Measuring Student Knowledge of Landscapes and Their Formation Timespans

Alison Jolley,[1] Francis Jones,[1,a] and Sara Harris[1]

ABSTRACT

Geologic time is a crucial component of any geoscientist's training. Essential knowledge of geologic time includes rates of geologic processes and the associated time it takes for geologic features to form, yet measuring conceptual thinking abilities in these domains is challenging. We describe development and initial application of the Landscape Identification and Formation Test (LIFT), a concept inventory for measuring abilities to identify landscapes and their formation timespans. Test development included careful choice of concept questions followed by a cycle of validation steps involving student and expert think-aloud interviews. We then administered the test, together with eight validated questions about geological time, to 96 university students in second year and fourth year geoscience courses. Results showed that students' abilities and confidence were more closely aligned with their general knowledge about geologic time than with the level of the course in which they were enrolled. Students were better at identifying landscapes than estimating how long they take to form, and both students and experts had the most difficulty with intermediate formation timespans. Details about students' errors, including common landscape misidentifications and systematic errors in estimating formation timespans, can help instructors prioritize the content and pedagogy of their courses. The LIFT is a validated concept inventory that is available for anyone to use as a pre–post, diagnostic, progress, or end-of-degree assessment that can provide valuable feedback about knowledge and learning to students, instructors and program administrators. © *2013 National Association of Geoscience Teachers*. [DOI: 10.5408/12-307.1]

Key words: concept inventory, landscapes, landforms, formation timespans, process rates

INTRODUCTION

An ability to work with geological time concepts is fundamental for geoscientists, evolutionary biologists, physicists, and anyone interested in comprehending the nature of our planet. In particular, geoscientists in all specializations must be able to work comfortably with widely varying ranges of time, including the rates at which landscapes change (Allen, 2008). Consequently, educators have flagged deep time concepts as crucial for geoscience learning, and most universities and professional organizations require experience with geologic time for completion of corresponding degrees or professional accreditation. Despite this widespread need for expertise there appears to be a shortage of studies that focused upon how people learn about the timespans involved in the formation of landscapes or surface features.

Our work was motivated partly by these general needs for understanding about geological time and rates, and partly by interviews with experts who identified geologic "rates and processes" as one of six key areas of competency regarding geological time (Rhajiak, 2009). Further motivation came in the form of results from the Student Perceptions about Earth Sciences Survey (SPESS; Jolley et al., 2012). In particular, student reactions to the SPESS statement "When I look at a landscape, I have an idea of how long it took to form" were sometimes mixed, and did not always reflect the expected growth when the test was used both before and after taking geology courses. In addition, while there are many previous and ongoing efforts to build concept tests that support learning and assessment in various aspects of the geological sciences (Libarkin, 2008; Cheek, 2010), there appear to be few studies that have focused on landscape identification and estimation of how long particular landscapes take to form.

In this paper, we report on efforts to test student knowledge about landscapes and their formation timespans. We detail the steps taken to construct a validated concept test, and report results from employing it in second year and fourth year geosciences courses taught to students with a variety of backgrounds. Implications regarding student knowledge and instruction about these subjects are also discussed.

Received 11 March 2012; revised 14 October 2012; accepted 10 December 2012; published online 14 May 2013.

[1]Department of Earth and Ocean Sciences, University of British Columbia, 6339 Stores Road, Vancouver, British Columbia, Canada, V6T 1Z4.

[a]Author to whom correspondence should be addressed. Electronic mail: fjones@eos.ubc.ca. Tel.: 604-822-2138.

BACKGROUND

Concept Inventories

In the science education community, there is a growing recognition of the importance of assessing a student's ability to work with key concepts of a discipline rather than simply testing recall of discipline specific facts and figures (e.g., Hestenes et al., 1992; Libarkin, 2008; Adams and Wieman, 2011; Arthurs, 2011). However, finding ways to measure whether learning at the conceptual level has occurred is difficult—and building reliable instruments is nontrivial (Adams and Wieman, 2011). Perhaps the most commonly cited example of these types of assessments is the Force Concept Inventory, which focuses on determining the ability of a student to work in a mature way with Newtonian mechanics concepts (Hestenes et al., 1992).

Several opportunities arise for educators once a validated concept inventory has been developed. Examples include diagnosing abilities of incoming students in diverse classes

(a common need in geoscience departments that offer elective courses for a broad range of students), measuring progress of students as they mature from novices toward experts in their chosen specialization, assessing the impact of courses or particular pedagogic initiatives, and comparing students across courses or institutions. Results from the Force Concept Inventory revealed how little university students learned in traditionally taught introductory physics courses (Hake, 1998) and catalyzed pedagogical change in physics at many institutions. Since then, concept tests have been developed for many science disciplines, to be used, as in physics, as measures of student learning in various situations (Libarkin, 2008).

Student Knowledge of Geologic Time

In the geosciences, the most well known validated assessment is the Geoscience Concept Inventory (GCI; Libarkin and Anderson, 2005), which assesses geologic knowledge at the introductory level, including some aspects of geologic time. A recent review of 79 concept studies in the geosciences (Cheek, 2010) concludes that teaching and learning about geologic time tends to focus on when certain events occur in Earth's history (Earth's formation, the dinosaur extinction, etc.), or in what sequence (e.g., Trend, 2000; Zen, 2001; Hidalgo et al., 2004; Libarkin et al., 2007). Only a few studies address timespans involved in formation of landscapes or surface features. For example, Dodick and Orion (2003a, 2003b), using the Geologic Time Aptitude Test (GeoTAT), found that most adolescent age students conceptualized strata of similar thickness as having formed over the same length of time. Kusnick (2002) analyzed essays written by preservice teachers to find that most far underestimated the time it takes rocks to form, with 29% saying it took years or less. Most of Kusnick's 24 subjects had already completed a university-level geology course. Rule (2005) found that less than one third of preservice teachers surveyed ($n = 67$) had scientifically accurate ideas about how long it takes petroleum to form. College and university students in New Zealand typically overestimated the time it took local soils to form (Happs, 1984).

As a further indicator that timespans of formation may not be in the forefront of people's thoughts about geology, Dove (1998) noted that students have trouble visualizing processes in the past, which might not be operating in the present, and cannot be directly observed. Ford (2003) reported that only 3% of sixth graders even mentioned time when asked "How do you think rocks are formed?" We are not aware of similar work with postsecondary students.

Although the studies reviewed by Cheek (2010) provide an indication of difficulties that students have with geologic timespans, the interview protocols and/or question sets used were not thoroughly validated with think-aloud interviews of students and experts, as recommended by Adams and Wieman (2011). A validated concept inventory about geologic time, the Geologic Time Assessment Tool (GTAT; Rhajiak, 2009), does exist but it does not include any questions regarding landscapes themselves or timespans of landscape formation processes.

Student Confidence

Results from the perception survey cited above suggested that students' confidence in their abilities to work with landscapes dropped after taking an upper-level geoscience course. This unexpected result is one reason we incorporated questions about "confidence" as part of our concept test. Studying confidence turns out to be challenging partly because the term is not always clearly defined. Schraw (1997) interprets confidence as the ability to predict one's performance accurately on test questions. Others (e.g., Bong, 2001; Burton and Mattietti, 2011) sometimes use the word "self-efficacy" to describe this type of confidence. This conflation of self-efficacy with confidence contrasts with Bandura's (1977) definition of confidence which can be understood as the belief in one's agentive capabilities; i.e., that one can produce given levels of attainment. For our study, we use the word "confidence" to mean surety that one's performance on a particular, domain-specific task is accurate or correct, which is consistent with how students in our validation interviews interpreted the term.

Other studies have attempted to tie confidence to achievement with specific tasks or concepts. For example, Jordan et al. (2009) focused on confidence regarding plate tectonic concepts, and found that men exhibited overconfidence. Burton and Mattietti (2011) showed that students displayed overconfidence on all but one of the questions asked about concepts of stratigraphy and geologic time. Exploring this complex realm of confidence in detail is beyond the scope of this work; however, we do report on basic results from a simple measure of students' own perception of confidence relating to specific questions about landscapes.

This Study

To our knowledge, there is not yet a validated test that assesses student abilities to estimate what spans of time are required for particular geologic features to form. Consequently, our goals for this work were to:

1. Develop a validated assessment to fill this gap, and
2. Use this assessment to explore knowledge and, to some extent, confidence among an initial group of beginner to advanced geology undergraduate students.

To pursue these goals, we developed the Landscape Identification and Formation Test (LIFT). For simplicity, in the text of this paper we use the term "landscapes" to describe "geological features" or "landforms." The LIFT was designed to quantify a person's ability to identify landscapes, to estimate the time it takes for them to form, and to estimate their confidence associated with those abilities. In order to enable a comparison of students' understanding of landscapes and timespans of formation with their understanding of geological or deep time, the LIFT included eight questions from Rhajiak's (2009) GTAT. We administered the LIFT in two geoscience classes, one offered primarily to second year students and the other primarily to fourth year students. Results from this first deployment allowed us to explore its potential for gathering useful data about student learning and knowledge about geological timespans and landscapes at different levels of geologic experience. Outcomes provide insight about varying abilities of students with a range of backgrounds and geologic maturities.

METHODS

The Landscape Identification and Formation Test (LIFT) was developed by following the general procedure recommended by Adams and Wieman (2011). The sequence of steps involves first identifying key concepts, then developing open-ended questions, conducting validation interviews with both students and experts, creating forced answer questions, validating these multiple choice versions of the questions, and finally, administering the test (Jolley, 2010). Appendix 1 provides an example of one complete question, and the complete LIFT can be found as a supplement to the online version of this article (available at: http://dx.doi.org/10.5408/12-307s1).

Establishing Important Topics

In interviews, instructors at our institution had identified "rates and processes" as one of six key concepts that geoscience students should be familiar with (Rhajiak, 2009). Specifically, they noted that beginner students (after their second year) should have knowledge of timespans of basic geologic processes (mountain building, volcanism, lava cooling, metamorphic events) and at the advanced level (before graduation), students should be able to quantify geological processes from chemical and physical rate laws. The LIFT aims to address these expectations for beginners, as well as incorporating the crucial first step of identifying the physical result of those processes—the specific landscapes. Sixteen geologic features were chosen for the primary validation stage based on information in a standard geomorphology textbook (Trenhaile, 2007), with formation timespans encompassing minutes through tens of millions of years.

Developing Open-Ended Questions

Based on the identified key concepts and landscapes, preliminary open-ended questions were developed for use in a first round of interviews with students. This step is important because it makes it possible to generate multiple choice options which are rooted in students' alternative conceptions, and which are written in language students commonly use (Adams and Wieman, 2011). It also helps ensure that images chosen were unambiguous (i.e., that students only made errors because features were unfamiliar and not because the images were confusing).

Ten validation interviews were conducted with paid volunteer students, five in their second year and five in their fourth year. After signing a consent form, the volunteer was presented with the landscape images, and asked to first consider the questions for each image silently. Most students spent less than 45 seconds considering each question. Then they went back through the images, verbalizing their thought processes used to determine answers. Questions that students asked of the researcher were not answered until the end of the interview. The test questions were revised as needed throughout the validation stages, both for content and format. This think-aloud interview procedure is described by many authors of qualitative methods, for example, Ericsson and Simon (1998), Adams and Wieman (2011), and Feig and Stokes (2011).

In this first round of interviews, students were asked to write a word or phrase identifying the feature in the image, and then provide an estimate of that landscape's formation timespan to the nearest order of magnitude (i.e., seconds, minutes, hours, days, weeks, years, 10 yrs, 10^2 yrs, 10^3 yrs, etc.). Interestingly, both second and fourth year students often gave ranges both smaller and larger than an order of magnitude, indicating their difficulty with timespan estimates.

The aim of the LIFT is not to quantify student knowledge surrounding the processes of formation (e.g., erosion, uplift, etc.), but instead to focus on how long these processes take. However, it is important to note that students do express thoughts about the processes themselves as they answer the questions, and do not merely attempt to recall answers about timespan of formation. As one example, a second year student considering the U-shaped valley image said: "Yeah, so there must have been a glacier in there that, um, sort of scraped out the middle of it and took along with it a bunch of rock to make that valley. Um, so, I wasn't really sure but then I remembered I think a Milankovitch, or Ice Age cycle is, uh, 40,000 years."

Based on these think-aloud interviews, 5 of the initial 16 images (karst, stromatolite mounds, Olympus Mons on Mars, salt flats, and a mesa) were removed due to lack of clarity. The best format for delivering images was also determined: All but one of the students preferred to view projected images instead of printed versions. One student suggested using the term "feature" (where relevant) instead of "landscape" as a blanket term. All subsequent students interviewed agreed this made questions clearer, so the change was implemented.

Regarding questions about confidence, it was considered crucial to find unambiguous wording and answer options for estimating confidence. This meant determining whether confidence ratings given by students corresponded to their visible verbal and physical cues. Both percentage and Likert scales were tried, and all but one student preferred a percentage over a Likert scale, with the scale divided into fifths rather than quarters. See Appendix I for the final format of questions.

Developing Forced Answer Questions

Because the landscape identification questions involve recognition, offering multiple-choice options would defeat the purpose of the test. Therefore the identification portion of the LIFT is open-ended, while the landscape formation timespan, geologic time, and confidence portions contain forced answers.

Once suitable landscapes were identified and question wording was refined via student interviews, experts were interviewed to determine appropriate true ranges of formation times. Expert interviews were also open-ended in order to determine the best possible answers without restriction. Six geology or geography experts took the test in an interview format, and one took the test using e-mail. All expert volunteers agreed the coverage of landscapes or features was compatible with knowledge of a competent graduating geoscience student. Similar to student interviews, experts were asked to give answers in terms of order of magnitude. In order to use expert responses as the basis for answer options, a range for best possible answers was constrained by dropping formation timespans given by two or fewer experts. After that, most landscapes' formation timespans were limited to two "orders of magnitude." Figure 1 summarizes expert responses and the time ranges chosen as options for the corresponding multiple-choice questions.

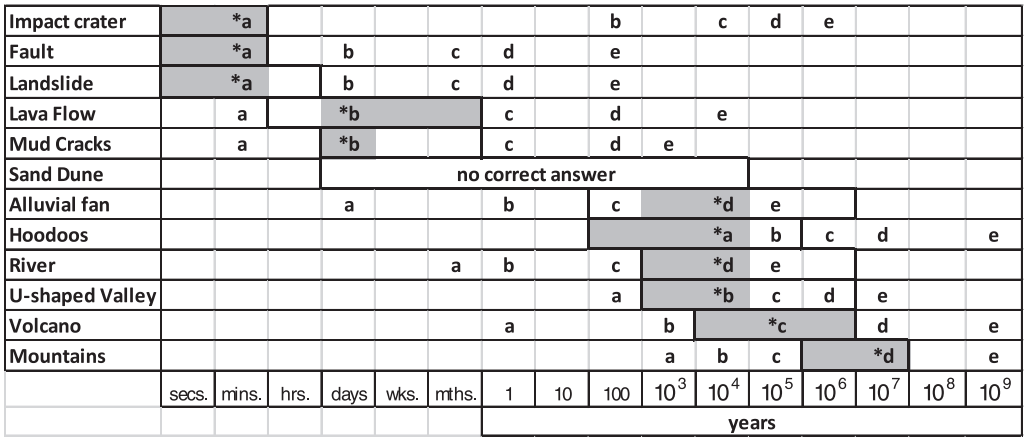

FIGURE 1: Range of expert responses and test answer options for the formation timespans of all twelve landscapes. Shaded spans are expert ranges after dropping open-ended responses given by two or fewer experts. Thick lined boxes indicate the full range of all responses given by experts. Letters "a" through "e" indicate options we chose for the forced-answer multiple choice questions based on student and expert interviews. Starred letters indicate the correct choice for each question. See text for further discussion.

(Note that when referring to "orders of magnitude" we are referring to columns in Fig. 1. This is not strictly correct but it is convenient.)

In order to generate the multiple choice options for each timespan question, the range of consensus from experts was used to define the correct answer, and results of interviews with students helped specify erroneous alternatives outside the range of acceptable answers. For landscapes with a wide range of possible formation timespans, the correct answer was set as one of those time ranges, then four distracters were chosen outside that range. All questions were clearly phrased requesting the "best" answer.

Because the answer choices for the formation timespans do not involve complicated wording and their structure was influenced by the open-ended interviews, the final timespan answer choices were not revalidated with another round of student interviews. The question stems remained as they were in the original suite of interviews, in which they were interpreted as intended.

In addition to the 12 landscape questions, 8 questions from the GTAT (Rhajiak, 2009) were selected in order to characterize each student's knowledge of geological time. These questions (validated in a similar manner as the LIFT questions) were chosen because, when used as part of a pre–post test, they were able to demonstrate positive learning gains of students in a third year course about geologic time. Since these questions measured learning in that course, it seemed reasonable to assume that they would help characterize the expertise of students taking the LIFT.

Data Collection

The LIFT was administered to two separate classes during a regular lecture period, and took approximately 30 minutes to complete in each class. None of the authors were instructors in either of the classes. Students were given a paper copy of the LIFT and each image was projected on a screen at the front of class for 45 seconds. This amount of time was deliberately chosen based on results of interviews. The identification question was open-ended, while the timespan and the two corresponding confidence questions were multiple-choice. After the landscape portion of the test was complete, ten minutes were given for the eight questions from the GTAT. Demographic information was also collected, including gender, age, major, and which other geosciences courses the student had taken.

These two classes are offered by a department that graduates approximately 100 undergraduate students per year. The institution is a large, research-oriented university with roughly 40,000 full time undergraduates. The two classes were a second year introductory mineralogy course for geoscience and geological engineering majors and a fourth year paleontology course offered to both geoscience majors and life science or general science majors (total $n = 96$). These classes were selected in order to collect responses from both introductory and advanced geoscience students, approximately characterizing students near the beginning and end of a geology undergraduate program. In terms of background, 2 of 71 students in the second year course had completed one or more classes related to geological time and/or landscapes. In the fourth year course, all students had one or more such courses and 9 of 25 had taken three or more courses. In the second year course, 20% were nongeoscience majors, most in other science disciplines. The remaining 80% spanned geology, geography, and geological engineering. In the fourth year course, 64% of students taking the LIFT were geological science majors while 36% were general life science or biological science majors who had one general prerequisite course about geological time.

For the most part, the administration of the LIFT ran smoothly and the preferences for its format that students expressed in the validation interviews held true with the larger groups. However, one image was not as clearly interpreted by students in the classroom as it was in the validation interviews (the landslide). Many of the students began to whisper when the image was projected, despite being instructed not to talk during the test and not having done so up to this point. It is useful for teachers who use such images to realize that an image that is seemingly clear to 10 different novices and 7 different experts when seen on

TABLE I: Scoring rubric for the identification section of the LIFT, and common incorrect responses.

Q#	Landscape/Feature	Other Acceptable Answers	Common Incorrect Responses
1	Alluvial fan	Fan, fluvial fan	Landslide (or other type of mass movement)
2	Lava flow	Lava bed/deposit/field, aa/pahoehoe, volcanic	Rock face/bed, rocky area, metamorphic rock
3	Impact crater	Meteor crater, crater	Volcanic crater, sinkhole
4	Hoodoos	Badlands	Limestone/karst column/cliff, canyon
5	Fault	Earthquake	Sinkhole, urban/city
6	Mountains	Mountain range	Glaciers, horn
7	Sand dunes	Dunes	Desert
8	Volcano	Stratovolcano, composite volcano	Forest, mountain
9	River	Meandering/braided river and/or oxbow lake, stream, fluvial	Delta
10	Mud cracks	Desiccation cracks, cracked soil	Dried up sediment, evaporate/evaporation
11	Landslide	Mass wasting, landslide scar, slide	Cliff, mountain, road
12	U-shaped valley	Glacial valley	Valley (not specific), glacier

a computer monitor may still be poorly recognized by some students when projected in the classroom.

Data Processing

All student-written data were transcribed and multiple choice answers were scored against the correct answers as determined by the expert interviews described above. The open-ended identification questions were marked by two of the authors as either correct or incorrect (no partial marks) based on a rubric (Table I). Interrater agreement was 97%.

When scoring confidence, a value was assigned from the middle of the confidence range chosen by the student. For example, the <20% range was represented by 10% confidence, and the 40%–60% range was represented by 50% confidence.

After scoring, students were grouped in four different ways: by course, major, GTAT score, and gender. Comparisons among groups were analyzed to identify patterns of knowledge about landscapes and formation timespans, and corresponding confidence in that knowledge.

Validity and Reliability of the LIFT

The LIFT is the product of several iterative development stages, all of which add successive evidence for the validity of the instrument. Three aspects of validity are typically discussed in concept inventory development: *construct validity*, *content validity*, and *criterion* (or *communication*) *validity* (e.g., Libarkin and Anderson, 2006). *Construct validity* refers to the evidence that the content being covered is of importance to experts in geoscience. This was addressed in Rhajiak's (2009) interviews of six instructors at the institution studied, establishing key concepts in geologic time for both second and fourth year geoscience students. *Content validity* refers to expert confirmation that the test questions measure the concept intended. This was addressed with the seven experts who took and commented on the LIFT during its development. Not only did they confirm that the questions measured the concepts of landscape identification and formation timespans effectively, their responses formed the basis for the correct answers on the final, forced answer test. *Criterion validity* is achieved by ensuring that the stem and answer choices are interpreted by the test takers as intended. This was achieved by having ten students take the open-ended version of the LIFT in the form of think aloud interviews. No further interviews were conducted with the forced answer version of the LIFT, due to the simple nature of the answer choices (order of magnitude timespans represented by numbers). The stems remained the same as in the interviews. Finally, after the test was administered, the suitability of final options was demonstrated using a frequency count of each question's options. Only 2 of the 60 options (5 options for each of 12 questions) were never chosen, and 53 options were chosen by four or more of the 96 students tested.

Content and criterion validity of the GTAT were established during the development of that assessment (Rhajiak, 2009). Although minor wording changes were made postdevelopment, these were grounded in expert experiences and were field tested with approximately 100 students in a course focusing on concepts of life in geologic time. Thus, it is safe to assume that the same questions would still be understood by majors with more content knowledge in geology.

The Cronbach's Alpha coefficient of reliability for the LIFT (including the GTAT questions) is 0.72. However, it should be noted that such measures of internal consistency are not necessarily the best representation of reliability for a concept inventory because they quantify the extent to which the questions measure a single construct (Adams and Wieman, 2011). Because one of the major goals of a concept inventory is to assess multiple constructs across a range of topics, an extremely high internal consistency is not desired. An additional measure of reliability can be provided by calculating a Ferguson's Delta value. Ferguson's Delta measures the discriminatory abilities of the instrument. In other words, it characterizes the extent to which the test is able to differentiate between persons of differing abilities (Ferguson, 1949; Hankins, 2008). It is desirable to have a Ferguson's Delta of >0.90. The Ferguson's Delta for the LIFT is 0.95.

RESULTS

Overall Scores on Identification of Landscapes

The percentage of all students ($n = 96$) who correctly identified each landscape used in the LIFT ranged from a low

TABLE II: Averaged scores by class, major, and experience.

	GTAT Score Mean %, S.E.%[1]	ID Score Mean %, S.E.%[1]	TS with correct ID[2] Mean %, S.E.%[1]
2nd Yr class ($n = 71$)	48,[5] 2.1	60, 2.1	55, 2.3
4th Yr class ($n = 25$)	66,[5] 3.5	65, 3.9	58, 3.8
2nd Yr Geo[3] ($n = 24$)	52,[5] 3.1	61,[4] 3.8	52, 3.8
4th Yr Geo[3] ($n = 16$)	70,[5] 3.4	72,[4] 4.7	55, 4.2
Beginner ($n = 55$)	38,[6] 2.4	57,[5] 1.6	52,[4] 2.7
Advanced ($n = 41$)	71,[6] 2.4	67,[5] 1.8	60,[4] 2.9

[1]Standard error of the mean.
[2]Percentage of correctly estimated timespans for those landscapes that were correctly identified.
[3]Geosciences major.
[4]$p < 0.05$.
[5]$p < 0.01$.
[6]$p < 0.001$.

of 20% for hoodoos to 98% for volcanoes. Landscapes that were identified by less than 60% of students were the landslide, lava flow, mud cracks, alluvial fan, hoodoos, and U-shaped valley. The range of success rates suggests that the collection of landscapes chosen for the LIFT was appropriate because none were impossible to identify, none were correctly identified by all students, and the success rates ranged quite uniformly from low to high. There were no significant differences in overall scores by gender.

Each of the 12 landscape images generated common incorrect answers (Table I). These are useful for geoscience educators as they reflect the kinds of misidentifications that are likely to be most prevalent among undergraduates.

Scores on Identification for Different Groups of Students

Students enrolled in a higher-level geology course might be expected to perform better on the LIFT than students in a lower-level geology course. The average identification scores for students in the fourth year course were indeed slightly higher than those in the second year course, but the differences were not statistically significant (Table II). The only significant difference in scores between the two courses was for the geologic time questions from the GTAT, for which fourth year students scored higher than second year students. The variety of backgrounds of students in these courses likely partially accounts for similarities in LIFT scores. Both courses included some nongeoscience majors, who were less likely to have formal experience with landforms regardless of the level of course in which they were enrolled. Also, 40% of students in the fourth year course had not taken or were not currently taking a course on landscapes. Thus, the mixed enrolment in these courses blurs the progress seen for students in geosciences programs, at least for landscape identification.

When considering only students registered as geoscience majors, those in the fourth year course did outperform those in the second year course on landscape identification ($p < 0.05$) and on the GTAT ($p < 0.01$), even though numbers in each of these partitions are small ($n = 16$ and $n = 24$, respectively; Table II).

Students, regardless of their major, who have a better grasp of geologic time concepts might be expected to perform better on the LIFT, particularly in estimating how long different landscapes take to form. Therefore, we chose to use student performance on the eight geologic time questions from the GTAT to partition students into two groups: "advanced" students who scored five out of eight or higher on the GTAT ($n = 41$), and "beginner" students who scored four or lower ($n = 55$). None of the GTAT questions expressly address landscape formation so we took GTAT scores as a general measure of experience with geologic time concepts. Based on this partitioning, 25 of 71 students in the second year course and 16 of 25 students in the fourth year course were placed in the advanced group.

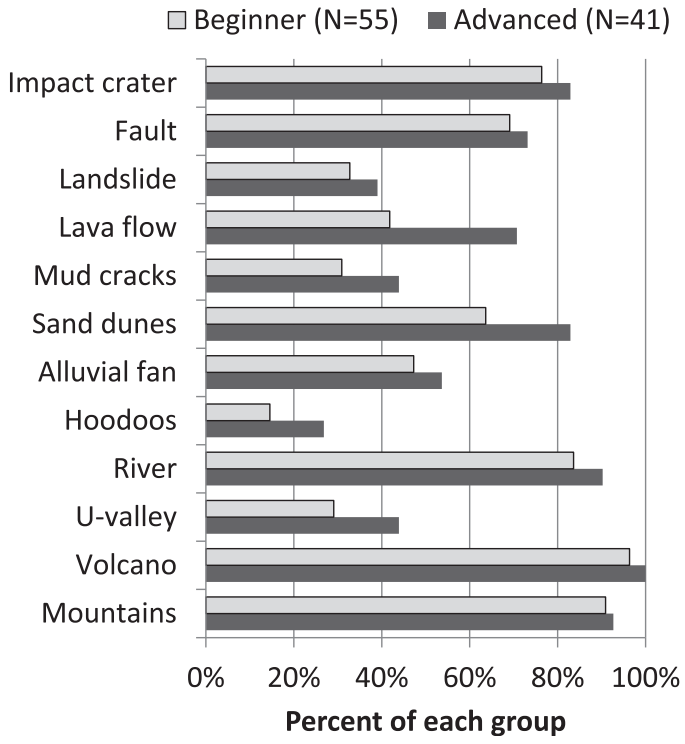

FIGURE 2: Percentage of students correctly answering landscape identification questions. Students were defined as advanced or beginner based on GTAT score. Landscapes are sorted with shortest formation time at the top (impact crater) and longest at the bottom (mountains).

Advanced students were better than beginners at identifying every landscape, with the overall difference between groups significant at $p < 0.05$, indicating that those with better general geologic time skills were also more familiar with this range of landscapes (Fig. 2). When scores on individual questions were analyzed separately, only the lava flow and sand dunes results yielded significant differences between advanced and beginner students, with $p = 0.002$ and $p = 0.038$ respectively. Finally, among advanced students, there were no significant gender differences in identification scores. Among beginner students, men scored significantly higher than women ($t(53) = 2.15$, $p = 0.04$).

Scores on Estimation of Formation Timespans of Landscapes

Only students who correctly identified the landscape were included in analysis of scores on estimating corresponding formation timespans. Overall, the percentage of all students correctly estimating formation timespans for landscapes that they had correctly identified ranged from 19% for the alluvial fan to 100% for the fault (Fig. 3). Advanced students significantly outperformed beginner students on estimating formation timescales (Table 2). Partitioning by course enrolled or year for geoscience majors showed no difference between groups (Table 2). There were no significant differences in overall scores for all students based on gender, nor were there gender differences among advanced students (based on GTAT scores). However among beginner students (based on GTAT score), men scored significantly higher on overall formation timespan estimates compared to women ($t(53) = 2.55$, $p = 0.01$).

Figure 3 illustrates the degree of success for all students at estimating timespans for individual landscapes. The three landscapes most reliably estimated by students are those with the most consistent expert estimates. For example, nearly 60% of students who correctly identified mountains also correctly estimated the corresponding formation timespan, and all expert estimates for mountains fell within two orders of magnitude. Results for sand dunes were not included since experts disagreed about which features in the figure were important.

For landscapes with low success rates on timespan estimates, there may be a concern that results could be have been obtained by random guessing. However, the hoodoos landscape was the only one for which this could be the case, with chi-square testing yielding a 16% probability of obtaining the results by random guessing. For all other landscapes, chi-square tests showed the probability of obtaining the distribution of answers randomly was less than 1%.

Next we consider which landscapes were correctly identified but for which corresponding formation timespans were incorrectly estimated. Landscapes for which more than 60% of students gave correct estimates of timespans were those with shortest formation timespans (impact crater, fault and landslide) and the one largest timespan (mountains). Making estimates for landscapes with intermediate formation timespans appeared to be more difficult for students. Students had difficulty identifying alluvial fans, hoodoos, and U-shaped valleys (Fig. 2) so the numbers of students included in the percentages in Fig. 3 are small. Despite small sample size for some landscapes, Fig. 3 suggests that the pattern of student difficulties with timespan estimation aligns with the degree of agreement among experts, who show greater agreement at the extremes of the timespan spectrum and less agreement at intermediate timespans. The impact crater, for example, elicited expert responses including only "seconds" and "minutes" (two time bins in Fig. 1), while the alluvial fan elicited responses ranging from "hundreds of years" to "millions of years" (five time bins in Fig. 1).

Experts had the widest range of disagreement on formation timespans for the sand dunes image (eight time bins in Fig. 1). This discrepancy was primarily due to different interpretations of the image (e.g., desert formation vs. formation of a single dune). The question was left on the test to see how students' ranges compared with expert's ranges, however, student responses regarding formation timespan for this image were not incorporated into any subsequent analysis.

Finally students were systematically wrong about formation timespans for some landscapes. Based on inspection of answering patterns, the most common wrong choices were adjacent to correct choices in five of the twelve questions. Most students who incorrectly estimated the formation timespans of the alluvial fan and the river chose a shorter timespan than the correct value. For the volcano and the mud cracks, incorrect estimates were usually a longer timespan than the correct value. Errors for mountains were equally likely to be one step shorter or one step longer. These are general remarks, constrained by the fact that there were only five options provided for each question, but it may be useful for instructors to recognize when there are such patterns in errors.

Confidence Results

Group averages, either partitioned by course or by knowledge of geologic time, showed a positive correlation between confidence and test scores. In all groups, average confidence in landscape identification was higher than confidence in formation timespan, consistent with higher average scores in landscape identification.

On identification questions, students generally reported higher confidence in their identification answers when they were correct compared to when they were incorrect. This relationship suggests that these students have a reasonable awareness of their own knowledge and limitations. There was no significant difference in these confidence patterns between advanced and beginner groups.

When considering only those timespan estimates from students who correctly identified the corresponding landscape, advanced students were significantly more confident in their correct formation timespan answers than beginner students ($\chi^2(4, N = 349) = 10.49$, $p = 0.03$). When their answer to formation timespan was wrong, beginner students were significantly more likely to express poor confidence ($\chi^2(4, N = 279) = 10.93$, $p = 0.03$). These results imply greater self-awareness among advanced students regarding what they do know, but better self-awareness by beginner students regarding what they do not know.

Regarding gender-related difference in confidence, men were more confident than women in their landscape identification when their identification was, in fact, wrong ($\chi^2(4, N = 397) = 12.4$, $p = 0.01$). In this case women chose the lowest confidence category more often than men.

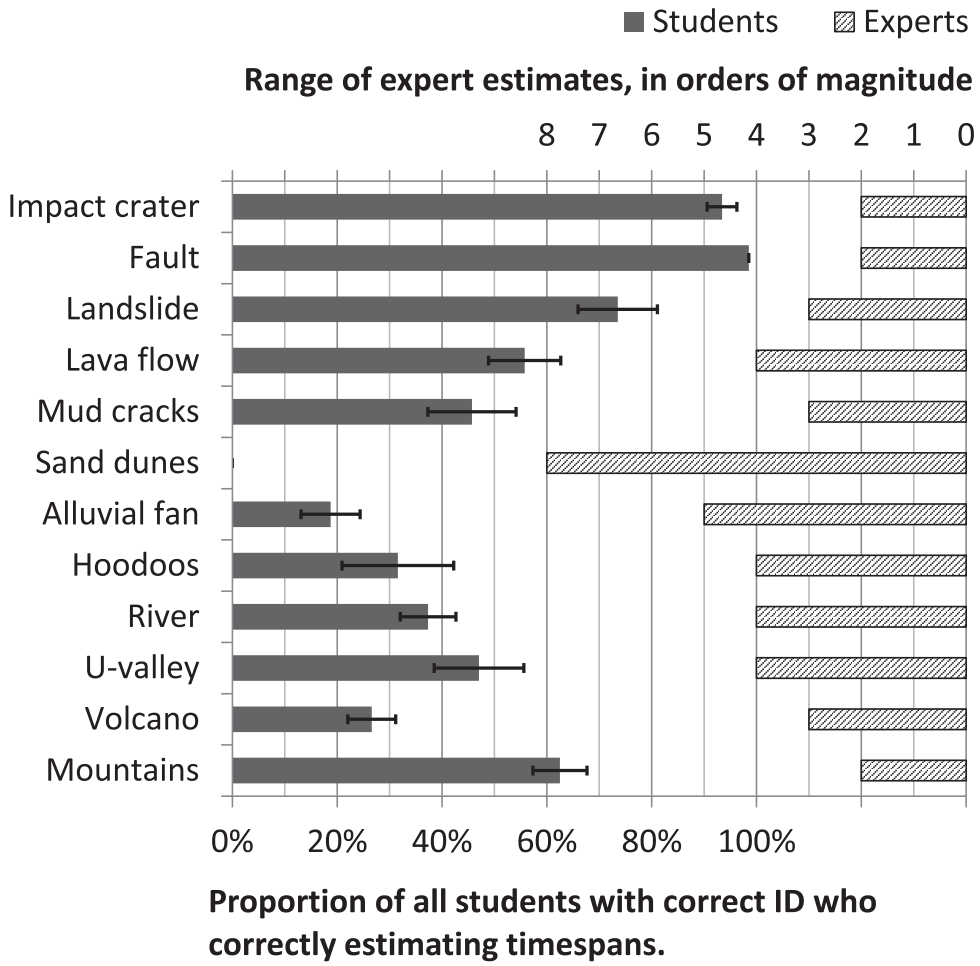

FIGURE 3: Solid bars and bottom scale: Percentage of students correctly identifying landscapes who also correctly estimated each landscape's formation timespan. Standard errors were calculated separately for each question since the number of students correctly identifying each landscape ("n" for the timespan calculations) was different for each question. Student estimates of the formation timespan for sand dunes were not used in the analysis because of the lack of consensus among experts. Patterned bars and top scale: The degree of agreement among experts, also depicted in Figure 1.

IMPLICATIONS FOR GEOSCIENCE EDUCATORS

A Validated Concept Test

A primary contribution of this paper to the geoscience education community is the LIFT itself. The test has undergone validation with both expert and student interviews and is ready for use. We used the LIFT to compare abilities of different students with a range of geoscience backgrounds. Other potential applications of the LIFT include: (1) as a pre–post test of student learning in courses that have objectives related to landscapes and formation timespans; (2) for diagnostic testing at the start of courses that expect students to have already learned these concepts; (3) to test individual development and learning, if administered longitudinally; or (4) as a program-level assessment, in cases in which graduates of a geoscience program are expected to be competent at identifying landscapes and corresponding formation timespans. For our purposes, it was important to include the general geologic time questions (GTAT) with the LIFT, because we sought to compare students based on their general abilities with geologic time. Depending on the goals of the instructor or department, the LIFT could be used with or without the GTAT questions. The item-specific confidence questions provide instructors with an opportunity to investigate one small aspect of students' metacognitive maturity—namely confidence about correctness of these answers. Metacognition is a key component of effective learning (Bransford et al., 2000) and may be of interest in geoscience education or cognition research. However, the confidence questions could be removed if the LIFT is being used specifically as a concept test about landforms.

Implications for Instruction: What Students Know

Evidence that students are better at identifying landscapes than at estimating corresponding timespans of formation is useful for instructors, even though this result is not surprising. Everyday life provides ample opportunity to recognize various landscapes, while exposure to timespans of landscape formation is likely more limited. Instructors might expect even beginning students to identify landscapes they are familiar with, but should be cautious of expecting

even advanced students to identify less common landscapes, even if those landscapes are in one's local area. Geoscience teaching often involves students looking at images of landscapes. Our testing of different images with students during the validation of LIFT questions indicates that specific images for teaching must be selected carefully. Therefore, another useful result of this study is that deliberate elicitation, during instruction, of what students see in images would provide valuable formative feedback to both students and instructors. In other words, instructors cannot assume that students see what experts see in an image.

Our LIFT results also serve as a clear example of one challenge instructors face when teaching senior courses to diverse audiences. We found that students in the fourth year course, in aggregate, were not better informed than students in the second year course, at least regarding landscapes. Because there was more than one pathway to entry into that fourth year course, this result is perhaps not surprising. But it does demonstrate that when students with differing backgrounds are allowed to take an "upper-level" course, instructors cannot assume a consistent level of student expertise in all relevant subdisciplines. Therefore, if prerequisite knowledge is expected in such courses, instructors and students would both benefit from diagnostic testing, especially if resources are provided for students to catch up. As noted above, the LIFT itself could be used as a diagnostic tool for this purpose.

More specifically, our LIFT results show a pattern of student difficulty with intermediate timespans. Why are students more successful at estimating landscape formation timespans at the very long and very short ends of the spectrum? Are they simply characterizing processes as "very rapid," "very slow," or "intermediate"? If true, this may be analogous to the result that preservice teachers tended to lump geologic events into either "extremely ancient," "moderately ancient," or "less ancient" (Trend, 2001). Do instructors (or textbooks) emphasize the extremes in their teaching, and omit (consciously or unconsciously) discussion of timespans in the middle? Given that the expert responses we elicited show the greatest ranges for landscapes of intermediate timespan, it seems likely that these formation timespans are either less well known or experts have looser agreement. Perhaps it is easy and highly relevant to include timespan information when learning about how a meteor impact crater forms (very short) or a mountain range forms (very long), yet more difficult, variable, or less relevant to discuss formation timespans of alluvial fans, hoodoos, and other intermediate timespans.

If intermediate timespans are relevant to, yet absent from, university-level instruction in landscape formation, the following opportunity may arise. In the terms of Windschitl et al. (2008), scientific knowledge is "testable, revisable, explanatory, conjectural, and generative" (p. 943). Therefore, learning about landscapes that form over intermediate timespans may represent an opportunity for students to practice revision, explanation, and conjecture, especially since experts are apparently also in poor agreement about some of these landscape formation timespans. Explicit practice at considering all factors needed to estimate how long it took a landscape to form could give students valuable experience both in the process of science and in working with geologic time in more uncertain, difficult scenarios about which experts might disagree. More specifically to geoscience education, work with intermediate timespans could strengthen future geoscientists' skills at working with deep time.

Implications for Instruction: Student Confidence

Our so-called advanced students were somewhat better at gauging when they were right, but our beginner students were somewhat better at gauging when they were wrong. In other words, it appears from LIFT results that both our beginner and advanced students could benefit from practice strengthening their self-awareness and self-assessment abilities. This is consistent with the understanding that explicit instruction and guided practice to develop metacognitive and self-assessment abilities can be beneficial for human learning in any context (see for example, Bransford, et al., 2000, p. 67–68).

Some previous work has shown that undergraduate men tend to be overconfident in their abilities (Lundeberg et al., 1994; cf. Beyer and Bowden, 1997), including within geological contexts (Jordan et al., 2009). Our data indicate that men and women taking these two courses were generally quite similar in ability and confidence, with one exception. There were gender differences in confidence when students were unable to correctly identify a landscape, with women appearing to be more self-aware than men in this situation. If recognizing one's own gaps and accurately self-assessing one's own knowledge is important for progress in learning, then these data suggest that men in particular may benefit from help and practice in realizing when they are incorrect.

Future Studies

Based on the validation process used during development, and on our preliminary experiences with administering the test, we believe the LIFT is a valid concept inventory that is ready to use as is. One useful adjustment, which would require only minimal revalidation of the test, might be to change timespan questions so they include the full range of "order of magnitude" timespan options for all questions. This would provide patterns of responses that could be more effective than standard five-option questions at revealing students' (and experts') understanding of formation timespans, and at enabling comparison of expert and novice knowledge.

Various pedagogies—both indoors and in the field—are used in geoscience teaching and learning. The relationships among students' geologic time knowledge, landscape knowledge, and confidence in answering specific questions could be further explored in the context of differing pedagogies. By measuring and comparing learning gains (and confidence) in a variety of scenarios, the most effective pathways for building knowledge about landscape processes can be identified.

While we plan to expand use of the LIFT in our setting, there are several reasons why we would like to see the LIFT applied at different institutions. First, investigating similar populations in other institutions would help confirm that our results are not unique. Second, student success at identifying landscapes might be related to familiarity. For example, students we tested live in an area where volcanoes, mountains, and river valleys are common, and these images were the ones that were most successfully identified. However, U-shaped valleys and landslides are also relatively

common near our university, yet these were among the least often correctly identified by students. The LIFT could be utilized, and perhaps revalidated, at schools in other regions, with different sets of common landscapes nearby.

The application of experiential learning with nearby common landscapes could also be assessed with the LIFT. As students have more purposeful encounters with landscapes and geological processes in their local environment, their scores on the LIFT may increase. Additionally, informal experiential learning may contribute to student knowledge of landscapes and their formation timespans. A demographic questionnaire (or interview) given to students tested requiring further information about informal experiences would be a useful addition in order to better reveal relevant experiences that may have occurred outside of the classroom.

In this paper, we only briefly explore confidence in personal knowledge about geologic concepts, as well as the effects of gender and other demographic variability on confidence. By incorporating confidence into the LIFT and similar concept tests, we should be able to illuminate these aspects of geoscience learning in ways that can usefully inform future pedagogy.

CONCLUSIONS

We have developed and validated a test (the LIFT) that is now available for use in a variety of geoscience education settings. The initial application of the LIFT explored differences in knowledge about landscapes and formation timespans among a population of students that spanned a range of geologic experience. Overall, results indicated that undergraduate students were generally competent with identifying most landscapes on the test, with some notable exceptions. Student scores were lower for estimating timespans of landscape formation than for identifying landscapes, especially for landscapes that take weeks to tens of thousands of years to form. Experts also tended to provide less consistent estimates for these intermediate timespans. Student knowledge of geologic time correlated better with landscape knowledge than did seniority or class year. Confidence expressed by advanced students correlated better with their performance when they were correct, however, confidence expressed by beginner students correlated better with their performance when they were incorrect. Men tended to be overconfident in incorrect answers compared to women.

The data presented in this work have important implications for curricula, pedagogy, and research about geoscience education. More time should perhaps be spent teaching concepts of landscape processes, especially of the intermediate range. Care should be taken with all images displayed in the classroom, as they are not always as clear as they appear to a trained or expert eye. The engagement of all genders in the classroom may require more deliberate consideration, although more work is needed to better understand this relationship. Finally, the level of a course may not correlate consistently with expertise of students in the course, and independent measures of knowledge such as the GTAT questions are recommended in cases where the background knowledge of students may vary.

Acknowledgments

We thank instructors and students who participated in interviews and testing. Invaluable feedback about the article was provided by Carl Wieman, Sarah Gilbert, Wendy Adams, and reviewers and editors at Journal of Geoscience Education. The project was supported by the Carl Wieman Science Education Initiative, the Department of Earth, Ocean, and Atmospheric Sciences, and a SkyLight Development Grant, all associated with the Faculty of Science, UBC.

REFERENCES

Adams, W.K., and Wieman, C.E. 2011. Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, 33:1289–1312.

Allen, P.A. 2008. From landscapes into geological history. *Nature*, 451:274–276.

Arthurs, L. 2011. What college-level students think: Student alternate conceptions and their cognitive models of geoscience concepts. *In* Feig, A.D., and Stokes, A., eds., Qualitative inquiry in geoscience education research. The Geological Society of America special paper 474. Boulder CO: GSA, p. 135–152.

Bandura, A. 1977. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84:191–215.

Beyer, S., and Bowden, E.M. 1997. Gender differences in self-perceptions: Convergent evidence from three measures of accuracy and bias. *Personality and Social Psychology Bulletin*, 23:157–172.

Bong, M. 2001. Role of self-efficacy and task-value in predicting college students' course performance and future enrollment intentions. *Contemporary Educational Psychology*, 26:553–570.

Bransford, J.D., Brown, A.L., and Cocking, R.R., eds. 2000. How people learn: Brain, mind, experience, and school: Expanded edition. Washington, DC: The National Academies Press.

Burton, E.P., and Mattietti, G.K. 2011. Cognition and self-efficacy of stratigraphy and geologic time: Implications for improving undergraduate student performance in geological reasoning. *Journal of Geoscience Education*, 59:163–173.

Cheek, K.A. 2010. Commentary: A summary and analysis of twenty-seven years of geoscience conceptions research. *Journal of Geoscience Education*, 58:122–134.

Dodick, J., and Orion, N. 2003a. Cognitive factors affecting student understanding of geologic time. *Journal of Research in Science Teaching*, 40:415–442.

Dodick, J., and Orion, N. 2003b. Measuring student understanding of geological time. *Science Education*, 87:708–731.

Dove, J. 1998. Students' alternative conceptions in Earth science: A review of research and implications for teaching and learning. *Research Papers in Education*, 13:183–201.

Ericsson, K.A., and Simon, H.A. 1998. How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, 5:178–186.

Feig, A.D., and Stokes, A., eds. 2011. Qualitative inquiry in geoscience education research. The Geological Society of America special paper 474. Boulder, CO: GSA.

Ferguson, G.A. 1949. On the theory of test discrimination. *Psychometrika*, 14:61–68.

Ford, D. 2003. Sixth graders' conceptions of rocks in their local environments. *Journal of Geoscience Education*, 51:365–372.

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*, 61:64–74.

Hankins, M. 2008. How discriminating are discriminative instruments? *Health and Quality of Life Outcomes*, 6:36–40.

Happs, J. 1984. Soil genesis and development: Views held by New Zealand students. *Journal of Geography*, 83:177–180.

Hestenes, D., Wells, M., and Swackhamer, G. 1992. Force Concept Inventory. *The Physics Teacher*, 30:141–158

Hidalgo, A.J., San Fernando, I.E.S., and José Otero, I.C.E. 2004. An analysis of the understanding of geological time by students at the secondary and post-secondary level. *International Journal of Science Education*, 26:845–857.

Jolley, A.R. 2010. Identifying landscapes and their formation timescales: Comparing knowledge and confidence of beginner and advanced geoscience undergraduate students. [B. Sc. (Hons.) Thesis]. Vancouver: University of British Columbia. http://hdl.handle.net/2429/23321

Jolley, A., Lane, E., Kennedy, B., and Frappé-Sénéclauze, T.-P. 2012. SPESS: A new instrument for measuring student perceptions in Earth and ocean science. *Journal of Geoscience Education*, 60:83–91.

Jordan, S., Libarkin, J.C., and Clark, S.K. 2009. Too much, too little, or just right? An investigation of confidence, demographics, and correctness [abstract]. *Geological Society of America Abstracts with Programs*, 41:667.

Kusnick, J. 2002. Growing pebbles and conceptual prisms—Understanding the source of student misconceptions about rock formation. *Journal of Geoscience Education*, 50:31–39.

Libarkin, J. 2008. Concept inventories in higher education science. Manuscript prepared for the National Research Council Promising Practices in Undergraduate STEM Education Workshop 2, Washington, DC, Oct. 13–14, 2008. Available from: http://sites.nationalacademies.org/dbasse/bose/dbasse_080106#.UUoFyldrb5S (accessed 20 March 2013).

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.

Libarkin, J.C., and Anderson, S.W. 2006. Development of the geoscience concept inventory. *In* Deeds, D., and Callen, B., eds., Proceedings of the National STEM Assessment Conference, Washington, DC, Oct. 19–21, 2006, p. 148–158.

Libarkin, J.C., Kurdziel, J.P., and Anderson, S.W. 2007. College student conceptions of geological time and the disconnect between ordering and scale. *Journal of Geoscience Education*, 55:413–422.

Lundeberg, M.A., Fox, P.W., and Puncochar, J. 1994. Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology*, 86:114–121.

Rhajiak, J.A.N. 2009. Understanding geological time: A proposed assessment mechanism for beginner and advanced geology students at the University of British Columbia (Vancouver) [B. Sc. (Hons.) Thesis]. Vancouver, BC, Canada: University of British Columbia. Available from: http://circle.ubc.ca/handle/2429/6655.

Rule, A. 2005. Elementary students' ideas concerning fossil fuel energy. *Journal of Geoscience Education*, 53:309–318.

Schraw, G. 1997. The effect of generalized metacognitive knowledge on test performance and confidence judgments. *The Journal of Experimental Education*, 65:135–146.

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

Trend, R.D. 2001. Deep time framework: A preliminary study of UK primary teachers' conceptions of geological time and perceptions of geoscience. *Journal of Research in Science Teaching,* 38:191–221.

Trenhaile, A.S. 2007. Geomorphology: A Canadian perspective. Oxford, UK: Oxford University Press.

Windschitl, M., Thompson, J., and Braaten, M. 2008. Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education*. 92:941–967.

Zen, E., 2001. What is deep time and why should anyone care? *Journal of Geoscience Education*, 49:5–9.

APPENDIX 1. Examples of questions from LIFT.

1. Image projected for 45 seconds. Students answer the following:

(Image Copyright © Marli Miller, University of Oregon. Image source: Earth Science World Image Bank; http://www.earthscienceworld.org/images. Used with permission.)

a) What type of feature is this? _____

b) How confident are you that you recognized the type of feature that is present in the image?

<20% 20-40% 40-60% 60-80% >80%

c) How long did this feature take to form? Choose the BEST answer.

a. days or less
b. years
c. 100s of years
d. 10s of 1000s of years
e. 100s of 1000s of years or more

d) How confident are you in your estimation of the time the feature took to form?

<20% 20-40% 40-60% 60-80% >80%

End of sequence for question 1.

Next image is then shown (total of 12 images used).

Eight geologic time multiple choice questions (from GTAT) after 12 landscape questions.

Demographic information as needed, at the end.