The Oceanography Concept Inventory: A Semicustomizable Assessment for Measuring Student Understanding of Oceanography

Leilani Arthurs,[1,a] Jennifer F. Hsia,[2] and William Schweinle[3]

ABSTRACT

We developed and evaluated an Oceanography Concept Inventory (OCI), which used a mixed-methods approach to test student achievement of 11 learning goals for an introductory-level oceanography course. The OCI was designed with expert input, grounded in research on student (mis)conceptions, written with minimal jargon, tested on 464 students, and evaluated for validity, reliability, and generalizability. The result was a valid and reliable, semicustomizable instrument, with longer 23-item and shorter 16-item versions, as well as flexible grading using either classical one-point-per-item scoring or item-difficulty–weighted scoring. This article is of utility to potential end users of the OCI (e.g., practitioners and researchers) and test developers considering constructing a concept inventory. © *2015 National Association of Geoscience Teachers*. [DOI: 10.5408/14-061.1]

Key words: assessment, concept inventory, classical test theory, item response theory

INTRODUCTION

Concept inventories test for conceptual understanding rather than factual recall (Hestenes et al., 1992). That is, they test individuals' abilities to apply fundamental first principles to answer new questions or problems that they had not previously encountered. Although studies in geoscience concept inventory development can be traced back several years (e.g., Dodick and Orion, 2003; Libarkin and Anderson, 2006; Parham et al., 2010), to our knowledge, such an instrument is lacking for oceanography courses. The existence and use of such an instrument would facilitate systematic analysis of students' prior knowledge (National Research Council, 2000) and individuals' conceptual change over the instructional period (Driver and Odham, 1986; Boyle and Monarch, 1992; Pearsall et al., 1997; Savinainen et al., 2005; Vosniadou, 2007; Lewis and Baker, 2010). When administered as a preinstruction test, a concept inventory can provide instructors with feedback about the students' preexisting knowledge and help inform instructional decisions about how much time to dedicate to certain concepts and how to teach those concepts. When administered as both a preinstruction and postinstruction test, concept inventories can measure learning gains (Hake, 1998; Thomson and Douglass, 2009).

Concept inventory instruments are multiple-choice tests that target a particular construct. A construct is "the concept or characteristic that a test is designed to measure" (American Educational Research Association et al., 1999, p. 5). Two examples of such constructs for cognitive instruments are "the understanding of astronomical concepts" and "the ability to design a scientific instrument" (Briggs et al., 2006, p. 38). These tests are developed based on student thinking and language, rather than being based solely on predetermined content (Hestenes et al., 1992). This is illustrated by the fact that incorrect answer options are not based on instructor speculation, assumptions, or anecdotal experiences but, instead, are developed through research into students' alternate conceptions or misconceptions (Arnaudin and Mintzes, 1985; Thijs, 1992; Arthurs, 2011). As such, the goal in crafting the incorrect answers in a concept inventory is to produce plausible "distractors" (Libarkin, 2008).

Concept inventories are currently available for a number of disciplines, such as astronomy (e.g., Lindell and Sommer, 2004; Lindell, 2005), biology (e.g., Odom and Barrow, 1995; Anderson et al., 2002; Knudson et al., 2003; Garvin-Doxas et al., 2007), chemistry (e.g., Tan et al., 2008), geology (e.g., Dodick and Orion, 2003; Libarkin and Anderson, 2006; Parham et al., 2010), and physics (e.g., Hestenes et al., 1992; Chabay and Sherwood, 2006). Within these disciplines, some of these concept inventories are "conceptually extensive," such as the Geoscience Concept Inventory (Libarkin and Anderson, 2006) and others, such as the Geological Time Aptitude Test (Dodick and Orion, 2003), are "conceptually intensive" (Parham et al., 2010). In other words, conceptually extensive inventories address a range of concepts within a discipline, whereas those that are conceptually intensive focus more deeply on a limited number of concepts. In developing a concept inventory for oceanography, we used the conceptually extensive approach.

Although there are well-established theories and methods from the field of psychometrics to inform test construction in general, there exists no single prescribed approach for developing concept inventories. Thus, we used both Classical Test Theory (CTT) and Item-Response Theory (IRT) to develop and evaluate the Oceanography Concept Inventory (OCI).

The purpose of this study was to evaluate the OCI, which was developed to measure student understanding of oceanographic concepts (Arthurs and Marchitto, 2011). As part of this process, we asked two research questions: (1) to

Received 14 November 2014; revised 27 July 2015 and 8 September 2015; accepted 10 September 2015; published online 14 December 2015.

[1]Department of Earth & Atmospheric Sciences, University of Nebraska-Lincoln, 330 Bessey Hall, Lincoln, Nebraska 68588-0340, USA
[2]Department of Psychology, University of South Dakota, South Dakota Union, 414 East Clark Street, Vermillion, South Dakota 57069, USA
[3]School of Health Sciences, University of South Dakota, Lee Medical Building, Room 302, 414 East Clark Street, Vermillion, South Dakota 57069, USA
[a]Author to whom correspondence should be addressed. Electronic mail: larthurs2@unl.edu. Tel.: 402-472-6353. Fax: 402-472-4917

TABLE I: Advantages and key underlying assumptions of Classical Test Theory and Item Response Theory. Sources: Lord and Novick, 1968; Wright and Stone, 1979; Lord, 1980; Hambleton and Jones, 1993; Embreston and Reise, 2000; de Ayala, 2009.

Metric	Classical Test Theory	Item Response Theory
Advantages	(a) Smaller sample sizes are needed.	(a) Scores describing respondents' ability are not dependent on the choice of items or on the item difficulty.
	(b) Simpler mathematical analyses are required.	(b) Item statistics are independent of the groups from which they were estimated.
	(c) Model parameter estimation is conceptually straightforward.	(c) Produces test models that provide a basis for matching test items to ability levels.
	(d) Associated analyses do not require strict goodness-of-fit studies to ensure a good fit of model to test data.	(d) Test models generated do not need parallel tests for assessing their reliability.
Key underlying assumptions	(a) True scores and error scores are uncorrelated.	(a) Unidimensionality: A single, continuous latent trait (θ) underlies responses to all items in a given test.
	(b) The average score in a population of examinees is zero.	(b) Local independence: A respondent's response on one test item is independent of the individual's responses to other items on the test.
	(c) Error scores on parallel tests are uncorrelated.	

what extent is the instrument valid and reliable, and (2) what potential for generalizability does the instrument possess for use in oceanography courses taught elsewhere.

Our approach to concept-inventory evaluation and the results of that process is of utility to both potential end users (e.g., practitioners and researchers) and test developers. Both groups, for example, need to know the instrument is valid and reliable, how such a determination was made, and how to score the instrument. The qualitative methods used to design the multiple-choice items for the OCI are described in Arthurs and Marchitto (2011); therefore, this article emphasizes the quantitative methods used to evaluate the instrument.

THEORETICAL FRAMEWORK

In this study, the two major stages of the OCI construction—development and evaluation—were guided by different theories. The development stage was informed by grounded theory and CTT, whereas the evaluation stage was guided by CTT and item-response theory (Nunnally and Bernstein, 1994; Bond and Fox, 2007; Boone et al., 2010).

Open Coding From Grounded Theory

Grounded theory is a qualitative, data-driven approach geared toward understanding a key issue (Creswell, 1998). The key issue addressed in this study was students' alternate conceptions or misconceptions of oceanographic concepts. Although we did not use grounded theory to present a particular theory, consistent with a grounded-theory approach, student conceptions were not defined a priori and were, instead, gathered using in-class exercises and think-aloud interviews. A common practice across the numerous permutations of grounded-theory approaches is the open coding of this kind of qualitative data (Charmaz, 2006), and this practice was central in the development of the OCI. According to Charmaz (2006), in this way, grounded theory can serve as an essential precursor to the development of quantitative instruments (e.g., a concept inventory).

CTT and Item-Response Theory

Instrument validity and reliability, which are critical to test developers and end users, are determined during the instrument's construction. An instrument's validity cannot be unequivocally proven, and instead, coherent arguments for its validity must be made for "a unifying concept … [for which] … the degree to which all accumulated evidence supports the intended interpretation of test scores for the proposed purpose" (American Educational Research Association et al., 1999, p. 9). On the other hand, a test is considered reliable when its measurements are consistent and reliable (O'Connor, 1972; DeVellis, 2006). Although validity is evaluated qualitatively, reliability is evaluated quantitatively using statistics. CTT and IRT can be used to evaluate instrument reliability.

CTT is one of the most commonly applied theories in the construction of tests (National Council on Measurement in Education and American Council of Education, 2006). Table I summarizes the advantages and underlying assumptions of CTT. Although CTT is valuable in the test-construction process, it has a key limitation: classical item statistics are sample dependent (Hambleton and Jones, 1993). In other words, true scores, item difficulty, and item discrimination are dependent on the choice of items or on the examinee sample.

IRT is not similarly limited, and its strengths (Battisi et al., 2009; Pathak, 2013) can complement the use of CTT. In particular, IRT attempts to model the relationship between an unobservable variable, conceptualized as a respondent's ability, and the probability of the respondent correctly answering any particular item on a test. Table I lists the advantages and underlying assumptions of IRT.

METHODS

Target Course

We received institutional review board approval for this study. This instrument was designed as a preinstruction and postinstruction test for an "Introduction to Oceanography"

college course, which satisfied one of the university's general education requirements for science. It was taught by Instructor A, had large enrollments (~165 students), and was taught every Spring semester. It was a lecture-based course that was designed based on the course textbook, *Oceanography: An Invitation to Marine Science* (Garrison, 2007). The course was divided into, and taught in, four modules addressing: (1) basics of physical geology, (2) geochemistry of ocean water, (3) physical oceanography, and (4) marine biology.

Locating the Researchers

Feig (2011) discusses the placement of researchers in the broader context of qualitative research. Here, *researchers* refers to those involved in the development and evaluation of the OCI. Instructor A, one of two test developers, was a researcher–participant because he was the instructor of the course for which the instrument was developed. J.F.H. and W.S. were two of three test evaluators and had no interaction with the students involved in this research. L.A. was a test developer and the third test evaluator. She was a researcher–observer because she interacted with students in the project during interviews and during the administration of in-class exercises and tests.

Population

Students enrolled in a large, research-intensive, west-central state university and enrolled in Instructor A's Spring 2008 oceanography course participated in the development phases of the instrument. For the evaluation phase of the instrument, it was administered four times, at least once in three different sections of introductory oceanography courses taught at the same large, west-central state university. Given that the instrument was developed during the Spring 2008 semester, it was administered to Instructor A's class only at the end of the Spring 2008 semester. Determining the effect of the interview process on the students who participated in both the interview and post-instruction administration of the OCI in Spring 2008 would introduce another level of research that was beyond the scope of this project, and we believe there was negligible to no effect for the following three reasons: (1) the interviews focused on what students thought, rather than what the correct answers were; (2) the interviewees were not given the correct answers to the items that were either used during their interviews or in the final version of the items used in the OCI's postinstruction administration; and (3) the interviews occurred more than 1 mo before the OCI's postinstruction administration, and it is thus questionable how much the students remembered about the interview by the end of the semester.

In Spring 2009, the instrument was administered in Instructor A's class at the beginning and the end of the semester. At the beginning of the Spring 2009 semester, the instrument was also administered in Instructor B's class to increase the number of preinstruction student responses available for evaluating the instrument. In other words, in the time available for data collection for this project, Instructor A's classes yielded one semester of preinstruction OCI responses and two semesters of postinstruction OCI responses. Therefore, we wanted an additional set of preinstruction responses. We also needed to obtain preinstruction responses from another similar course to use a modified test–retest method for checking the instrument's reliability. A demographic description of the students in these courses is summarized in Table II.

In the three course sections in which the instrument was administered during the evaluation phase, student participation was voluntary, and no demographic information was collected from the participants. Table II describes the overall demographics of the courses in which students completed the OCI and not the specific demographics of the actual survey respondents. Respondents' demographic data were not collected because they were not crucial for this project. Given the optional nature of study participation and the low class attendance on the day that the OCI was administered in Instructor A's Spring 2008 class, only 50% of the class completed the OCI at the end of the semester. Of the 164 students in Instructor A's Spring 2009 class, 93% completed the test at the beginning of the semester and 85% completed it at the end of the semester. Of the 93 students in Instructor B's Spring 2009 class, 99% completed the test at the beginning of the semester. Overall, 464 student responses were collected for each of the 23 items in the instrument.

Design Framework for OCI Construction

Figure 1 illustrates the design framework used for the OCI's construction. Two stages characterize the construction of the OCI: development and evaluation. Each stage comprised sequential and sometimes iterative phases. In the development stage, the methods were mainly qualitative, whereas quantitative methods were used in the evaluation stage.

The initial analysis of the test items used CTT methods. To supplement the classical test statistics with analyses that were not sample dependent, we also applied item-response statistics to the collected test data. We applied the one-parameter, or Rasch, model to the test data. The one-parameter model is further discussed in *Supplemental File 1: Quantitative Methods Used to Evaluate the OCI*.[4] Evaluation of the OCI addressed item difficulty (P), item discrimination (r_{pb}), reliability using test–retest method and coefficient of stability (r), internal consistency (α), and goodness of fit (Akaike information criterion [AIC]). How they were evaluated is described in *Supplemental File 1*.[4]

RESULTS

OCI Administered as a Pretest–Posttest

The original 23-item version of the OCI was administered as a preinstruction and postinstruction test in Instructor A's Spring 2009 class. Figure 2 compares the preinstruction and postinstruction item-difficulty index. The item-difficulty index when multiplied by 100 represents the percentage of the respondents who answered a given item correctly. Thus, when taken as a whole, Figure 2 suggests that there was an overall gain in student understanding of oceanography concepts. The greatest gains observed were in the concepts of isostatic equilibrium, convection, Coriolis effect, deep and shallow waves, and limitations on productivity. The least gains were observed in the concepts of

[4] The Supplemental File 1: Quantitative Methods Used to Evaluate the OCI is available online at http://dx.doi.org/10.5408/14-061s1.

TABLE II: Demographic makeup of participating instructors' courses. Not all students in these courses participated in the study. Actual study participants were a subset of the instructor courses listed.[1]

When	Instructor	Total	Male	Female	Freshman	Sophomore	Junior	Senior	5th-y Senior	Graduate Student	Nondegree	STEM Major	Non-STEM Major	Undeclared Major
Development stage														
During Spring 2008	A	162	104	58	4	32	54	59	9	1	3	43	108	11
Evaluation stage														
Post-Spring 2008	A	Same course demographics as listed above												
Pre/post Spring 2009	A	164	86	78	7	32	53	65	0	0	0	62	92	8
Pre-Spring 2009	B	93	51	42	42	11	45	32	4	0	0	24	69	0

[1]STEM = science, technology, engineering, and mathematics.

density stratification, heat and temperature, biogeochemical cycling, and food-chain efficiency.

We also examined possible correlations with OCI pre- and postinstruction scores and learning with different measures of student learning, including average homework scores, final clicker scores, average exam scores, and final course grades. There was no correlation between OCI pre- and postscores and learning gains with average homework scores (r^2 values = 0; Fig. 3). There was a weak positive correlation between OCI pre- and postscores and learning gains with final clicker scores (r^2 values ranged from 0.02 to 0.06; Fig. 4). Figure 5 shows weak positive correlations with average exam scores (r^2 values ranged from 0.06 to 0.25), and Fig. 6 shows weak positive correlations with final course grades (r^2 values ranged from 0.04 to 0.25).

Original Version: 23-Item OCI

Table III lists the items, concepts, and learning goals critical for understanding oceanography in Instructor A's introductory-level oceanography course and for which open coding of students' open-ended responses to in-class premodule exercises yielded sufficient variability of student alternate conceptions to develop multiple-choice items.

We used the modified test–retest method described earlier and applied classical test statistics to subsets of data collected from Instructor A (n = 151, Spring 2009) and Instructor B's (n = 93, Spring 2009) courses. Although these two courses have notably different numbers of freshmen, the courses were similar enough for the purposes of this study. In particular, both courses were designed for and targeted toward students without prior background in oceanography, were introductory-level college courses, and were courses that satisfied the university's general science education requirement. These criteria for inclusion were of far greater relevance to this study than students' year in college. The results are summarized in Table IV. A χ^2 analysis of the distribution of correct and incorrect answers that students in each course selected helped to determine how much students from these two courses differed in their preferences for particular distractors, and only one item (Item 3) showed a significantly different spread of answers between the two courses (p = 0.025). Reliability was also checked by calculating r. The two courses had consistent r measurements (Instructor A, r = 0.347; Instructor B, r = 0.334). Internal consistency was also checked by calculating Cronbach's α (α = 0.74).

Shorter Version: 16-item OCI

The Rasch model was applied to the test data collected from the 464 respondents to further evaluate and identify ways to refine or modify the test for use in introductory-level oceanography courses elsewhere.

Before the Rasch model was applied, we determined that the assumptions of unidimensionality and local independence were met in so far as was reasonable for the scope of this study (that is, conducting detailed studies of item order effects and factor analyses were beyond the scope of this particular study). The sample of 464 students' responses were collected for each of the 23 items on the original version of the test was considered a sufficiently large sample for IRT analyses (c.f. Embreston and Reise, 2000).

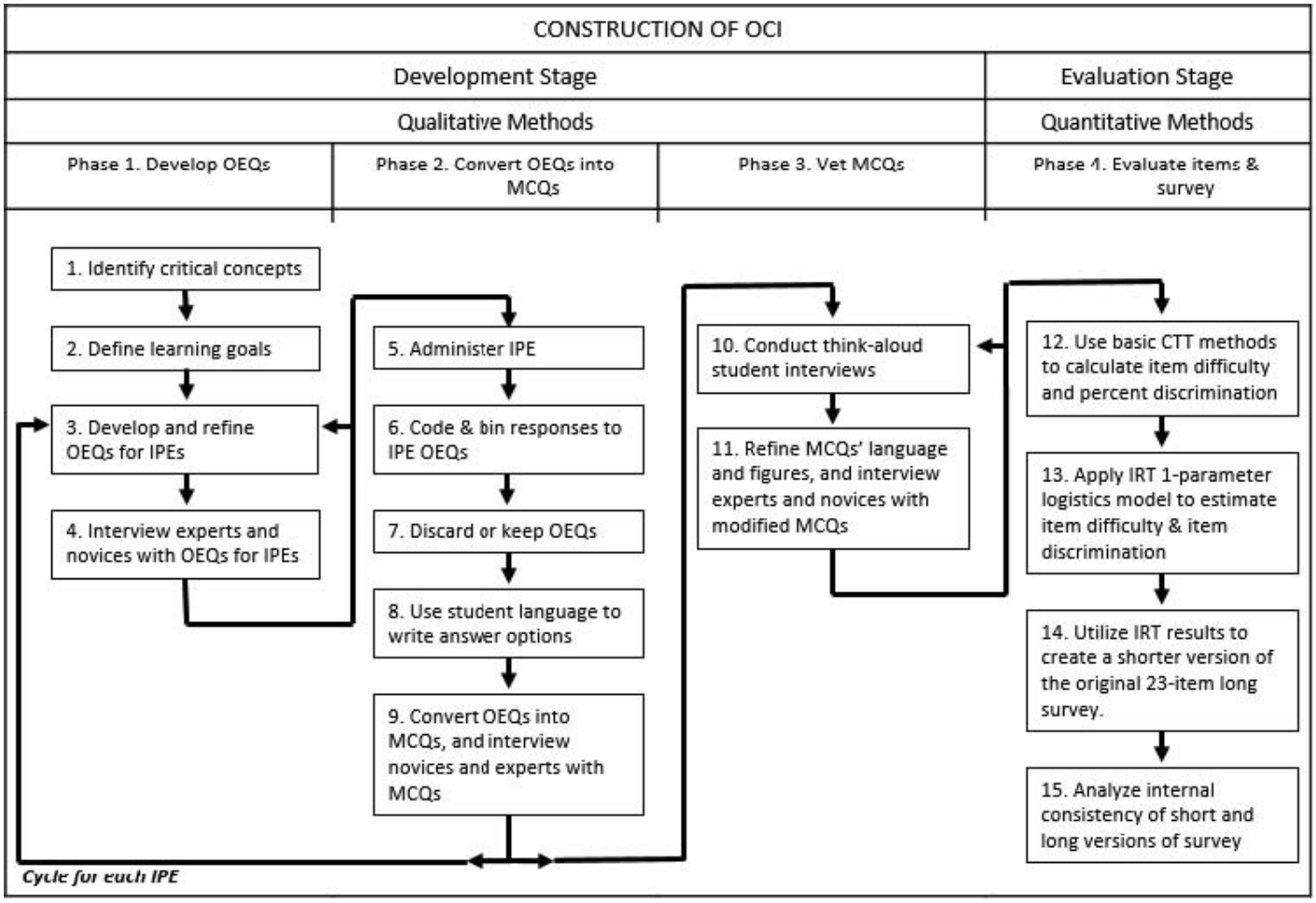

FIGURE 1: Design framework for OCI construction. Two stages characterize the construction of the OCI: development and evaluation. Each stage comprised sequential and sometimes iterative phases. In the development stage, the methods were mainly qualitative, whereas, in the evaluation stage, the methods were quantitative. Abbreviations: OEQs = open-ended questions. MCQs = multiple-choice items, IPE = in-class premodule exercise. Adapted from "Qualitative methods applied in the development of an introductory oceanography concept inventory survey" by L. Arthurs and T. Marchitto, 2011. *In* Feig, A. D. and Stokes, A., eds., Qualitative inquiry in geoscience education research. Boulder, CO: GSA Special Paper 474, p. 101. Copyright 2011 by the Geological Society of America.

Results of the Rasch model for the collected data are summarized in Table V, which shows that the item-difficulty estimates (β) ranged from −2.0 to 1.9. These estimates were used to inform decisions on how to shorten the original 23-item test. For instance, Item 8 was found to be unusually difficult and was removed from the 16-item test. Based only on apparent redundancy in β, eight items were considered for removal from the test. Upon closer, expert (i.e., content expert) examination of this recommendation, only seven items were removed from the original 23-item test because removal of the eighth item would have also removed representation for one of the critical concepts and its associated learning goals. In other words, items with a similar β may measure different aspects of the test construct and cannot be removed based only on β. In this way, we reasonably shortened the 23-item test to a 16-item test, thus providing instructors with the choice between using the longer original version or the shorter version.

Comparison of the Original and Shorter Versions

The shorter 16-item test was evaluated and compared with the original 23-item test. Item difficulty estimates (β) for the 16-item test generated by the Rasch model ranged from −2.0 to 1.1 (see Table V). Based on calculations of Cronbach's α, the 16-item test had slightly lower internal consistency ($\alpha = 0.69$) than the original 23-item test did ($\alpha = 0.74$). The AIC value for the 23-item Rasch model was 12,919.40, whereas the AIC for the 16-item Rasch model was 9,044.26.

Multiple Versions of OCI

Our CTT analyses showed that the original 23-item version of the OCI was valid and reliable for the course population for which it was designed. Evidence for the validity argument, outlined in Arthurs and Marchitto (2011), was derived using qualitative methods originating in grounded theory. Given that the 16-item version is a

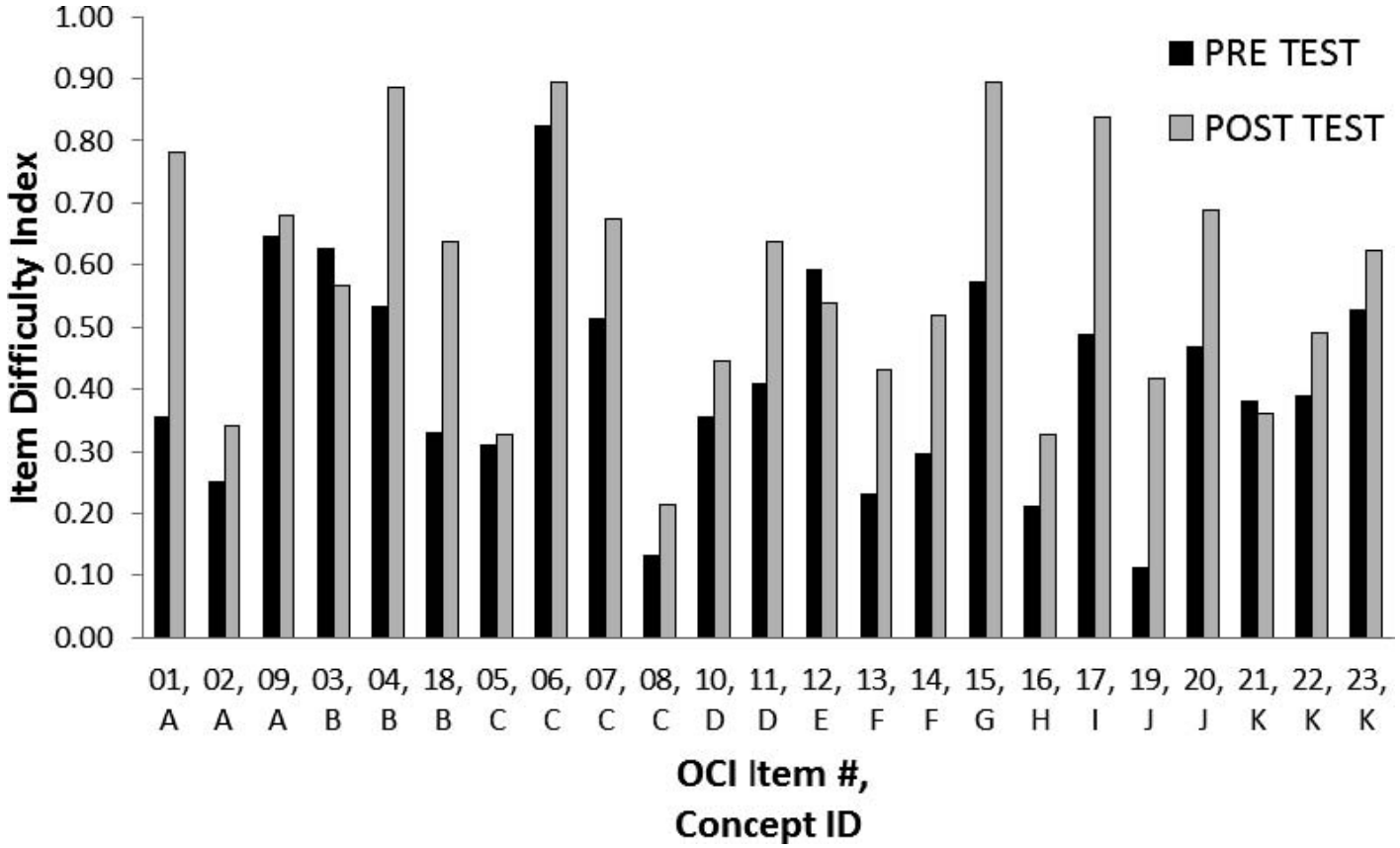

FIGURE 2: Comparison of item difficulty index from preinstruction and postinstruction OCI tests. The Concept ID is cross-linked with Table 3, where one can find the concepts and learning goals associated with each OCI Item number. $n = 122$ matching preinstruction–postinstruction scores.

representative subset of the original 23 items, the same validity argument applies to the 16-item version. Application of classical test statistics provides evidence of the instrument's reliability.

Application of both classical test and item response statistics allowed us to generate a shorter 16-item version of the OCI that is psychometrically similar to the original 23-item version in terms of the range of item difficulties (spanning >3 SD of respondent ability), internal consistency, and coverage of the 11 critical concepts that the test was originally designed to assess.

In addition, the shorter version is composed of 10 fixed items from the original version (Items 1, 6, 11, 12, 13, 15, 16, 17, 19, and 21) and has the option for some flexibility via interchangeable pairs for the remaining six items of the actual 16-item test administered. Pairs were deemed interchangeable if they met the following criteria: the paired items had similar item difficulties (β); the absolute difference in item difficulties between the paired items (β_{diff}) were in the lowest range of β_{diff} among all possible paired item combinations (the low cutoff was arbitrarily set at $\beta_{diff} \leq 0.06$); and the items were mutually exclusive pairs (i.e., a

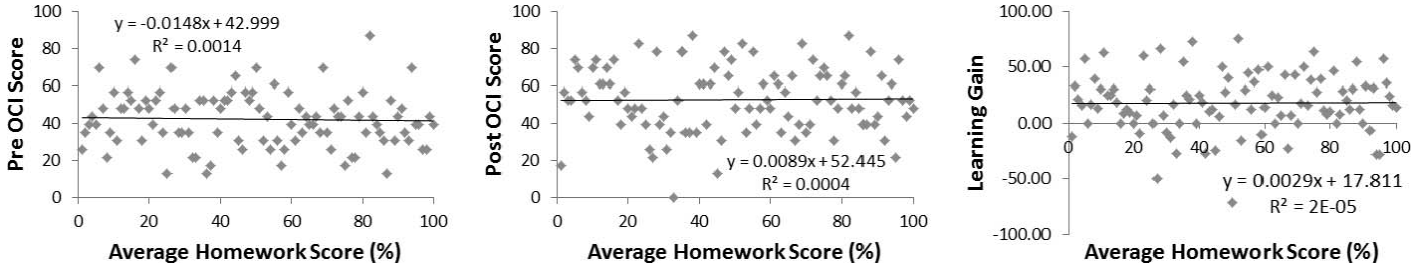

FIGURE 3: Correlation of preinstruction–postinstruction OCI scores and learning gains with average homework scores. The equation of the trend line and r^2 value are displayed on each graph. $n = 122$ matching preinstruction–postinstruction scores.

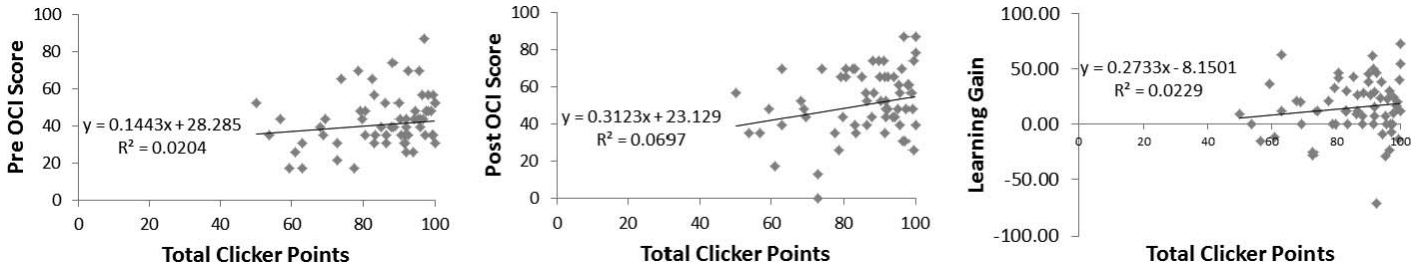

FIGURE 4: Correlation of preinstruction–postinstruction OCI scores and learning gains with average final clicker scores. The equation of the trend line and R^2 value are displayed on each graph. $n = 122$ matching preinstruction–postinstruction scores.

given item occurred in one and only one interchangeable pair). The interchangeable item pairs are Items 2 and 5 ($\beta_{diff} = 0.00$), Items 20 and 23 ($\beta_{diff} = 0.00$), Items 10 and 22 ($\beta_{diff} = 0.03$), Items 9 and 17 ($\beta_{diff} = 0.03$), Items 13 and 19 ($\beta_{diff} = 0.05$), and Items 3 and 7 ($\beta_{diff} = 0.06$). Thus, instructors can select one item from each interchangeable pair and generate alternate versions of this shorter 16-item test that are equivalent in terms of item difficulties and still assess student learning of all 11 concepts.

Scoring Options

The OCI not only has flexibility in the versions that an instructor can use but also has flexibility in its scoring. As with other conceptual tests identified in the "Introduction" section, this test can be classically scored by assigning equal point values to all test items. This type of scoring, although useful, does not account for variability in item difficulty.

One advantage of IRT models is the ability to score tests in such a way that accounts for the variation in item difficulty that is independent of respondent ability. This is typically done using maximum-likelihood methods and IRT scoring software. For example, developing a regression with nonlinear raw scores and IRT latent abilities would be an appropriate way to develop a scoring table so that users can easily convert the examinees' raw response patterns into IRT scores. For IRT logit scores to be interpretable, a rescaling of logit scores into a more-conventional scoring range, such as 0 to 100, is often necessary. However, another possible way to achieve difficulty-weighted scoring without special software is to use the empirically derived item difficulties (β) generated by the IRT model. For example, the lowest item difficulty for the OCI (rounded to the nearest tenth) was -2.0; so, one could add a value of 3.0 to each item difficulty in the 16-item test and assign points for each correct response according to the item's difficulty estimate plus 3.0 (i.e., the total points for an item $= \beta_i + 3.0$). Using this scoring method, correct responses to the least-difficult item on the test ($\beta = -2.0$) are worth 1 point, and the most difficult item ($\beta = 1.1$) are worth 4.1 points. Although this technically is not probabilistic IRT scoring, this differential weighting of items does more-objectively estimate the extent of a student's understanding of the concepts addressed in a test.

DISCUSSION

In our discussion, we address our two original research questions: (1) to what extent is the instrument valid and reliable, (2) what potential for generalizability does the instrument possess for use in oceanography courses elsewhere? We also discuss the OCI's ability to distinguish between lower- and higher-performing students as well as its uses and applications.

Validity and Reliability

Validity is concerned with an instrument's ability to measure what it is intended to measure. Knowledge of student ideas about oceanographic concepts were researched (discussion of these conceptions are to be presented in a future paper) and used to inform the development of the OCI through an iterative process (see Fig. 1). Two different versions of the OCI were developed,

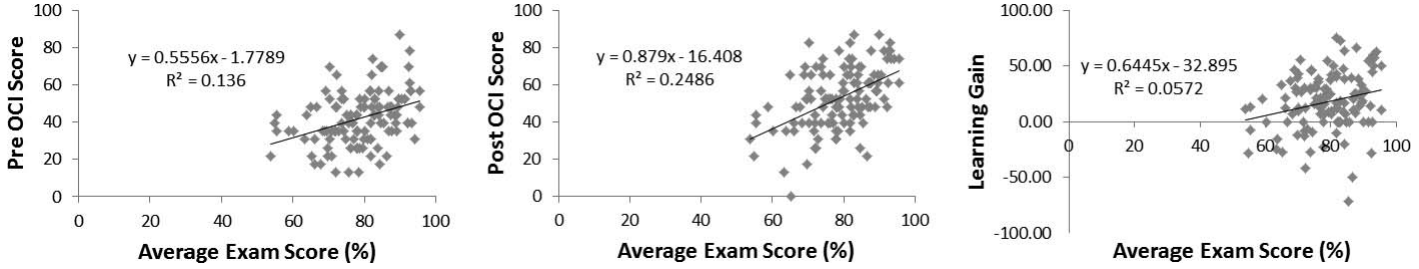

FIGURE 5: Correlation of preinstruction–postinstruction OCI scores and learning gains with average exam scores. The equation of the trend line and R^2 value are displayed on each graph. $n = 122$ matching preinstruction–postinstruction scores.

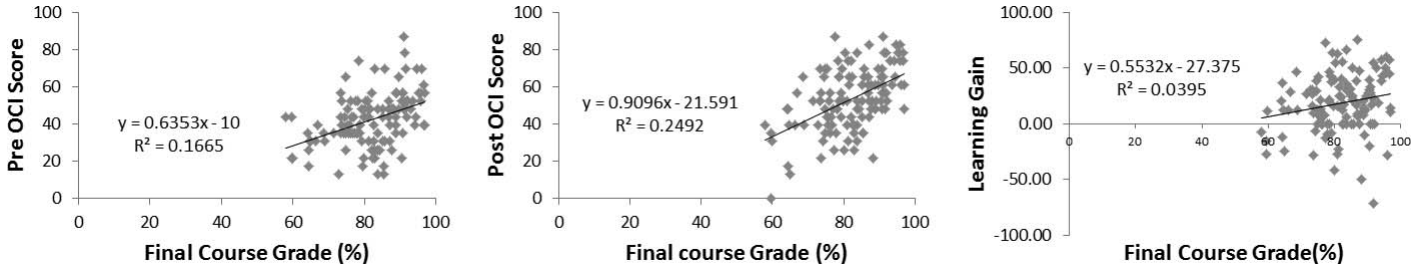

FIGURE 6: Correlation of preinstruction–postinstruction OCI scores and learning gains with final course grades. The equation of the trend line and R^2 value are displayed on each graph. $n = 122$ matching preinstruction–postinstruction scores.

the original 23-item version and a shorter 16-item version. The OCI was valid based on evidence described elsewhere (Arthurs and Marchitto, 2011); so, we focus on instrument reliability here. The reliability of the OCI was supported by multiple lines of evidence.

Reliability is an expression of an instrument's consistency or reproducibility. To characterize the OCI's reliability, we used two different commonly used classical test-theory approaches. The first was the modified test–retest method, and the second was Cronbach's α.

The test–retest method was used to find consistency in two different administrations of the same instrument (e.g., Webb et al., 2006). Thus, one can estimate consistency by examining dis/similarities in the spread of correct and incorrect answers between the two different administrations. In the traditional test–retest method (Sadowski et al., 1992; Kline, 2000), the two administrations are conducted with the same population at two different points in time and assumes consistency of the measurement, such as ability, from the first occasion to the second occasion of administration (e.g., Webb et al., 2006). The very nature of using a concept inventory as a pre- and postinstruction test might seem to ideally lend itself to using the traditional test–retest method because of the two occasions of administration with the same population. However, this is not the case because the main purpose for using a concept inventory as a pre- and postinstruction test, unlike most other types of tests, is to measure *changes in ability*. Thus, we used a modified test–retest method where the same test was administered at the same point in time with two different but similar populations, two introductory-level oceanography college courses designed to satisfy the university's general science education requirement.

The modified test–retest analysis of the preinstruction test data from two oceanography courses taught in Spring 2009 showed consistency between the two administrations. This consistency was evidenced in the range of correct and incorrect answers observed in each administration as well as in the similarities in their coefficient of stability (r) values. Respondents in both courses typically selected a similar range of correct and incorrect answers to each item. Only one item, Item 3, showed a significantly different spread of

TABLE III: List of concepts and learning goals used to develop OCI items.

Concept	Concept ID	Learning Goal	Item No.
Isostatic equilibrium	A	Explain how isostatic equilibrium accounts for the existence of ocean basins.	1, 2, 9
Convection	B	Describe the conditions necessary for the development of a convection cell.	3, 4, 18
Density stratification	C	Describe what causes density stratification and what it leads to; explain the behavior of neutrally buoyant material.	5, 6, 7, 8
Heat and temperature	D	Distinguish between temperature and heat.	10, 11
Biogeochemical cycling	E	Explain the importance of nutrient cycling through seawater, biota, and sediments.	12
Thermohaline flow	F	Explain why and what energy is ultimately required to drive the thermohaline circulation and under what surface conditions deep waters may form.	13, 14
Coriolis effect	G	Describe how the direction and magnitude of the Coriolis effect vary with latitude and velocity.	15
Geostrophic flow	H	Apply geostrophic flow to predict surface water movement	16
Deep and shallow waves	I	Distinguish between deep-water and shallow-water waves on the basis of wavelength and water depth.	17
Limitations on productivity	J	Compare and contrast photosynthesis and chemosynthesis.	19, 20
Food chain efficiency	K	Explain why harvesting older fish has both benefits and risks.	21, 22, 23

TABLE IV: Preinstruction administration of the OCI in Spring 2009 in courses of Instructor A and B show the item difficulties (*P*) and item discriminations (r_{pb}) associated with both courses, providing evidence of the OCI's reliability and reproduceability.[1]

	Item Difficulty (*P*)		Item Discrimination (r_{pb})	
Item No.	A	B	A	B
1	0.36	0.37	0.38	0.39
2	0.25	0.23	0.32	0.33
3	0.63	0.55	0.31	0.51
4	0.53	0.70	0.44	0.33
5	0.31	0.25	0.47	0.51
6	0.82	0.74	0.25	0.35
7	0.51	0.55	0.37	0.18
8	0.13	0.14	0.14	0.09
9	0.65	0.69	0.36	0.40
10	0.36	0.45	0.42	0.25
11	0.41	0.37	0.35	0.29
12	0.59	0.52	0.35	0.38
13	0.23	0.19	0.10	0.34
14	0.30	0.28	0.36	0.36
15	0.57	0.64	0.40	0.45
16	0.21	0.22	0.20	0.34
17	0.49	0.60	0.30	0.34
18	0.33	0.37	0.12	0.03
19	0.11	0.17	0.17	0.34
20	0.47	0.45	0.19	0.23
21	0.38	0.30	0.25	0.32
22	0.39	0.25	0.39	0.21
23	0.53	0.46	0.37	0.35

[1]Instructor A's course, $n = 153$; Instructor B's course, $n = 92$.

TABLE V: Item difficulties (β) and their associated standard errors (σ_β) for the original 23-item and reduced 16-item oceanography test based on a Rasch model.[1]

	Original 23 Items		16 Retained Items	
Item	β	σ_β	β	σ_β
1	−0.2205	0.1108	−0.2237	0.1115
2	0.9555	0.1182	E	E
3	−0.5179	0.1123	−0.5247	0.1131
4	−1.1093	0.1199	−1.1259	0.1207
5	0.9555	0.1182	0.9699	0.119
6	−1.9531	0.1431	−1.9801	0.1439
7	−0.4553	0.1119	E	E
8	1.8689	0.141	E	E
9	−0.9321	0.1170	E	E
10	0.3146	0.1114	E	E
11	−0.0294	0.1105	−0.03	0.1113
12	−0.3526	0.1113	−0.3591	0.1121
13	0.8513	0.1166	0.864	0.1174
14	0.5120	0.1128	0.5203	0.1136
15	−1.2467	0.1226	−1.2651	0.1234
16	1.0879	0.1205	1.1042	0.1213
17	−0.8978	0.1165	−0.9112	0.1173
18	0.1821	0.1109	0.1845	0.1116
19	0.8968	0.1177	E	E
20	−0.3848	0.1117	E	E
21	0.6292	0.1139	0.6392	0.1147
22	0.3454	0.1116	0.3502	0.1124
23	−0.3833	0.1115	−0.3899	0.1122

[1]Note: $n = 464$. E = items omitted from the 16-item logistic model.

answers between the two courses ($p = 0.025$). Nevertheless, there is precedent for retaining such items in concept inventories. For example, a similar analysis was done in the development of another concept inventory, and the developers of that instrument found that four of their 25 items showed a significantly ($p < 0.05$) different spread between the two courses surveyed and the items were retained (Smith et al., 2008).

Using the test–retest method, the coefficient of stability (r) for the two courses were also consistent (Instructor A's course, $r = 0.347$; Instructor B's course, $r = 0.334$), with a mean of 0.340 and a spread of ±0.007. Although the two r values are similar, they also appear low, especially in comparison to commercially available tests where r ranges from 0.8 to 0.9 (Crocker and Algina, 1986). It is difficult to say what explains the notable difference between our derived values and those for commercially available standardized tests. It may be that our process was smaller in scale; that is, we had neither the financial nor human resources to administer the OCI on a scale comparable to that of commercial test-development organizations, who can collect larger data sets for analyzing individual items and tests as a whole. Whatever the reason for the difference, it is more important to note that making such a direct comparison of r values may not be appropriate when evaluating the reliability of a concept inventory because the purpose of commercially available standardized tests and concept inventories are fundamentally different—commercially available standardized tests, such as the Law School Admissions Test, are designed to examine essentially the same set of skills at one point in time, whereas concept inventories are aimed at measuring changes in individual ability over time. Furthermore, psychometric literature states that, although the higher the r value, the better, there is no cutoff for r values (i.e., there is no standard for what is too low), and making correct interpretations of low values for tests that are not expressly about personal history or behavior is not always possible (Madsen, 2004).

As another measure of the OCI's reliability, we calculated the α coefficient. It is the most widely used reliability coefficient and estimates test-score reliability from a single test administration (Webb et al., 2006). It estimates a test's reliability based on the covariation among items

internal to the test and is, thus, also called an *internal-consistency coefficient* (Webb et al., 2006). Although there is no absolute standard cutoff, values of $\alpha = 0.70$ or greater are favorable and indicate strong reliability (Cortina, 1993). The α value for the original version of the OCI was 0.74 and for the shorter version 0.69; therefore, our calculated α values suggest that both versions of the OCI are internally consistent and provide additional evidence for the OCI's reliability.

Ability to Distinguish Students

Tests that measure learning are most useful when they can distinguish between lower- and higher-performing students (e.g., Smith et al., 2008). In other words, if the test is too easy, then all respondents score well and, if the test is too difficult, then all respondents score poorly. For a test to be a useful discriminator of lower- and higher-performing students, its items should reflect a range of item difficulty.

Our pre- and postinstruction results provide evidence that the items on the original 23-item test range in difficulty. Figure 2 is based on classical statistics to calculate item difficulty (*P*) from the preadministration and postadministration of the OCI in the same course, and it indicates that the items span a range of difficulties. Using an IRT approach, we also see that item difficulties ($\beta = -1.95$ to 1.87) range widely in both the original, longer version and the shorter version of the OCI, suggesting that both versions can discriminate between lower- and higher-performing students.

Our test–retest results provide evidence of item discrimination. That is, the average of each item's r_{pb} values from the two courses indicate that all but two items are very good to reasonably good items. Furthermore, based on suggested psychometric cutoffs, none of the items were deemed poor enough to be rejected.

Generalizability and Semicustomizability

Using the IRT framework, we used a Rasch model to further evaluate the original 23-item version of the OCI and check its potential for generalizability. This generated a shorter 16-item version of the OCI and also provided evidence of both versions' reliability (as already discussed above). The original 23-item and the shorter 16-item versions of the OCI are psychometrically similar in their coverage of the original 11 critical concepts that the OCI was designed to assess coefficients of stability (r_{pb}), internal consistencies (α), and range of item difficulties (β, spanning more than 3 SD of respondent ability). In terms of their AIC values, the shorter version shows a better and more-parsimonious fit to the data than the longer version does, although there may be circumstances where the longer version is more appropriate or useful.

AIC values provide relative, rather than absolute, values for model fit. They are used to compare multiple models against (i.e., relative to) one another. For this reason, there are no absolute cutoff values for fit, and absolute AIC values alone are not indicators of fit (Bozdogan, 2000). For our study, the purpose was to use IRT to develop a psychometrically sound, shorter version of the OCI as an alternative option to the original version for test administrators.

In particular, the shorter version requires less administration time and possesses the flexibility to interchange items. The shorter version has six interchangeable pairs of items, which provide administrators of the OCI some flexibility to emphasize or deemphasize certain oceanography concepts through the choice of the interchangeable items. Thus, an instructor can select one item from each pair without compromising the instruments' overall validity and reliability. The OCI is semicustomizable for different administrator needs, in the sense that two versions exist (i.e., long and short versions), the short version has interchangeable item pairs, and the OCI can be scored using the classical one-point-per-item way or in an item-difficulty-weighted way.

The short and long versions of the OCI are backed with evidence that support their validity and reliability. The shorter version, however, may have greater appeal to instructors or researchers with less in-class time to administer the OCI as a pre- and postinstruction test or who want to emphasize or deemphasize certain concepts by selecting from among the interchangeable items.

Regardless of the version that is selected, test administrators also have flexibility in scoring the OCI. The OCI may be scored in the classical way in which all items on the test are assigned equal point values or the difficulty-weighted method to test items based on their β values (as described in the "Results" section). To make direct comparisons in pre- and postinstruction administrations of the OCI in the same group of students, however, the same version of the OCI and the same scoring method should be used before and after instruction to measure potential learning gains that occurred during the period of instruction. The same recommendation holds for those who wish to use the OCI to compare student performance across different sections or courses of introductory-level oceanography.

Uses and Applications

As with other concept inventories, the OCI can assess conceptual understanding before and after instruction. In other words, it can be used to assess the extent to which learning goals were achieved *and* learning gains were made (Thomson and Douglass, 2009).

Such analyses of student performance can help inform course-level and department-level instructional decisions and help instructors identify the impact of certain instructional approaches on student learning (e.g., Savinainen and Scott, 2002; Klymkowski and Garvin-Doxas, 2008; Petcovic and Ruhf, 2008; Arthurs and Templeton, 2009; Marbach-Ad et al., 2009). Similar analyses can be conducted on larger scales, such as at interinstitutional levels for larger-scale research projects (e.g., Hake, 1998).

Although the OCI can be used widely, preadministration validity checks are recommended as part of standard practice. This is to ensure, for example, that the language used in the instrument is understood in the manner intended by the populations to which it is administered. There may be cultural or linguistic differences in how students interpret items that were not already resolved. Greenfield (1997), for example, conducted research that showed ignoring the epistemologies inherent to cultures and languages may lead to misjudgments about an individual's abilities. Recommending preadministration validity checks does not undermine the OCI's trustworthiness; rather, it is standard and responsible practice for instruments to be used when conditions are notably different from the conditions

under which the instrument was developed (Lederman et al., 2002).

Future Work

Although introductory-oceanography courses are not as prevalent as their equivalents in, say, physics, chemistry, and biology, oceanography instruction is nevertheless important enough that ocean literacy standards have been defined (National Geographic Society's Oceans and Life Initiative and National Oceanic and Atmospheric Administration, 2007). Potential studies using the OCI could involve investigating a number of questions, including (1) the oceanography concepts that students find particularly challenging to master; and (2) the instructional approaches most useful for facilitating student learning of oceanography concepts, especially for the most challenging concepts.

Although every effort was made to ensure that items were interpreted in the intended manner, we did have limited resources and, therefore, did not include students in different locations in the United States or internationally. Given that language, culture, and local customs might influence students' interpretations of items in unexpected ways that may then negatively affect their apparent abilities (e.g., Greenfield, 1997), another possible avenue for future research is an expanded study of the validity and reliability of the OCI if is administered across multiple introductory-level courses taught by multiple instructors at different institutions.

CONCLUSIONS

The OCI is a concept inventory test that can be used by oceanography instructors to measure student-learning gains and to evaluate the effect of different instructional approaches on student learning. The OCI is available in its original 23-item version and its shorter semicustomizable, 16-item version. Both versions were demonstrated to be valid and reliable within the context of the study and have potential for broader use. To minimize concern that circulation of the OCI may diminish its value to instructors, we did not include the full set of test items in this article. We would, however, like to see the OCI used more widely, and we will supply the full set of test items and their answers upon request. Interested instructors and researchers should contact the corresponding author.

Feedback from OCI users is welcome and will be used to further evaluate and refine the instrument. Instructors who obtain the full set of test items and their answers from us will be notified of updated versions that become available.

Acknowledgments

We would like to thank Derek Briggs, Katherine Perkins, and Wendy Smith for their guidance and advice on test construction; the instructors who permitted the administration of this survey in their classes; the students who participated in the OCI exercises and surveys; and the experts and novices who provided regular feedback on individual items as they were developed.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Anderson, D.L., Fisher, K.M., and Norman, G.J. 2002. Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*, 39:952–978.

Arnaudin, M.W. and Mintzes, J.J. 1985. Students' alternative conceptions of the human circulatory system: A cross-age study. *Science Education*, 69(5):721–733

Arthurs, L. and Templeton, A. 2009. Coupled collaborative in-class activities and individual follow-up homework promote interactive engagement and improve student learning outcomes in a college-level environmental geology course. *Journal of Geoscience Education*, 57:356–371.

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

Arthurs, L. and Marchitto, T. 2011. Qualitative methods applied in the development of an introductory oceanography concept inventory survey. *In* Feig, A.D. and Stokes, A. eds., Qualitative inquiry in geoscience education research. Boulder, CO: Geological Society of America Special Paper 474, p. 97–111.

Battisti, B.T., Hanegan, N., Sudweeks, R., and Cates, R. 2009. Using item response theory to conduct a distracter analysis on conceptual inventory of natural selection. *International Journal of Science and Mathematics Education*, 8:845–868.

Bond, T. and Fox, C.M. 2007. Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, NJ: Lawrence Erlbaum Associates.

Boone, W.J., Townsend, J.S., and Staver, J. 2010. Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education*, 95(2):258–280.

Boyle, F. and Monarch, I.A. 1992. Studying conceptual change in learning physics. *Science Education*, 76(6):615–652.

Bozdogan, H. 2000. Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, 44(1):62–91.

Briggs, C.B., Alonzo, A.C., Schwab, C., and Wilson, M. 2006. Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, 11(1):33–63.

Charmaz, K. 2006. Constructing grounded theory: A practical guide through qualitative analysis. Washington, DC: Sage Publications Ltd.

Creswell, J. 1998. Qualitative inquiry and research design: Choosing among five traditions. Thousand Oaks, CA: Sage Publications.

Cortina, J.M. 1993. What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1):98–104.

Crocker, L., and Algina, J. 1986. Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.

de Ayala, R. J. 2009. The theory and practice of item response theory. New York: The Guilford Press.

DeVellis, R.F. 2006. Classical test theory. *Medical Care*, 44(11):S50–S59.

Ding, L., Chabay, R., Sherwood, B., and Beichner, R. 2006. Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical Review Special Topics–Physics Education Research*, 2(1):010105-1–010105-7.

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

Efron, B. 1982. The jackknife, the bootstrap, and other resampling

plans. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Driver, R. and Odham, V. 1986. A constructivist approach to curriculum development in science. *Studies in Science Education*, 13:105–122.

Embretson, S.E. and Reise, S.P. 2000. Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Feig, A.D. 2011. Methodology and location in context of qualitative data and theoretical frameworks in geoscience education research. *In* Feig, A.D. and Stokes, A., eds., Qualitative inquiry in geoscience education research. Boulder, CO: Geological Society of America Special Paper 474, p. 1–10.

Garrison, T. 2007. Oceanography: An invitation to marine science. Belmont, CA: Thomson Brooks/Cole.

Garvin-Doxas, K., Klymkowski, M., and Elrod, S. 2007. Building, using, and maximizing the impact of concept inventories in the biological sciences: Report on a National Science Foundation–sponsored conference on the construction of concept inventories in the biological sciences. *CBE Life Sciences Education*, 6:277–282.

Greenfield, P.M. 1997. Culture as process: Empirical methods for cultural psychology. *In* Berry, J.W., Poortinga, Y.H., and Pandey, J., eds., Handbook of cross-cultural psychology. 2nd ed., vol. 1: Theory and method. Needham Heights, MA: Allyn & Bacon, p. 301–346.

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

Hambleton, R.K. and Jones, R.W. 1993. An NCME Instructional Module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurements: Issues and Practice*, 12(3):38–47.

Hestenes, D., Wells, M., and Swachhamer, G. 1992. Force concept inventory. *The Physics Teacher*, 30:141–158.

Kline, P. 2000. A psychometrics primer. London, UK: Free Association Books.

Klymkowski, M.W. and Garvin-Doxas, K. 2008. Recognizing student misconceptions through Ed's tools and the biology concept inventory. *PLoS Biology*, 6(1):0014–0017.

Knudson. D., Noffal. G., Bauer. J., McGinnis. P., Bird. M., Chow. J., Bahamonde, R., Blackwell, J., Strohmeyer, S., and Abendroth-Smith, J. 2003. Development and evaluation of a biomechanics concept inventory. *Sports Biomechanics*, 2(2):267–277.

Lederman, N.G., Abd-El-Khalick, F., Bell, R.L., and Schwartz, R. 2002. Views of nature of science questionnaire: Toward valid and meaningful assessment of learner's conceptions of nature of science. *Journal of Research in Science Teaching*, 39(6):497–521.

Lewis, E.B. and Baker, D.R. 2010. A call for a new geoscience education research agenda. *Journal of Research in Science Teaching*, 47(2):121–129.

Libarkin, J. 2008. Concept inventories in higher education science. Paper presented at the National Research Council Workshop 2: Promising Practices in Undergraduate STEM Education. Washington, DC, 13–14 October.

Libarkin, J.C. and Anderson, S.W. 2006. The geoscience concept inventory: Application of Rasch analysis to concept inventory development in higher education. *In* Liu, X. and Boone, W.J., eds., Applications of Rasch measurement in science education. Maple Grove, MN: JAM Press.

Lindell, R.S. 2005. Measuring conceptual change in college students' understanding of lunar phase. *American Institute of Physics Conference Proceedings*, 790(1):53–56.

Lindell, R.S. and Sommer, S.R. 2004. Using the lunar phases concept inventory to investigate college students' pre-instructional mental models of lunar phases. *American Institute of Physics Conference Proceedings*, 720(1):73–76.

Lord, F.M. 1980. Applications of item response theory to practical testing problems. Hillside, NJ: Erlbaum.

Lord, F.M. and Novick, M.R. 1968. Statistical theories of mental test scores. Charlotte, NC: Addison-Wesley Publishing Company, Inc.

Madsen, D. 2004. Stability coefficient. *In* Lewis-Beck, M.S., Bryman, A., and Liao, T.F. eds., The SAGE encyclopedia of social science research methods. Available at http://dx.doi.org/10.4135/9781412950589 (accessed 15 July 2015).

Marbach-Ad, G., Briken, V., El-Sayed, N.M., Frauwirth, K., Fredericksen, B., Hutcheson, S., Gao, L.-Y., Joseph, S., Lee, V.T., McIver, K.S., Mosser, D., Quimby, B.B., Shields, P., Song, W., Stein, D.C., Yuan, R.T., and Smith, A.C. 2009. Assessing student understanding of host pathogen interactions using a concept inventory. *Journal of Microbiology and Biology Education*, 10:43–50.

National Council on Measurement in Education and American Council on Education. 2006. Brennan, R.L., ed., Educational measurement. Westport, CT: Praeger Publishers.

National Geographic Society's Oceans for Life Initiative and National Oceanic and Atmospheric Administration. 2007. Ocean literacy: The essential principles of ocean sciences, Washington, DC: National Geographic Society.

National Research Council. 2000. How people learn: Brain, mind, experience, and school. Bransford, J.D., Brown, A.L., and Cocking, R.R., eds. Washington, DC: National Academy Press.

Nunnally, J.C. and Bernstein, I.H. 1994. Psychometric theory, 3rd ed. New York: McGraw-Hill.

O'Connor, E.F., Jr., 1972. Extending classical test theory to the measurement of change. *Review of Educational Research*, 42(1):73–97.

Odom, A.L. and Barrow, L.H. 1995. Development and application of a two–tier diagnostic test measuring college biology students' understanding of diffusion and osmosis after a course of instruction. *Journal of Research in Science Teaching*, 32:45–61.

Parham Jr., T.L., Cervato, C., Gallus, W.A., Jr., Larsen, M., Hobbs, J., Stelling, P., Greenbowe, T., Gupta, T., Knox, J.A., and Gill, T.E. 2010. The InVEST volcanic concept survey: Exploring student understanding about volcanoes. *Journal of Geoscience Education*, 58(3):177–187.

Pathak, A., Patro, K., Pathak, K., and Valecha, M. 2013. Item response theory. *International Journal of Computer Science and Mobile Computing*, 2(11):7–11.

Pearsall, N.R., Skipper, J.E.J., and Mintzes, J.J. 1997. Knowledge restructuring in the life sciences: A longitudinal study of conceptual change in biology. *Science Education*, 81(2):193–215.

Petcovic, H.L. and Ruhf, R.J. 2008. Geoscience conceptual knowledge of preservice elementary teachers: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education*, 56(3):251–260.

Sadowski, C.J. and Gulgoz, S. 1992. Internal consistency and test-retest reliability of the need for cognition scale. *Perceptual and Motor Skills*, 74(2):610–610.

Savinainen, A. and Scott, P. 2002. The force concept inventory: A tool for monitoring student learning. *Physics Education* 37(1):45–52.

Savinainen, A., Scott, P., and Viiri, J. 2005. Using a bridging representation and social interactions to foster conceptual change: Designing and evaluating an instructional sequence for Newton's third law. *Science Education*, 89(2):175–195.

Smith, K.S., Wood, W.B., and Knight, J.K. 2008. The genetics concept assessment: A new concept inventory for gauging student understanding of genetics. *CBE Life Sciences Education*, 7:422–430.

Tan, K.C.D., Taber. K.S., Xiufeng. L., Coli. R.K., Lorenzo. M., Jia, L., Goh, N.K., and Chia, L.S. 2008. Students' conceptions of ionization energy: A cross-cultural study. *International Journal of Science Education*, 30(2):265–285.

Thijs, G.D. 1992. Evaluation of an introductory course on force considering students' preconceptions. *Science Education*, 76(2):155–174.

Thomson, G. and Douglass, J.A. 2009. Decoding learning gains: Measuring outcomes and the pivotal role the major and student backgrounds. Berkeley, CA: University of California at Berkeley's Center for Studies in Higher Education Research & Occasional Paper Series: CSHE.5.09.

Vosniadou, S. 2007. Conceptual change and education. *Human Development*, 50:47–54.

Webb, N.M., Shavelson, R.J., and Haertel, E.H. 2006. Reliability coefficients and generalizability theory. *In* Rao, C.R., ed., Handbook of statistics, Vol. 26 on Psychometrics. Amsterdam, Netherlands: Elsevier, p. 81–124.

Wright, B.D. and Stone, M. H. 1979. Best test design. Chicago, IL: MESA Press.