the most abundant atmospheric gas, and oxygen was the most selected response.

Using the same cut-off values to analyze for areas of strength and weakness for particular subgroups yielded similar results with some minor differences (Tables 7 and 8). Male students scored above 92.0% and below 61.8% on the same questions as all 852 students, with the exception that males also scored above 92.0% on Question 28, which was a math question about estimating the meter unit to the nearest English unit.

Female students scored above 92.0% on Question 24 only and scored below 61.8% on Questions 37, 40, 41, 30, 36, and 2 (Tables 7 and 8). Questions 40 and 41 were paired logic questions regarding conceptual understanding of displacement of equal volume spheres that had unequal masses. The correct answer was most selected, and the next most selected choice was based on greater mass leading to greater displacement. Question 30 was a general science question on remembering latitude vs. longitude. The correct answer was most selected, and the other answer was the next most selected. Question 36 was a math question on the definition of kilogram, with the correct answer being most selected, and the next most selected choice being off by a factor of 10. Question 2 was a geology question about surface runoff, with the correct answer being most selected, and the next most selected choice being an incorrect relationship between surface runoff and gradient.

All		Males		Females		SMT Majors		Non-SMT Majors		Freshmen		Seniors	
N = 852		N = 455		N = 397		N = 153		N = 668		N = 347		N = 81	
DIFF	Q #	DIFF	Q #	DIFF	Q #	DIFF	Q #	DIFF	Q #	DIFF	Q #	DIFF	Q #
94.2	24	95.6	24	92.7	24	93.5	29	94.6	24	96.0	24	97.5	29
92.7	29	94.9	29			93.5	24	92.5	29	91.9	29	95.1	24
		93.2	28										

Table 7. Questions indicating areas of student strength on 19 common questions (DIFF = Item Difficulty value).

All		Males		Females		SMT Majors		Non-SMT Majors		Freshman		Seniors	
N = 852		N = 455		N = 397		N = 153		N = 668		N = 347		N = 81	
DIFF	Q #	DIFF	Q #	DIFF	Q #	DIFF	Q #	DIFF	Q #	DIFF	Q #	DIFF	Q #
37.4	37	45.7	37	28.0	37	38.6	37	36.7	37	34.6	37	43.2	37
				55.2	40	58.6	2			61.4	15		
				56.2	41	58.8	30						
				57.7	30								
				57.9	36								
				60.4	2								

Table 8. Questions indicating areas of student weakness on 19 common questions (DIFF = Item Difficulty value).

	All	SMT Males	SMT Females
N Students	852	107	46
AVE %	76.9	83.8	78.4
SD %	15.1	14.8	13.8
N fail	183	11	9
N pass	669	96	37
% passing	78.5	89.7	80.4

Table 9. Averages and standard deviations for the 19 common questions, pass-fail numbers, and passing percentages for all students and male and female SMT majors.

SMT majors scored above 92.0% and below 61.8% on the same questions as all 852 students, with the exception that SMT majors also scored below 61.8% on Questions 2 and 30, two questions for which female students also scored below 61.8%. Non-SMT majors scored above 92.0% and below 61.8% on the same questions as all 852 students, an expected result because of the large number of non-SMT majors in the course. Freshmen students scored above 92.0% on Question 24 only and scored below 61.8% on Questions 37 and 15. Question 15 was a logic question about the direct relationship between pressure and melting point of a substance, with correct answer being most selected, and the next most selected choice based on confusing ambient temperature with melting point of a substance. Seniors scored above 92.0% and below 61.8% on the same questions as all 852 students.

Analysis of SMT Majors by Gender - The responses by SMT majors on the 19 common questions were analyzed by gender (Table 9). A t-test comparing the mean scores by gender yielded statistical evidence that male SMT majors score higher on the diagnostic test than female SMT majors (t=2.127, df=151, p=0.035), however, an ANOVA test showed no evidence for no interaction effects between the factors of gender, major, and pass-fail status. The results from a Pearson chi-square test comparing the pass-fail proportions by gender were not statistically significant (2=2.441, df=1, p=0.118).

The same cut-off values of 92.0% for areas of strength and 61.8% for areas of weakness were used to study student performance by major and gender on the 19 common questions (Tables 10 and 11). Although male SMT majors scored above 92.0% on the same questions as all 852 students, they additionally showed Question 34 as an area of strength. Question 34 was a general science question on identifying the unit of volume within a list of various units. Female SMT majors showed areas of strength much different than male SMT majors and much different than female non-SMT majors. Question 28 was the math question on estimating the meter unit to the nearest English unit, an area of strength identified for male students in general. Question 25 was a general science question on identifying the element within a list of metal alloys and one metal. Examining the data for non-SMT majors by gender yielded the same areas of strength identified previously for students of that gender, as expected due to the large number of non-SMT majors.

Male SMT majors scored below 61.8% on the same questions as all 852 students, all male students, and male non-SMT majors. Female SMT majors scored below 61.8% on the same questions as all female students. Female non-SMT majors scored below 61.8% on the same questions as all female students, except Question 15 replaced Question 2 as an area of weakness. Question 15 was the logic question on the direct relationship between pressure and melting point of a substance, which was identified as an area of weakness for freshmen, and female non-SMT majors had similar difficulty.

DISCUSSION

Overall Performance and Relationship to Pass-Fail - On each diagnostic test and on the 19 common questions, males scored higher than females, SMT majors scored higher than non-SMT majors, and seniors scored higher

All		SMT Males		SMT Females		n-SMT Males		n-SMT Females	
N = 852		N = 107		N = 46		N = 355		N = 333	
DIFF	Q #	DIFF	Q #	DIFF	Q #	DIFF	Q #	DIFF	Q #
94.2	24	94.9	24	93.2	28	95.8	24	93.2	24
92.7	29	94.9	29	93.2	25	95.2	29		
		92.4	34			93.8	28		

Table 10. Questions indicating areas of student strength for SMT majors on 19 common questions (DIFF = Item Difficulty value).

All		SMT Males		SMT Females		N-SMT Males		N-SMT Females	
N = 852		N = 107		N = 46		N = 335		N = 333	
DIFF	Q #	DIFF	Q #	DIFF	Q #	DIFF	Q #	DIFF	Q #
37.4	37	44.3	37	32.4	37	45.1	37	27.0	37
				48.6	40			56.6	40
				50.0	36			56.6	41
				51.4	30			59.5	30
				52.1	2			59.8	36
				54.1	41			61.7	15

Table 11. Questions indicating areas of student weakness for SMT majors on 19 common questions (DIFF = Item Difficulty value).

than freshmen. These results can be reasonably expected. The well-known gender difference in performance on standardized tests based on math and spatial visualization skills (Gallagher, 2004, p. 129) leads to male students scoring higher than female students. SMT majors likely have taken more science and math courses than non-SMT majors, and therefore score higher than non-SMT majors. Seniors are more prepared than freshmen, whether through completion of more science and math courses or more coursework in general, so they score higher.

The results also show that students who eventually pass the course score higher than students who fail the course. The passing students had more incoming knowledge, more natural ability, or some combination of both that enabled greater success on the diagnostic test and in the course. The diagnostic test therefore had some ability to predict which students will pass the course and could be a potential advisement tool, similar to the CCDT, with significant development efforts.

Reliability - As a whole and within subgroups, students' diagnostic test averages were statistically the same on the common questions used in 2002 and 2003. The questions were apparently reliable from year to year, supporting the reliability of the diagnostic test from year to year, and this result makes sense because the questions were from past New York State Regents Earth Science Exams that were available on line. Presumably, these exams were well-developed and well-tested for many years, and thus, they successfully served as the basis for the diagnostic test in this study. The 2002 diagnostic test had a KR-20 reliability of 0.69, as did the 2003 test, and the combined data also yielded a KR-20 reliability of 0.69. Instruments with KR-20 reliabilities greater than 0.70 are generally considered to possess statistical reliability, therefore these diagnostic tests are on the borderline. Because of the wide variety of topics and levels of Bloom's Taxonomy covered by the questions selected, it is not an unexpected result that the KR-20 reliability is on the borderline. Still, the New York Regents exams may serve as a possible prototype for a national standardized exam, at least for introductory physical geology courses like the one used for this study. However, the recently developed questions in the Geoscience Concept Inventory (Libarkin and Anderson, 2005) potentially represent a better starting point for such national testing since they provide a broader selection of questions, suitable for a range of introductory geology courses.

Strengths and Weaknesses - The analysis of specific student strengths and weaknesses showed only a few areas that students found to be extremely easy or truly difficult. Students were able to convert from scientific notation into standard notation and were familiar with a definition of evolution. Students most likely found these questions to be easy because the questions were at the simple recall level of Bloom's Taxonomy (Table 6) and the topics were most likely thoroughly addressed at the high school level. Students generally had the common misconception that oxygen is the most abundant gas in the Earth's atmosphere, and although the test question was also at the simple recall level, this misconception is pervasive and resistant to change. Female students had more difficulty with the relationship between water displacement and an object's mass and volume, and probably reflected the general gender difference in spatial and math ability, rather than any greater difficulty due to the question level. Female students and SMT majors had difficulty with remembering the difference between latitude and longitude, and both groups also had difficulty with factors affecting surface runoff. Both of these questions are at the lower level of Bloom's Taxonomy, but most likely students simply confused the terms latitude and longitude and/or had not encountered these topics. The possibility of not being exposed to the topics, particularly in the case of SMT majors, is high because of the lack of a high school Earth science requirement. Although this initial evidence is limited, the results indicate the potential of using a standardized test to aid in curriculum planning and curriculum assessment, including testing for student competency with respect to national Earth science learning standards.

Examining SMT majors by gender uncovered few new findings, except that female SMT majors had much different areas of strength than other students as a whole and as subgroups. This subgroup had strengths in estimating the meter unit to the nearest English unit, an area of strength identified for male students, and in identifying an element from a list of substances. With so few questions identified as a strength, it may be better in the future to apply results from a test of SMT ability, rather than a student's stated major, to better understand how gender and SMT ability could interact to impact diagnostic test performance.

CONCLUSIONS

Our initial attempt to develop and implement a diagnostic test for an introductory geology course produced a test that matches known and expected trends in student performance on standardized tests with respect to gender and background preparation. The test has some ability to predict student success in passing the course, appears to be reliable from year-to-year, and uncovers specific student strengths and weaknesses. This study provides evidence for the feasibility of developing one or more national standardized tests in Earth sciences.

Such tests would be valuable for Earth science education for many reasons. A diagnostic test could be used for student advising, particularly if geoscience educators produce a test with the greater predictive power of the CCDT as shown by Legg et al. (2001). This power should be achievable through optimization of the test and restricting questions to a single subject area, such as geology, instead of including a broad spectrum of science questions. The test results can be used to advise students about their chances of succeeding in a geology course and suggesting that the student should possibly seek additional preparation and/or tutoring.

A diagnostic test could be used in curriculum planning by providing instructors with data that is specific to the instructor and course. Such information may guide instructors toward addressing different learning styles and backgrounds of different subgroups of students. The same or a different test could be used for curriculum assessment, especially if used as part of a pre-post testing approach. Instructors can use the tests as a method for monitoring teaching effectiveness and student success in the specific course. Lastly, instructors can use the tests to make national comparisons regarding their students' incoming preparation and outgoing level of achievement, which can then also be matched to national Earth science learning standards.

Taken together, these outcomes would lead to significant enhancement of geoscience education efforts at the national level. The tests can facilitate the efforts of instructors concerned with what they are teaching, how they are teaching, and how their students are learning. Based on these potential benefits and our initial encouraging results, state and national efforts to create and implement diagnostic testing for Earth science students should be initiated and pursued.

ACKNOWLEDGMENTS

We thank Robert Ridky for sharing with us his thoughts on national Earth science standards and his experience with the Earth science curriculum exams, and Tom Greenbowe for the GALT exams. We also thank Jim Shea for reviewing an earlier version of this manuscript, and Julie Libarkin and Heather Petrovic for their useful comments. The authors gratefully acknowledge the help and dedication of the late Amy Rodgers, who assisted us with data entry and compilation. Partial funding for this study was provided by two National Science Foundation Awards (DUE-0228491 and 0231246).

REFERENCES

Bloom, B.S., and Krathwohl, D.R., 1956, Taxonomy of Educational Objectives, Handbook I: Cognitive Domain, New York, Longman, 205 p.

Drummond, C., 2003, Editorial, Journal of Geoscience Education, v. 51, p.1.

Gallagher, A.M., 2004, Gender differences in mathematics, Cambridge University Press, 368 p.

Krishnan, S.R. and Howe, A.C., 1994, The mole concept: developing an instrument to assess conceptual understanding, Journal of Chemical Education, v. 71, p. 653-655.

Legg, M.J., Legg, J.C. and Greenbowe, T.J., 2001, Analysis of success in general chemistry based on diagnostic testing using logistic regression, Journal of Chemical Education, v. 78, p. 1117-1121.

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

Libarkin, J.C., and Anderson, S.W., 2006, The Geoscience Concept Inventory: Application of Rasch analysis to concept inventory development in higher education; Rasch Applications in Science Education, ed. X. Liu, JAM Publishers, p. 45-73.

McDonnell, D.A., Steer, D.N., Owens, K.D., Knott, J.R., Van Horn, S., Borowski, W., Dick, J., Foos, A., Malone, M., McGrew, H., Greer, L., and Heaney, P.J., 2006, Using ConcepTests to assess and improve student conceptual understanding in introductory geoscience courses, Journal of Geoscience Education, 54, p. 61-68.

McFate, C. and Olmsted, J., 1999, Assessing student preparation through placement test, Journal of Chemical Education, v. 76, p. 562-565.

National Research Council, 1996, National Science Education Standards, National Academy Press, Washington, D.C.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W., 1996, Applied Linear Regression Models, The McGraw-Hill Companies, USA.

Olson, L., 2006, An alternative approach to gauging readiness: Coalition of small schools in N.Y. uses performance assessments, Education Week, 25, p. 28.

Roadrangka, V., Yeany, R.H., and Padilla, M.J., 1983. The construction and validation of Group Assessment of Logical Thinking (GALT), Annual meeting of the National Association for Research in Science Teaching, Dallas, TX.

Russell, A.A., 1994, A rationally designed general chemistry diagnostic test, Journal of Chemical Education, v. 71, p. 314-317.

Steinberg, R.N. and Sabella, M.S., 1997, Performance on multiple-choice diagnostics and complementary exam problems, Force concept inventory, The Physics Teacher, v. 35, p. 150-155.