Sample WINSTEPS Code

```
&INST                                  ; shows this is a control file (optional)
TITLE='Temperature and Heat Post-test'  ; Report title
NI=23                                   ; Number of items
NAME1=1                                ; First column of person label in data file
ITEM1=10                               ; First column of responses in data file
CODES=abcde                            ; How the answers were coded in the assessment
KEY1 = aeabcbedebcebdbaddbedcb          ; Answer key
PERSON=PERSON                          ; What a "person" was called
ITEM=ITEM                              ; What an "item" was called
&END
1                                      ; name of the items
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
END NAMES                           ; END NAMES or END LABELS must come at end of list
21015700 aeabcbedacddbdbaedbedce; person label with that persons responses
21010149 aeabcbedaccabdeaedbedcb
22054640 addbaabceddccbbacdccbab
22057562 aeaecaedebdebdbaedbedab
22060206 aebacaeaecdebbabecacbab
22064732 aebbcaedecddbdabedcedeb
High-performing persons
Difficult items
Low-performing persons
Easy items
```

Chapter 3

The Geoscience Concept Inventory: Application of Rasch Analysis to Concept Inventory Development in Higher Education

Julie C. Libarkin
Ohio University

Steven W. Anderson
Black Hills State University

The purpose of this study was the development of a concept inventory for use in entry-level college geoscience courses nationwide. Techniques gleaned from previous work in concept inventory development, scale development theory, grounded theory, and item response theory were used in validating 69 items on the Geoscience Concept Inventory (GCI). This chapter details the application of Rasch and differential item functioning (DIF) approaches to investigation of item stability and item bias. In addition, we establish a methodology for development of GCI subtests, creating an innovative solution to the problem of assessing the diversity of content covered in entry-level geoscience courses. The care taken in

developing and validating the GCI has resulted in a flexible instrument that provides the only standardized method for evaluating conceptual change in the geosciences.

Research Problem/Objectives

Student conceptual understanding and the impacts of instruction on student ideas is a burgeoning field of research (e.g., Oliva, 2003). At the college level, significant work has been done in mathematics (e.g., Adams, 1997), physics (e.g., Lewis and Linn, 1994), and chemistry (e.g., Basili and Sanford, 1991), with researchers in biology (e.g., Windschitl and Andre, 1998); (Anderson, Fisher, and Norman, 2002) and geology (e.g., Dodick and Orion, 2003; Libarkin, Anderson, Dahl, Beilfuss, Boone, and Kurdziel, 2005) rapidly catching up to other disciplines. Assessment of conceptual learning in the geosciences has traditionally focused on K-12 students, with studies of college students or other adults only recently emerging (DeLaughter, Stein, Stein, and Bain, 1998; Trend, 2000; Dahl, Anderson, and Libarkin, 2005; Libarkin et al., 2005). Qualitative studies are concentrated outside of the U.S. (e.g., Happs, 1984; Marques and Thompson, 1997; Dodick and Orion, 2003; Trend, 2000), 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, Steer, and Owens, 2003). Quantitative study of student conceptual understanding in the geosciences (Libarkin and Anderson, 2005; Lambert, 2005) lags far behind other disciplines.

While the majority of work in conceptual change has utilized qualitative methods to ascertain student ideas, mixed methods approaches and quantitative studies are becoming increasingly common (e.g., Gobert, 2000; Libarkin and Anderson, 2005). Assessment test development for the college student population is also receiving more attention, with significant work emerging across science, technology, engineering, and mathematics (STEM) disciplines. The development of the Force Concept Inventory (FCI; Hestenes, Wells, and Swackhamer, 1992) in the early 1990s dramatically changed the way physicists viewed teaching and learning in college level physics courses, and marked a new beginning for concept inventory development and use in higher education. 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 of 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 et al., 2002), physics (Yeo and Zadnik, 2001), astronomy (Zeilik, Schau, and Mattern, 1999; Lindell and Olsen, 2002), and the geosciences (Libarkin and Anderson, 2005; Table 1) and may significantly impact the way faculty view classrooms as loci for research. Other initiatives to create concept inventories on a variety of topics, such as the Foundation Coalition's drive to create concept inventories in engineering (e.g., Rhoads and Roedel, 1999), are ongoing.

We set out to design an assessment instrument that would be a valid tool for use with all entry-level students nationwide, and which could be applied to a wide range of courses covering a variety of topics relevant to the Earth Sciences. The resulting instrument, the Geoscience Concept Inventory (GCI), is unique in higher education in the use of Rasch approaches in its development, validation and use. The ability to statistically place all GCI items on a single Rasch scale means that faculty and researchers are able to create sub-tests that are tied to a single scale, and which will therefore have comparable scores. This then allows faculty nationwide freedom of test design relative to their course content, without sacrificing the ability to compare student learning and instructional approaches used in different courses. Ongoing development of the GCI builds upon existing studies and incorporates additional methodologies for development and validation, blending three theoretical bases, scale development theory, grounded theory, and item response theory, and utilizes a diverse population of students and institutions during piloting (Table 1). We report here on results from testing of 76 inventory questions, called items, that were developed, piloted, and evaluated in 2002-2004.

Theoretical Framework

Quantitative assessment instruments in higher education have been developed in a wide range of science disciplines, starting with physics. The publication of the FCI (Hestenes et al., 1992) heralded a new age in concept inventory development and use in higher education STEM disciplines. The importance of this instrument in physics education is clearly well-recognized by the community (e.g., Wieman and Perkins, 2005), primarily for the vast number of courses that have been assessed by this instrument and the FCI's power as a method for identifying "effective" instruction. A number of additional conceptual assessment instruments in other STEM disciplines have been developed in the past two decades. Faculty and researchers alike have used these conceptual inventories to ascertain the effectiveness of new instructional interventions, to evaluate student learning, to compare innovative with traditional pedagogies, and to test the transferability of curriculum from one discipline to another. "Learning" has been evaluated through comparison of pre- vs. post-test averages, use of raw or normalized gain scores, comparison of change in individual scores, implementation of model analysis theory, or comparison of scores scaled via Rasch analysis. Many concept inventories have been placed online through FLAG (Field-Tested Learning Assessment Guide; www.flaguide.org) or through independent websites (e.g., Geoscience Concept Inventory or Lunar Phases Concept Inventory).

Existing concept inventories in higher education science utilize a number of common approaches, including (Table 1): 1) Using a pre-determined content focus, usually through expert opinion or a review of texts; 2) Designing alternative responses (distracters) to multiple-choice items based upon developer experiences in the classroom, a review of existing literature, open-ended questionnaires, and/or interviews with students; and 3) Choosing participating institutions (existing studies had N = 1-5 institutions) according to type (e.g., large state schools) or similar geographical area (e.g., schools in Arizona). Studies differ on the number of students tested during initial piloting and on the type of statistical analyses performed. For studies discussed here, the number of college students tested during piloting ranged from 50 to 750 students and measurement of reliability, factor analysis, and model analysis were performed in some studies but not others. A wide range of individual item difficulties is reported on each of these instruments.

Table 1

Comparison of existing concept inventories and the Geoscience Concept Inventory (GCI)

Typical Concept Inventory Development*	Development of the GCI	Comments on GCI Approach
Predetermined content	Test content is based upon ideas presented by students	Questions are grounded in data gathered from college students
Alternative choices based on existing studies, questionnaires, and/or interviews (N < 30)	Data from ~20 existing studies 1000+ questionnaires 75+ interviews 10 institutions	Analysis (coding) of qualitative data allowed development of authentic "incorrect" choices
50-750 college students tested during piloting	Fall 2002: N = 2219 pre-tests Fall 2003: N = 1376 pre-tests	For N >~300, statistical sampling of sub-populations is usually possible
Institutions of similar type or locality (N = 1-8)	Colleges: 5 community or tribal, 44 public or private, 60 courses, 8-250 students per course.	The GCI should be generalizeable to all populations of students.
Statistical analyses either not performed, or reliability scores only ‡	Item Response Theory (Rasch) analysis performed. Some items removed due to statistical bias.	Raw scores can be re-scaled relative to test difficulty, providing a more accurate measure of changes.

Notes: * Blend of development strategies utilized by Hestenes et al., 1992, Zeilik et al., 1999, Yeo and Zadnik, 2001, Anderson et al., 2002, and Lindell and Olsen, 2002. ‡Anderson et al. (2002) perform a factor analysis to ascertain internal validity.

The significant research into concept test development in some areas of higher education, STEM, provides an excellent basis for test development in other disciplines (Table 1). As many researchers have discovered, development of high quality tools for assessing student learning is needed to accurately reveal links between learning and teaching. Recent advances in the field suggest that the incorporation of cognitive models and model development into assessment instrument development holds the greatest promise for developing tools that will yield useful results. The Cognition-Observation-Interpretation assessment triangle is one validated approach to assessment that considers the link between thinking, measurement, and statistical analysis (National Research Council, 2001). In this triangle, Cognition refers to a model for student knowledge or skill development, Observation refers to a situation through which student performance can be observed, and Interpretation refers to a method for scoring observations and drawing conclusions, such as a qualitative rubric or a statistical model. In reference to the GCI one could consider a) conceptual change in the geosciences as the model being evaluated; b) the GCI itself as the task that elicits thought; and c) Rasch analysis and DIF as the interpretive agent.

Cognition

A variety of models exist that relate to how people understand the world around them, and how understanding can change as a result of experience. The National Research Council (NRC, 2001) lays out four "perspectives" of how we understand knowledge and learning, some of which align perfectly with conceptual change models that were developed prior to the NRC report. Of most importance to our discussion of cognition with relation to the GCI are the Differential and Cognitive Perspectives (National Research Council, 2001), and the related conceptual change models of Posner, Strike, Hewson, and Gertzog (1981) and Vosniadou and Brewer (1992). Differential approaches to evaluating cognition drive most assessment, including testing in the typical college classroom. This approach is an attempt to characterize individual performance, often in relation to other people, and usually after a learning experience. Conceptual change as viewed by Posner et al. (1981) lays out a foundation for effectively engendering conceptual change, and, if assessment aligns with learning, improving individual performance. Cognitive approaches to evaluating cognition, on the other hand, focus on the connections between old and new knowledge, and the methods by which people construct knowledge. From a conceptual change perspective, this is similar to the model of Vosniadou and Brewer (1992) wherein prior knowledge and new knowledge are synthesized together as people build new mental models to explain phenomena. A theory of cognition specific to the domain of geosciences is not well-established, although recent work on conceptual understanding suggests that Earth-related concepts develop from both observations of the world and instructional experiences (Libarkin et al., 2005). Research in conceptual change from naïve models toward scientific models in Earth Science suggests a blending of models as learning occurs, rather than wholesale shift from pre-existing to new ideas.

Observation

In creating the GCI, we sought to develop an instrument that would allow us to document student conceptual understanding of fundamental ideas in the geosciences. We were most interested in documenting student performance (the Differential Perspective, above), and comparing student performance before and after instruction, and across different instructional interventions. The Geoscience Concept Inventory is the tool that we used to observe student performance in this study, and to document mental models of Earth processes held by students at different instructional stages. The method by which the GCI was created helps illuminate its usefulness as an observational tool.

We envision two primary uses for concept inventories: As diagnostics of conceptual understanding and as assessments of student learning. We believe that valid and reliable assessment tools are most applicable to a wide variety of college students when test questions and answers are developed through careful qualitative evaluation of student conceptions, following pioneering work in other fields; and careful attention is paid to both scale development and grounded theory.

Scale Development Theory. Measurement of psychological phenomena, such as learning, is a well-established field of research (psychometrics) and has specific procedures for ensuring validity and reliability. A variety of forms of validity were incorporated into the development of the GCI, including construct validity, content or face validity, criterion validity, external validity, and internal validity. Development of the GCI involved an iterative process of qualitative data collection (questionnaires and interviews), question development, review by educators and geoscientists, pilot testing, statistical analysis, revision, testing, and further collection of Think-Aloud Interviews (e.g., Zeilik et al., 1999) to ascertain the reasons behind student responses (Figure 1). Each step in the process was designed to achieve as a high level of reliability and validity as possible. This paper focuses on the statistical analysis of GCI data, particularly Rasch and DIF approaches. Further discussion of the methods used in developing the GCI, including a full development of validity and reliability measures, will be expanded in a future publication.

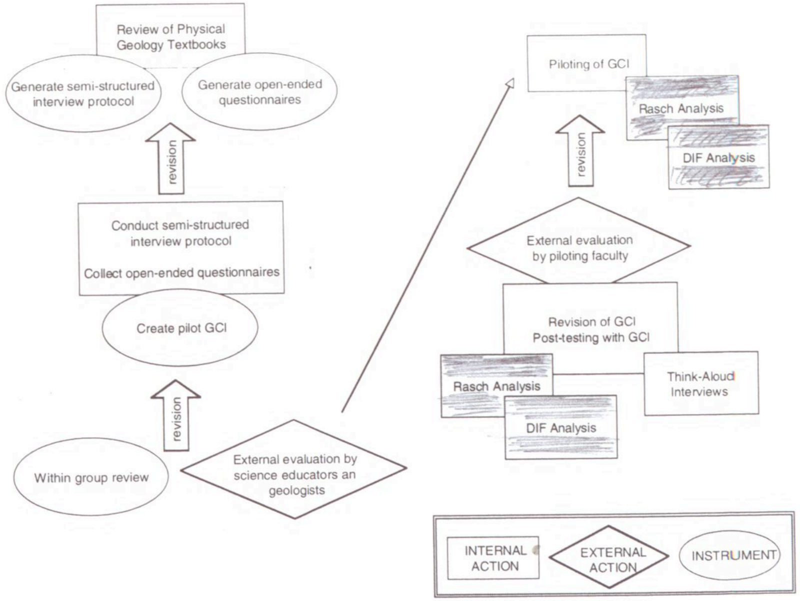

Review of Physical Geology Textbooks
Generate semi-structured interview protocol
Generate open-ended questionnaires
revision
Conduct semi-structured interview protocol
Collect open-ended questionnaires
Create pilot GCI
revision
Within group review
External evaluation by science educators an geologists
Piloting of GCI
Rasch Analysis
DIF Analysis
revision
External evaluation by piloting faculty
Revision of GCI Post-testing with GCI
Rasch Analysis
DIF Analysis
Think-Aloud Interviews
INTERNAL ACTION
EXTERNAL ACTION
INSTRUMENT

Figure 1. Schematic of steps taken in developing the Geoscience Concept Inventory (GCI). Internal actions are those taken by the research team and external actions are those taken by study participants or experts during validation. Instrument refers to the GCI. Tasks in gray indicating those items, specifically Rasch and DIF analysis, discussed in this paper.

Grounded Theory. Grounded theory rests on the idea that the perceptions and perspective of the study population is fundamental to interpreting research data, conclusions will only be meaningful if the study population is considered, and that researchers must be careful to avoid imposing their own perspectives onto research findings. While grounded theory is most often applied to qualitative studies, we argue that a fundamental limitation in the development of most concept tests is the lack of attention paid to grounded theory during the initial phases of test development. The goal of grounded theory research is to construct theories in order to understand phenomena, rather than testing predetermined theories through data collection and analysis. Although it has been developed and principally used within the field of sociology, grounded theory has been used successfully in a variety of different disciplines. These include education, nursing, politics, and psychology. Glaser and Strauss (1967) do not regard the procedures of grounded theory as discipline specific, and many researchers have applied the methodology within their own disciplines (Haig, 1996). Our use of grounded theory specifically related to the mining of interview data for conceptual questions and responses, rather than driving the research design itself.

Grounded theory hinges on the notion that studies should be embedded in the perceptions and reality of the study population. This ensures that conclusions based upon research will be meaningful, and that the researcher does not superimpose their own perspective on research findings. While grounded theory is most often applied to stand-alone qualitative studies, we argue that a fundamental limitation in the development of most concept tests is the lack of attention paid to grounded theory during the initial phases of test development. Although use of qualitative data ensures that test items will be grounded, test items are generally pre-designed. That is, a review of existing tests, scientific theory, or expert opinion is used to predetermine the content of test questions, and these questions are generally used verbatim in interviews or open-ended questions used in qualitative phases of test development. In addition, because test item composition is predetermined, questions often contain technical language that may limit the researcher's ability to uncover alternative conceptions. In fact, many existing tests assume understanding of fundamental concepts in assessing specific target information (Table 1). We hoped to embed our research, i.e., the development of the GCI, in the experiences and perspectives of the population being studied. As a consequence, the GCI test items were only partially based upon predetermined content, and were primarily developed from coding of questionnaire and interview data. Analysis of emerging themes in our interview data led to the development of a specific set of concepts important for the entry-level geosciences. Interview data were subsequently reanalyzed to identify student-generated distracters.

Interpretation

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 latent 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 % 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, discrimination 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.

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. We applied a specific IRT approach, Rasch analysis, to test development and interpretation. Equivalent changes in raw score for multiple students may not translate to equivalent changes in conceptual understanding. Using a statistically calculated Rasch scale allows the determination of test scores that more accurately reflect "understanding". Rasch and related approaches can also be used to study differential item functioning (do sub-populations, women, for instance, perform differently

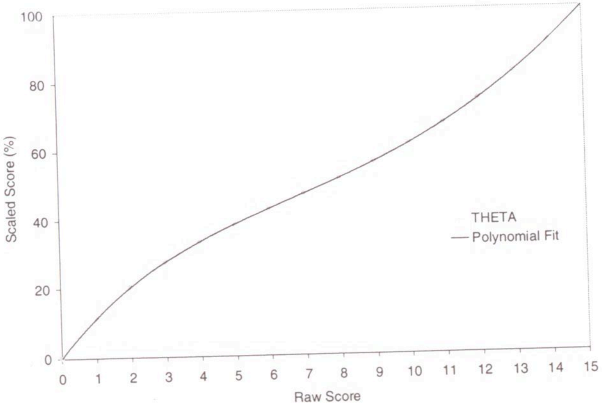


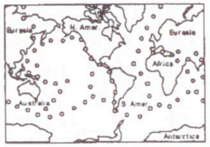
Figure 4. Comparison of raw GCI scores and scores scaled using one-dimensional Rasch modeling. Gray dots are raw score-theta conversions calculated from logit values generated by Quest (Adams and Khoo, 1996). Theta values have been converted to scaled scores on a 0-100% scale. The solid line is the polynomial fit as in eq. (1).

tual difference between 5 and 10. The transformation to scaled scores reflects this conceptual difference.

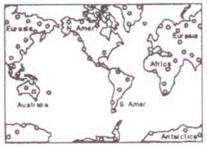
Overall, students in the test population found the test difficult. For example, during 2002 testing students had similar but statistically different pre- and post-test means (pre = 42.2±12; post = 45.8±13; $n = 1295$ student). Analysis of sub-populations indicates that students with low pre-tests (<40%, $n = 388$) dominate this effect (pre = 32±7, post = 41±10), with extreme significance on a t-test ($t_{stat} = 1.96 < t_{crit} = 15$) and an effect size of 0.46 (Libarkin and Anderson, 2005). The distribution of test items along the Rasch scale is itself interesting, and provides some insight into the conceptual understanding of entry-level geoscience students. In particular, the presence of rote knowledge does not necessarily imply conceptual understanding. In this study, 82% of students ($N = 2192$) correctly indicated that the Earth is ~4 billion years old (Figure 5-b) prior to instruction. However, after instruction ($N = 1907$) only 18% of students were able to correctly indicate that absolute age dating is the primary technique employed to calculate the Earth's age. In addition, only 5% knew that U-Pb is the only method among those listed that can be used when dating the Earth (Figure 5-c). Rather, the majority of students believed that the comparison of fossils or rock layers or the use of carbon were primary absolute age dating techniques. These data indicate that while students are familiar with geologic topics prior to instruction, in this case the Earth's age, few understand underlying concepts vital to complete conceptual understanding, even after instruction.

a. The following maps show the position of the Earth's continents and oceans. The o's on each map mark the locations where earthquakes occur most frequently. Which map do you think best represents where earthquakes occur most frequently on Earth?

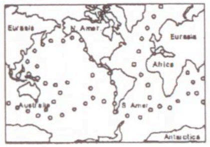
Circle one: A B C D E



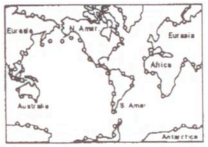
A. In continental and oceanic crust, and along continental margins



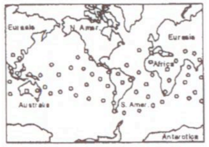
B. Mostly in continental crust



C. Mostly in oceanic crust



D. Mostly along continental margins



E. Mostly in warm climates

b. If you could travel back in time to when the Earth first formed as a planet, how many years back in time would you have to travel?

(A) 4 hundred years
(B) 4 hundred-thousand years
(C) 4 million years
(D) 4 billion years
(E) 4 trillion years

c. Some scientists claim that they can determine when the Earth first formed as a planet. Which technique(s) do scientists use today to determine when the Earth first formed? Choose all that apply.

(A) Comparison of fossils found in rocks
(B) Comparison of different layers of rock
(C) Analysis of uranium and lead in rock
(D) Analysis of carbon in rock
(E) Scientists cannot calculate the age of the Earth

Figure 5. Sample questions from the GCI. Bias and assertions about difficulty are based upon IRT analyses performed using Quest. a) Question removed because of bias. b) Third easiest of the 69 validated questions. C) Most difficult of the 69 validated questions. Notice that the easier question deals with knowledge of the Earth's age, while the most difficult question asks how that knowledge was established.

Anderson (2005) report an average pre-test score of 41.5% for 2493 students and 43 courses tested nationwide, similar to GeoJourney average. GeoJourney gains far outstrip the 4% gain observed in 29 post-testing courses, however. Preliminary results suggest that students are making exceptional gains, from scaled pre-course scores of 43% to post-instruction scores of 59% (Lyle-Elkins and Elkins, 2004; Joe Elkins, personal communication). The potential of the GCI is its power as both a diagnostic and an assessment instrument in individual courses and as a mechanism for comparing learning across the wide variety of entry-level geoscience courses. With such an assessment instrument, we have a mechanism for evaluating learning, discriminating between effective and less effective teaching approaches, and ultimately transferring teaching deemed effective, with this label applied based upon data as opposed to anecdotal evidence, into a variety of instructional and educational settings. Ultimately, as has happened in other disciplines (see Kurdziel and Libarkin, 2001, for a discussion), the availability and use of a valid and reliable assessment instrument can produce a real shift in the ways that faculty approach teaching.

Design and Methodology

Participants

Data were collected from a wide range of institutions scattered across the U.S. as an approach to ensuring external validity, the generalizability of the GCI to entry-level college students nationwide. GCI test items were completed at over 42 institutions nationwide, with 3595 students participating early in the academic year (usually on the first day of class and no later than two weeks into the year). We also have post-test data from ~1750 students, collected during the last week of class or during the final examination. In all, 59 courses in Physical Geology, Environmental Science, Oceanography, and Historical Geology, and with class sizes ranging from 8-210 were included in the study. These courses stemmed from 42 different institutions in 23 states across the country. Of these institutions, 6 were community colleges and one was a tribal college, and 35 were four-year institutions of which 8 were privately funded. Participants were almost equally split between men and women, and about 20% of the students were non-Caucasians (Table 2). Specific information about some of the courses and students participating in this study are detailed in Libarkin and Anderson (2005).

Administration and Data Analysis

Items were created and administered at different times and on tests of different length, and linking items were used to place item difficulty estimates on the same scale. During 2002, 30 questions were piloted on two tests of twenty questions, with each containing eleven common items. During 2003, two tests, one with 29 and another with 30 questions, were piloted. Each of these tests contained six common items drawn from the pool of the original common eleven. Two of the items included in GCI piloting were pilot questions for a study in phenomenography and were removed from further analysis. A third item relating to the location of cloud formation was removed because of faculty concerns about the accuracy and interpretation of the item. Data sets were merged by treating students who did not answer certain items as missing. The resulting file contained 73 items and responses from 3595 students. Individual questions were answered by as many as 3595 students and as few as 306 students. Calibration of item difficulty estimates was performed using Quest (Adams and Khoo, 1996) software for Rasch modeling, utilizing the one-parameter logistic model for Rasch analysis, and Mantel-Haenszel approximation of differential item functioning (DIF).

Rasch Analysis

Quest (Adams and Khoo, 1996) utilizes a Rasch model that is easily applied to dichotomous data sets. We limited the current analysis to a standard one-parameter logistic model most common when utilizing Rasch approaches. This approach simply applies a "right-wrong" to each test item, where all wrong answers are equivalently scored regardless of specific distracters chosen. The one-parameter logistic model takes the form (e.g., Embretson and Reise, 2000):

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}, i = 1, 2, \ldots n, \qquad (1)$$

where $P_i(\theta)$ is the probability that a random examinee with ability, θ, answers item *i* correctly. $\theta - b_i$ is the difference between ability and item difficulty. b_i is the difficulty parameter, or threshold, for item *i*. (Please note that the variables used to represent person ability and item difficulty are different than those used in many of the other chapters in this book.) This threshold is the point where the probability of a correct response to the specific item *i* is 50%. Eq. (1) implies an s-shaped curve between 0 and 1 over the ability scale. The one-parameter logistic model assumes unidimensionality in the data, suggesting that items are measuring a single latent trait. This unidimensionality can be evaluated through factor analysis. The presence of a single dominant factor, even given the existence of less important sub-factors, satisfies the condition of unidimensionality (Embretson and Reise, 2000).

Quest initially estimates approximate item difficulties and then uses these values to calculate person abilities. The difference between an individual's actual score on an item and the expected response is called the response residual. The iterative process of approximating item difficulties is repeated until the item and person residuals are reduced to lowest possible values. Quest repeats this iteration until the change in residuals that occurs as the result of a single iteration is 0.005 logits. Although residuals can never be brought to zero with the one-parameter logistic model, they can be reduced to minimum values. Application of higher order logistic models beyond the simple one-parameter Rasch model may further reduce residual values and are worth considering for more complex analyses.

Differential Item Functioning (DIF)

Bias in testing is an important consideration when developing instruments and interpreting results. Tests should be fair to all takers, and should perform in similar ways for people of diverse backgrounds. Bias generally exists because a test item is measuring more than the specific latent trait under consideration, and the introduction of additional sources of difficulty can affect the performance of specific groups. The use of differential item functioning, or DIF, to evaluate differences in item performance relative to demographic variables such as race or gender is becoming a standard in psychometric test evaluation (Zumbo, 1999; Bond and Fox, 2001). In general, individuals at the same level of understanding should perform similarly on unbiased items regardless of other variables (e.g., age, gender, race). DIF compares test-takers with similar abilities and different demographics, looking for dramatic differences in difficulty that are unrelated to the overall test performance of an individual.

A variety of DIF-detection approaches exist, some of which are specifically based on latent-variable analysis, and are therefore technically IRT-based, and others that use slightly different, non-IRT approaches. In our case, we rely on the Mantel-Haenszel statistic to approximate DIF. This approach was chosen because it is an extremely popular and common approach used in the research literature (e.g., Clauser and Mazor, 1998), and is also a routine calculation easily performed using Quest software.

Evidence of DIF does not necessarily mean that test items need to be removed. In some studies, evidence of DIF is exactly what researchers are interested in, and variation in item performance across groups can indicate powerful differences in understanding that are meaningful in their own right. DIF differences may reflect either 1) actual differences between groups; or 2) differences related to measurement of variables other than the latent trait, often labeled "bias". In the case of the GCI, it is quite difficult to differentiate between these two possibilities. Our poor understanding of cognitive processes and conceptual change in the domain of the geosciences limits our ability to meaningfully understand DIF results. However, we can consider the DIF analysis from the perspective of meaningful testing. We are most interested in providing an assessment tool that will be valid and reliable for all populations. We are faced with a quandary: Is it best to leave DIF items on the GCI, as would be done for items that are simply measuring true differences, or is it best to remove DIF items, potentially lowering the ability of the instrument to differentiate between populations yet removing any possibility of bias? In the end, we chose a policy of "better-safe-than-sorry" and removed items that showed DIF on both the pre- and post-test administrations, ensuring minimal bias. These items were not removed completely from the test bank, but were rather labeled as DIF items, and the possibility for revisiting their inclusion on the GCI once our understanding of geoscience cognition improves still remains.

For the GCI, we were interested in developing an instrument that would be applicable to the widest range of students possible. In keeping with this purpose, we used DIF to remove items that performed extremely differently for different identifiable groups. Two different DIF analyses, one for groups subdivided by gender and another for groups based on ethnicity/race (Caucasian vs. others) were performed. Standardized differences that fell outside of ±2 were considered significant DIF, and DIF was evaluated for both pre- and post-test data. DIF that remained significant for both test administrations was stable enough to warrant classification of test items as biased.

Findings

This study is focused on the Rasch analysis of student test data collected during creation and validation of the Geoscience Concept Inventory (GCI). Factor analysis of these data indicates one dominant factor, satisfying the condition of unidimensionality required for use of the Rasch model. In conducting the Rasch analysis we are most interested in item estimates, as opposed to results from individual students. The item separation reliability estimated by QUEST was 0.99, indicating that most of the observed estimate variance is considered true (Wright and Masters, 1982). The items on the GCI were therefore spread out enough along the Rasch scale to allow for consistent modeling of item responses.

Estimates of item fit to the model suggest that no items are mis-fitting enough to warrant removal from the test (Table 3). INFT MNSQ and OUTFT MNSQ are infit and outfit mean squares, respectively. INFT *t* and OUTFT *t* are the normalized versions of the mean square statistics. Both the INFT MNSQ and OUTFT MNSQ have expected values of one for data that perfectly conform to the model, and values of 1±x indicate that the data set is x% away from the model. As a rule of thumb for multiple-choice tests, INFT MNSQ and OUTFT MNSQ should be between 0.8 and 1.2 (Bond and Fox, 2001). INFT *t* and OUTFT *t* are absolute values of *z*-scores and are expected to be zero for a perfect data-model fit, and within ±2 assuming a 95% confidence interval.

Applying the criteria above to our data indicates a good match between the Rasch model and the test data. The values for INFT MNSQ range from 0.8-1.17, with the majority of the items very close to the expected value of one. The highest departure of 17% overfit applied to an item that was removed from the test for anomalous DIF, and the 12% underfit for question 65 is well within the acceptable range for mean square values. The INFT *t* and OUTFT *t* indicate that a few items may not have good model-data fit as their absolute t values are greater than 2. However, none of the items showed consistent misfit on all four statistics, suggesting that items have an overall good fit to the model (e.g., Liu and McKeough, 2005). Future interpretations based upon specific questions, particularly those few with both significant (>4) INFT *t* and OUTFT *t* misfit, should incorporate more detailed analysis of item fit to the one-parameter logistic model.

A significant difficulty gap of 0.78, calculated as the difference between item measures, exists between the first and second hardest items on the GCI, and a gap of 0.73 exists between the 2nd and 3rd hardest questions. In contrast, the easiest and second easiest questions have a difficulty gap of only 0.23. This indicates that the harder end of the Rasch scale is sparsely populated by items, and future additions to the scale should concentrate on this end, if possible. Analysis of the relationship between item and case

Table 3

Rasch modeling fit statistics, difficulty estimates, standard errors, and % correct based on pre-test data (N = 3595).

Item	INFIT MNSQ	OUTFIT MNSQ	INFT *t*	OUTFT *t*	Measure	*S.E.*	% CORRECT
1	0.97	1.22	–0.2	1.0	3.73	0.14	2.7
2	1.01	0.92	0.1	–0.1	2.95	0.33	2.7
3	1.11	1.84	1.0	3.9	2.22	0.13	6.2
4	0.95	1.02	–1.0	0.2	1.86	0.07	10.5
5	**1.08**	**1.49**	**0.5**	**1.9**	**1.88**	**0.2**	**7.3**
6	1.14	2.12	1.6	6.1	1.83	0.12	8.8
7	0.97	1.12	–0.2	0.6	1.59	0.18	9.5
8	0.94	0.92	–0.5	–0.4	1.35	0.17	11.6
9	1.02	1.01	0.6	0.3	1.29	0.06	20.3
10	0.97	0.97	–0.2	–0.2	1.09	0.15	14.4
11	**1.15**	**1.29**	**2.7**	**2.9**	**1.03**	**0.09**	**16.8**
12	1.09	1.29	1.7	2.8	1.07	0.09	16.3
13	1.12	1.21	3.9	3.7	1.13	0.05	22.6
14	1.06	1.15	1.7	2.2	0.83	0.07	26.9
15	1.01	1.03	0.2	0.5	0.98	0.07	24.5
16	1.03	1.08	0.9	1.2	0.88	0.07	27.0
17	1.15	1.29	3.8	3.9	0.96	0.08	25.5
18	0.93	0.88	–1.6	–1.6	0.68	0.08	21.9
19	0.94	0.90	–1.8	–1.6	0.71	0.07	28.9
20	0.97	0.95	–0.8	–0.8	0.81	0.07	27.4
21	0.96	0.96	–1.1	–0.7	0.66	0.07	29.8
22	1.06	1.02	0.8	0.3	0.51	0.13	22.4
23	0.92	0.95	–2.2	–0.9	0.54	0.07	23.4
24	1.01	1.09	0.4	2.0	0.63	0.05	30.8
25	1.04	1.05	1.1	0.8	0.56	0.07	32.5
26	0.91	0.9	–5.1	–3.2	0.43	0.04	31.0
27	1.03	1.11	0.9	1.6	0.43	0.08	25.8
28	0.94	0.95	–2.1	–0.9	0.38	0.07	36.0
29	1.01	1.00	0.2	0.1	0.33	0.13	25.5
30	1.01	1.01	0.2	0.1	0.32	0.08	27.7
31	0.93	0.95	–2.4	–0.9	0.17	0.07	30.5
32	0.97	0.98	–1.1	–0.4	0.01	0.07	33.4
33	1.1	1.17	2.0	2.0	–0.06	0.12	32.7
34	1.07	1.10	1.5	1.1	–0.05	0.12	32.5
35	0.94	0.92	–3.3	–2.4	–0.06	0.05	44.4
36	0.98	0.97	–0.9	–0.7	–0.06	0.07	45.1
37	1.13	1.14	4.9	2.8	–0.3	0.07	39.9
38	0.98	0.95	–0.7	–0.9	–0.18	0.07	37.3
39	0.9	0.87	–4.2	–2.9	–0.25	0.07	48.0
40	1.1	1.18	5.8	5.3	–0.17	0.05	46.9
41	1.01	1.03	0.6	0.6	–0.28	0.07	49.5
42	0.93	0.92	–5.3	–3.4	–0.45	0.04	48.7

(*Table 3 continued on next page*)

Table 3 (*continued from previous page*)

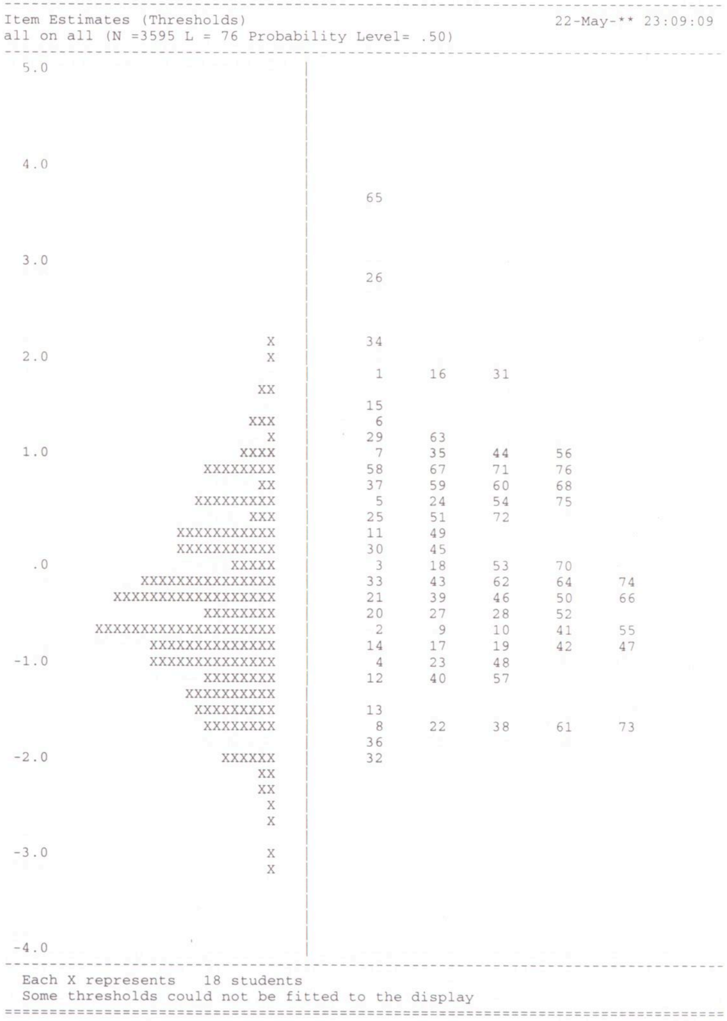
Rasch modeling fit statistics, difficulty estimates, standard errors, and % correct based on pre-test data (N = 3595).

Item	INFIT MNSQ	OUTFIT MNSQ	INFT *t*	OUTFT *t*	Measure	*S.E.*	% CORRECT
43	1.03	1.06	1.2	1.3	−0.35	0.07	40.8
44	1.06	1.06	2.4	1.2	−0.42	0.07	42.3
45	0.93	0.89	−2.7	−2.3	−0.35	0.07	41.0
46	0.98	0.97	−1.4	−1.0	−0.39	0.05	51.4
47	0.94	0.93	−1.5	−1.0	−0.51	0.11	42.3
48	0.94	0.92	−1.8	−1.1	−0.48	0.11	41.6
49	1.04	1.09	2.6	2.7	−0.56	0.05	54.6
50	1.05	1.06	1.6	0.9	−0.73	0.11	47.2
51	0.95	0.94	−1.6	−0.8	−0.66	0.11	45.5
52	0.95	0.95	−2.1	−1.1	−0.66	0.07	47.5
53	1.01	0.99	0.4	−0.2	−0.77	0.05	59.1
54	0.96	0.94	−1.4	−0.8	−0.89	0.11	51.0
55	0.98	0.97	−0.6	−0.4	−0.86	0.11	50.1
56	0.94	0.94	−2.0	−0.9	−0.82	0.11	49.3
57	1.1	1.11	4.0	2.3	−0.86	0.07	51.7
58	0.9	0.88	−4.4	−2.8	−0.96	0.07	54.0
59	0.91	0.89	−3.5	−2.1	−0.94	0.07	63.3
60	0.99	1.00	−0.2	0.1	−1.03	0.11	54.0
61	1.09	1.13	4.5	3.3	−1.07	0.06	56.0
62	0.95	0.93	−2.0	−1.6	−1.06	0.07	56.1
63	0.99	1.00	−0.3	0	−1.16	0.11	56.9
64	0.94	0.9	−2.2	−2.0	−1.22	0.07	59.4
65	1.04	1.04	1.5	0.8	−1.21	0.07	67.1
66	0.88	0.83	−3.0	−2.2	−1.5	0.11	64.2
67	1.01	1.04	0.2	0.5	−1.74	0.12	69.1
68	0.94	0.89	−2.7	−2.5	−1.75	0.05	73.4
69	0.92	0.86	−2.8	−2.5	−1.67	0.07	68.6
70	0.97	1.08	−0.8	1.0	−1.73	0.08	76.3
71	0.93	0.84	−1.7	−2.1	−1.76	0.08	77.8
72	1.06	1.11	1.8	1.8	−1.82	0.07	71.3
73	0.98	0.99	−0.5	−0.1	−2.05	0.08	75.4
P	1.02	1.02	0.2	0.2	−0.67	0.11	45.8
P	1.07	1.16	2.0	2.2	−0.57	0.11	43.6
C	1.17	1.19	4.8	2.7	1.26	0.16	12.4
Mean	1.00	1.05	−0.1	0.3	0.00		
SD	0.07	0.20	2.3	2.0	1.15		

Notes: Items in bold were subsequently removed for anomalous differential item functioning relative to gender and/or race/ethnicity. P are two items related to phenomenography and C is a questions related to cloud formation that was deemed confusing by reviewers; these questions were removed from further analysis. Specific items can be accessed on the GCI website, http://newton.bhsu.edu/eps/gci.html.

estimates also illustrates the extreme difficulty of the hardest two questions relative to the population of test-takers (Figure 2).

IRT analysis based on Rasch modeling (one parameter logistic model) of the 73 piloted test questions and DIF analysis provided useful information about the stability of questions and question bias. Overall, Rasch



```
--------------------------------------------------------------------------------
Item Estimates (Thresholds)                                   22-May-** 23:09:09
all on all (N =3595 L = 76 Probability Level= .50)
--------------------------------------------------------------------------------
  5.0                                  |
                                       |
                                       |
                                       |
                                       |
                                       |
  4.0                                  |
                                       |
                                       |     65
                                       |
                                       |
                                       |
  3.0                                  |
                                       |     26
                                       |
                                       |
                                       |
                                      X|     34
  2.0                                 X|
                                       |      1     16     31
                                     XX|
                                       |     15
                                    XXX|      6
                                      X|     29     63
  1.0                              XXXX|      7     35     44     56
                               XXXXXXXX|     58     67     71     76
                                     XX|     37     59     60     68
                              XXXXXXXXX|      5     24     54     75
                                    XXX|     25     51     72
                            XXXXXXXXXXX|     11     49
                            XXXXXXXXXXX|     30     45
   .0                             XXXXX|      3     18     53     70
                        XXXXXXXXXXXXXXX|     33     43     62     64     74
                     XXXXXXXXXXXXXXXXXX|     21     39     46     50     66
                               XXXXXXXX|     20     27     28     52
                   XXXXXXXXXXXXXXXXXXXX|      2      9     10     41     55
                         XXXXXXXXXXXXXX|     14     17     19     42     47
 -1.0                    XXXXXXXXXXXXXX|      4     23     48
                               XXXXXXXX|     12     40     57
                             XXXXXXXXXX|
                              XXXXXXXXX|     13
                               XXXXXXXX|      8     22     38     61     73
                                       |     36
 -2.0                            XXXXXX|     32
                                     XX|
                                     XX|
                                      X|
                                      X|
                                       |
 -3.0                                 X|
                                      X|
                                       |
                                       |
                                       |
                                       |
 -4.0                                  |
--------------------------------------------------------------------------------
 Each X represents   18 students
 Some thresholds could not be fitted to the display
================================================================================
```

Figure 2. Item and case estimates for GCI pre-test results (*N* = 3595). Each X represents 16 people.

results indicate that test items represent a range of conceptual understanding (Figure 3). This range is loosely correlated with difficulty, a traditional statistic often reported when discussing concept tests. Four items were removed due to gender and/or race-based DIF, determined via Mantel-Haenszel differential item functioning estimation. For example, a question (Figure 5-a) related to the global distribution of earthquakes exhibited DIF relative to gender on both the pre- and post-tests. Men outperformed women on this question, regardless of individual test results. While it is not clear if this question was biased, and therefore testing something other than conceptual understanding, it was temporarily removed from the test bank to ensure fairness.

Rasch analysis has been used to facilitate the development of statistically similar 15-item GCI sub-tests from the bank of GCI questions. The statistical similarity falls from the development of a single scaling function that applies to all sub-tests. Four items were chosen as anchor items for the sub-tests, specifically the two most difficult items, the easiest item, and one intermediate item. The remaining 65 validated items were divided into 11 bins, where each bin is made up of items that are closely grouped on the Rasch scale. The first step in estimating a scaling function for the sub-scales is the averaging of item difficulty estimates within each bin. Item difficulty estimates within a single bin generally have standard deviations on the order of ±0.15, with highest deviations related to the ends of the Rasch scale. Average item difficulties for each bin where then used to determine a relationship between true score, on a 0-15 scale, and theta. The resulting relationship between raw score and scaled score, as fit by the statistical package JMP, looks like ($R^2 = 1$):

$$S_{GCI} = 16.76 + 4.30\,R_{GCI} + 0.115(R_{GCI}-7.5)^2 + 0.042(R_{GCI}-7.5)^3 - 0.0017\,(R_{GCI}-7.5)^4, \tag{2}$$

where S_{GCI} is the scaled score on a 0-100% scale and R_{GCI} is the raw score on a 15-item GCI sub-test. For more details on creating a GCI sub-test, please see http://newton.bhsu.edu/eps/gci.html.

Equation 2 is used to convert raw scores to scaled scores that are assumed to be directly correlated to conceptual understanding (Figure 4). Notice that the scale is linear in the middle, but elongated on the edges; this elongation means that larger conceptual leaps are required for changes in score at the low or high ends. In essence, the "conceptual" difference between a raw score of 0 and a raw of 5 is much greater than the concep-

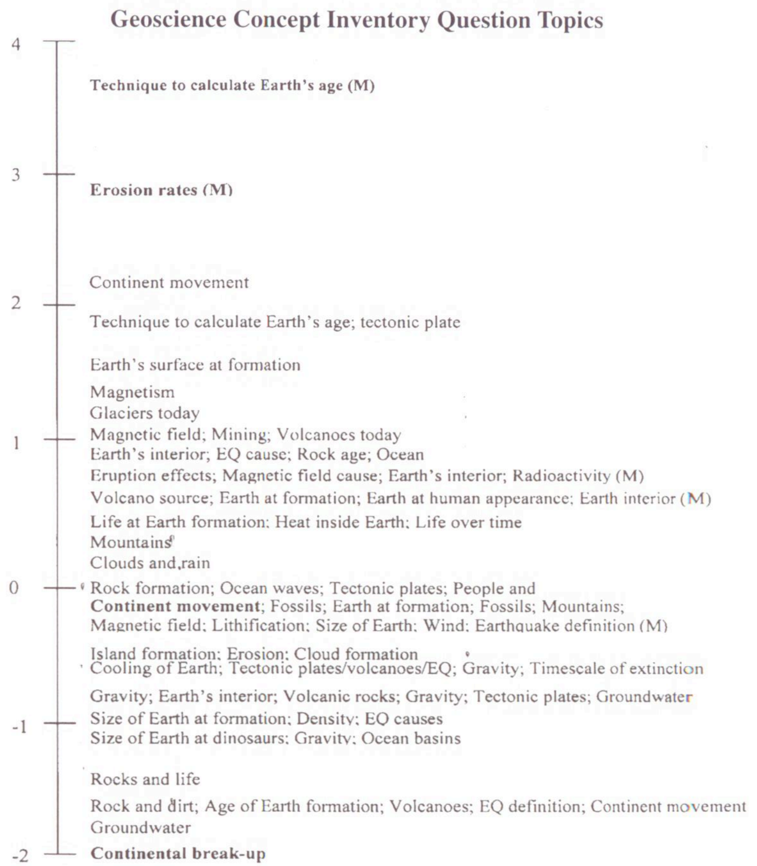


Figure 3. Rasch scale developed for GCI. Rasch analysis of test questions from Fall 2002 and 2003 pre-testing results on revised questions (*N*=3955 college students); these results were very similar to post-test Rasch results. The Rasch scale runs from –2 to +4, where positive values represent the "difficult" end of the scale. Four items in bold are anchor items that appear on all GCI sub-tests. A Rasch level of 0 implies that students scoring a 50% on the inventory have a greater than 50% chance of getting items falling below 0 correct, and less than 50% chance of getting items that fall above zero correct. M refers to questions where students were prompted to choose all possible answers. Labels refer to topics of each GCI question evaluated. M refers to questions where students were prompted to choose all possible answers. Analyses were performed using Quest (Adams and Khoo, 1996). Questions are available online at: http://newton.bhsu.edu/eps/gci.html.

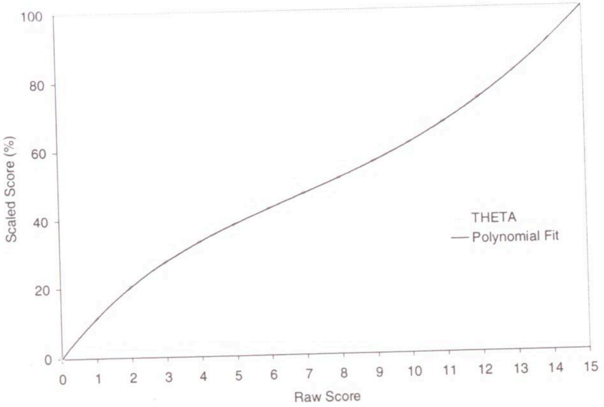


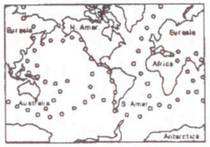
Figure 4. Comparison of raw GCI scores and scores scaled using one-dimensional Rasch modeling. Gray dots are raw score-theta conversions calculated from logit values generated by Quest (Adams and Khoo, 1996). Theta values have been converted to scaled scores on a 0-100% scale. The solid line is the polynomial fit as in eq. (1).

tual difference between 5 and 10. The transformation to scaled scores reflects this conceptual difference.

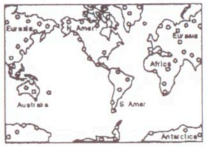
Overall, students in the test population found the test difficult. For example, during 2002 testing students had similar but statistically different pre- and post-test means (pre = 42.2±12; post = 45.8±13; n = 1295 student). Analysis of sub-populations indicates that students with low pre-tests (<40%, n = 388) dominate this effect (pre = 32±7, post = 41±10), with extreme significance on a t-test ($t_{stat} = 1.96 < t_{crit} = 15$) and an effect size of 0.46 (Libarkin and Anderson, 2005). The distribution of test items along the Rasch scale is itself interesting, and provides some insight into the conceptual understanding of entry-level geoscience students. In particular, the presence of rote knowledge does not necessarily imply conceptual understanding. In this study, 82% of students (N = 2192) correctly indicated that the Earth is ~4 billion years old (Figure 5-b) prior to instruction. However, after instruction (N = 1907) only 18% of students were able to correctly indicate that absolute age dating is the primary technique employed to calculate the Earth's age. In addition, only 5% knew that U-Pb is the only method among those listed that can be used when dating the Earth (Figure 5-c). Rather, the majority of students believed that the comparison of fossils or rock layers or the use of carbon were primary absolute age dating techniques. These data indicate that while students are familiar with geologic topics prior to instruction, in this case the Earth's age, few understand underlying concepts vital to complete conceptual understanding, even after instruction.

a. The following maps show the position of the Earth's continents and oceans. The o's on each map mark the locations where earthquakes occur most frequently. Which map do you think best represents where earthquakes occur most frequently on Earth?

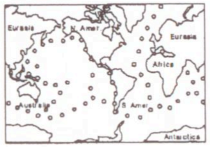
Circle one: A B C D E



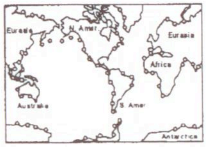
A. In continental and oceanic crust, and along continental margins



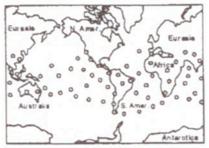
B. Mostly in continental crust



C. Mostly in oceanic crust



D. Mostly along continental margins



E. Mostly in warm climates

b. If you could travel back in time to when the Earth first formed as a planet, how many years back in time would you have to travel?

(A) 4 hundred years
(B) 4 hundred-thousand years
(C) 4 million years
(D) 4 billion years
(E) 4 trillion years

c. Some scientists claim that they can determine when the Earth first formed as a planet. Which technique(s) do scientists use today to determine when the Earth first formed? Choose all that apply.

(A) Comparison of fossils found in rocks
(B) Comparison of different layers of rock
(C) Analysis of uranium and lead in rock
(D) Analysis of carbon in rock
(E) Scientists cannot calculate the age of the Earth

Figure 5. Sample questions from the GCI. Bias and assertions about difficulty are based upon IRT analyses performed using Quest. a) Question removed because of bias. b) Third easiest of the 69 validated questions. C) Most difficult of the 69 validated questions. Notice that the easier question deals with knowledge of the Earth's age, while the most difficult question asks how that knowledge was established.

Discussion and Conclusion

The geosciences is a rich and complex field that encompasses a wide range of disciplines relevant to the biological, chemical, physical, and astronomical STEM fields. Entry-level college courses can focus on a wide range of topics, and may survey a variety of concepts or may focus on topical subjects (Macdonald et al., 2005). As a consequence of this diversity in instructional content in higher education geosciences, assessment of entry-level geoscience courses has lagged behind other STEM fields. The GCI, with its flexibility in specific content, provides a mechanism for assessment and comparison of a range of courses.

The application of three distinct theoretical foundations to concept inventory development was vital in producing the Geoscience Concept Inventory. Although time consuming, the rich information gained from collection and analysis of qualitative data suggested questions and distracters that were not obvious from classroom experience or a review of existing research. In addition, the development of question stems based upon questionnaire and interview responses allowed creation of items that were not initially obvious to the developers. Consequently, the grounding of both questions and responses in student experiences and perceptions allowed us to create an assessment instrument that is generalizable to a large and diverse population of students. Finally, the use of IRT analysis to convert raw test scores to scaled scores provides a more meaningful look at student conceptual understanding and gain. Specifically, ongoing analysis of GCI results offers insight into student conceptions both before and after instruction, and we and other researchers are currently exploring the relationship between conceptual change, as measured by pre- and post-instruction scores on the GCI, and instruction in entry-level geoscience classrooms nationwide.

The use of Rasch and DIF analyses in development of the GCI suggests that IRT approaches have a rich potential in concept inventory development in higher education. A unique methodology was developed for creation of 15-item GCI sub-tests that can be scaled to equivalent scores. This has the unique application of allowing assessment of similar courses that cover different content using comparable measures. Future applications of this approach will incorporate other concept inventories onto the established GCI Rasch scale. In particular, concepts in physics, chemistry, and biology are often fundamentally important to understanding in geosciences. Items from established assessment instruments in these fields can be evaluated for both differential item functioning and for item estimates. Faculty teaching geophysics courses, for example, will then be able to incorporate FCI questions into their GCI sub-test, maintaining the comparability of sub-tests while enhancing the content flexibility of the GCI item bank. Ultimately, it may be possible to compare assessment results in different disciplines, opening up the door to myriad questions about the relationship between instructional approaches and disciplinary differences in conceptual change.

Acknowledgements

The authors wish to thank all teachers, students, and scientists who participated in this study. Research collaborators, particularly Julie Dahl, are also thanked for their support in collection and analysis of qualitative data. This study was funded by the National Science Foundation through an Assessment of Student Achievement in Undergraduate Education grant (DUE-0127765; DUE-0350395) to Libarkin and Anderson. Libarkin was partially supported by an NSF Postdoctoral Fellowship in Science, Mathematics, Engineering and Technology Education (DGE-9906479) during the first years of this project. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Adams, R. J., and Khoo, S. T. (1996). Quest: The interactive test analysis system, Version 2.1 [Computer software]. Camberwell, Australia: Australian Council for Educational Research.

Adams, T. L. (1997). Addressing students difficulties with the concept of function: Applying graphing calculators and a model of conceptualized chance. *Focus on Learning Problems in Mathematics*, *19*(2), 43-57.

Anderson, D. L., Fisher, K. M., and Norman, G. L. (2002). Development and validation of the conceptual inventory of natural selection. *Jounal of Research in Science Teaching*, *39*, 952-978.

Basili, P. A., and Sanford, J. P. (1991). Conceptual change strategies and cooperative group work in chemistry. *Journal of Research in Science Teaching*, *28*, 293-304.

Bond, T. G., and Fox, C. M. (2001). *Applying the rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Clauser, B. E., and Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning items. *Educational Measurement: Issues and Practice*, *17*, 31-44.

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, 12*, 65-68.

DeLaughter, J. E., Stein, S., Stein, C. A., and Bain, K. R. (1998). Preconceptions about earth science among students in an introductory course. *EOS, 79*(36), 429-432.

Dodick, J., and Orion, N. (2003). Cognitive factors affecting students understanding of geological time. *Journal of Research in Science Teaching, 40*(4), 415-442.

Embretson, S. E., and Reise, S. P. (2000). *Item response theroy for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Glaser, B., and Strauss, A. (1967). *The discovery of grounded theory.* Chicago: Aldine Press.

Gobert, J. D. (2000). A typology of causal models for plate tectonics: Inferential power and barriers to understanding. *International Journal of Science Education, 22*, 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, 36*(1), 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, 4*, 5-7.

Gronlund, N. E. (1993). *How to make achievement tests and assessments* (5th ed.). Boston: Allyn and Bacon.

Haig, B. D. (1996). Grounded theory as scientific method. *Philosophy of Education 1995: Current Issues* (pp. 281-290). Urbana, IL: University of Illinois Press.

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.

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, 50*, 5-14.

Halloun, I. A., and Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics, 53*(11), 1043-1055.

Happs, J. C. (1984). Soil genesis and development: Views held by New Zealand students. *Journal of Geography*, 177-180.

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

Kurdziel, J., and Libarkin, J. C. (2001). Research methodologies in science education: Assessing students' alternative conceptions. *Journal of Geoscience Education, 49*, 378-383.

Lambert, J. (2005). Students' conceptual understandings of science after participating in a high school marine science course. *Journal of Geoscience Education, 53*(5), 531-539.

Lewis, E. L., and Linn, M. C. (1994). Heat energy and temperature concepts of adolescents, adults, and experts: Implications for curricular improvements. *Journal of Research in Science Teaching, 31*(6), 657-677.

Libarkin, J. C. (2001). Development of an assessment of student conception of the nature of science. *Journal of Geoscience Education, 49*(5), 435-442.

Libarkin, J. C., Anderson, S., Dahl, J., Beilfuss, M., Boone, W., and Kurdziel, J. (2005). College students' ideas about geologic time, earth's interior, and earth's crust. *Journal of Geoscience Education, 53*(1), 17-26.

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*, 394-401.

Lindell, R. S., and Olsen, J. P. (2002). Developing the lunar phases concept inventory. *Proceedings of the 2002 Physics Education Research Conference.* New York: PERC Publishing.

Liu, X., and McKeough, A. (2005). Developmental growth in students' concept of energy: Analysis of selected items from the timss database. *Journal of Research in Science Teaching, 42*(5), 493-517.

Lyle-Elkins, N. M., and Elkins, J. T. (2004). Assessment of the effectiveness of interdisciplinary expeditionary field trips on student misconceptions in geosciences. *Geological Society of America Abstracts with Programs, 36*(5), 554.

Macdonald, R. H., Manduca, C. A., Mogk, D. W., and Tewksbury, B. J. (2005). Teaching methods in undergraduate geoscience courses: Results of the 2004 on the cutting edge survey of U.S. faculty. *Journal of Geoscience Education, 53*(3), 215-219.

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 Technoogical Education, 15*(2), 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, 51*, 205-216.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Oliva, J. M. (2003). The structural coherence of students' conceptions in mechanics and conceptual change. *International Journal of Science Education, 25*, 539-561.

Posner, G. J., Strike, K. A., Hewson, W. P., and Gertzog, W. A. (1982). Accommodation of a scientific conceptions: Toward a theory of conceptual change. *Science Education, 66*(2), 211-227.

Rhoads, T. R., and Roedel, R. J. (1999). *The wave concept inventory: A cognitive instrument based on Bloom's taxonomy of the frontiers in eduction conference.* San Juan, PR: IEEE Publications.

Schoon, K. J. (1992). Students' alternative conceptions of earth and space. *Journal of Geological Education, 40*, 209-214.

Steer, D. N., Knight, C. C., Owens, K. D., and McDonnell, D. A. (2005). Challenging students ideas about earth's interior structure using a model-based, conceptual change approach in a large class setting. *Journal of Geoscience Education, 53*(4), 415-421.

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, 22*(5), 539-555.

Vosniadou, S., and Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24*, 535-585.

Wieman, C., and Perkins, K. (2005). Transforming physics education. *Physics Today, 58*(11), 36-41.

Windschitl, M., and Andre, T. (1998). Using computer simulations to enhance conceptual change: The roles of constructivist instruction and student epistemological beliefs. *Journal of Research in Science Teaching, 35*, 145-160.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Yeo, S., and Zadnik, M. (2001). Introductory thermal concept evaluation. *The Physics Teacher, 39*, 496-503.

Zeilik, M., Schau, C., and Mattern, N. (1999). Conceptual astronomy. II. Replicating conceptual gains, probing attitude changes across three semesters. *American Journal of Physics, 67*(10), 923-972.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (dif): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa, Canada: Directorate of Human Resources Research and Evaluation National Defense Headquarters.

Appendix

Quest codes and annotations

```
1.     Title GCIpost
2.     data_file post.dat
3.     codes 0123456789
4.     format course 1-4 gender 5 race 6 items 7-33
5.     group (gender=1) ! male
6.     group (gender=2) ! female
7.     group (race=0) ! white
8.     group (race=1) ! nonwhite
9.     key 111111111111111111111111111 !score=1
10.    estimate
11.    show >>postshow.out
12.    show items >>postitems.out
13.    show cases >>postcases.out
14.    logit_table >>postlogit.out
15.    compare item_ests ! group=male,female >>postgender.out
16.    compare item_ests ! group=white,nonwhite >>postrace.out
17.    itanal >>postraw.out
18.    itanal scored >>postscored.out
19.    correlate course,gender,race >>postcorr.out
```

Line 1. "Title" specifies a heading that will appear at the top of all output.

Line 2. "data_file" indicates the name of the file that contains the data set to be analyzed.

Line 3. "codes" indicates item response data which are to be considered valid for analysis

Line 4. "format" indicates the format of the data. In this case, course information is in columns 1-4, gender and race are specified in columns 5 and 6, respectively, and item responses appear in columns 7-33.

Lines 5-8. "group" specifies subgroups that will be used later in the analysis.

Line 9. "key" indicates the scoring key.

Line 10. "estimate begins the Rasch model estimation.

Line 11-13. "show" allows the output data to be viewed, and in this instance is being sent to output files.

Line 14. "logit_table" produces equivalence tables between raw scores and Rasch estimates.

Line 15-16. "compare item_ests" calculates item bias indices, including a Mantel-Haenszel test of DIF

Line 17-18. "itanal" provides a response alternatie analysis

Line 19. "correlate" provides product moment correlations, in this case between identification variables.