Development and Validation of the Homeostasis Concept Inventory

Jenny L. McFarland,[†‡] Rebecca M. Price,[†§*] Mary Pat Wenderoth,[†∥] Patrícia Martinková,[†¶] William Cliff,[#] Joel Michael,[@] Harold Modell,[**] and Ann Wright[††‡‡]

[†]Biology Department, Edmonds Community College, Lynnwood, WA 98036; [§]School of Interdisciplinary Arts and Sciences, University of Washington, Bothell, Bothell, WA 98011; [∥]Department of Biology, University of Washington, Seattle, WA 98195; [¶]Institute of Computer Science, Czech Academy of Sciences, 18207 Prague, Czech Republic; [#]Department of Biology, Niagara University, Niagara, NY 14109; [@]Department of Molecular Biophysics and Physiology, Rush Medical College, Chicago, IL 60612; **Physiology Educational Research Consortium, Seattle, WA 98115; [††]Department of Biology, Canisius College, Buffalo, NY 14208

ABSTRACT

We present the Homeostasis Concept Inventory (HCI), a 20-item multiple-choice instrument that assesses how well undergraduates understand this critical physiological concept. We used an iterative process to develop a set of questions based on elements in the Homeostasis Concept Framework. This process involved faculty experts and undergraduate students from associate's colleges, primarily undergraduate institutions, regional and research-intensive universities, and professional schools. Statistical results provided strong evidence for the validity and reliability of the HCI. We found that graduate students performed better than undergraduates, biology majors performed better than nonmajors, and students performed better after receiving instruction about homeostasis. We used differential item analysis to assess whether students from different genders, races/ethnicities, and English language status performed differently on individual items of the HCI. We found no evidence of differential item functioning, suggesting that the items do not incorporate cultural or gender biases that would impact students' performance on the test. Instructors can use the HCI to guide their teaching and student learning of homeostasis, a core concept of physiology.

INTRODUCTION

Traditional biology education has long been criticized for emphasizing memorization of facts and terminology, particularly in the face of mounting evidence that students benefit from and require a more transferable and enduring educational experience (Valverde and Schmidt, 1997; National Research Council, 2000; Zheng *et al.*, 2008). A curriculum that focuses on facts does not prepare students fully for life science careers in which they must rely on deep conceptual understanding and strong scientific reasoning skills to solve problems and adapt to the rapid changes in their fields (National Research Council, 2009). On the other hand, undergraduate biology textbooks are forever increasing in length, incorporating more factual knowledge (e.g., Michael *et al.*, 2009). For example, physiology textbooks now serve as encyclopedic references rather than guides to instruction. It is therefore challenging to move students from simple rote memorization of material to deep and meaningful learning (Michael, 2001; Michael and Modell, 2003; Knight and Wood, 2005; Momsen *et al.*, 2010).

Identifying the core concepts of a discipline is one way to help focus breadth of coverage to allow for more depth (American Association for the Advancement of Science [AAAS], 2011; National Research Council, 2012). Instructors can organize their courses around core concepts, directing student attention to phenomena that recur in a discipline. In physiology, for example, homeostatic regulation of blood pressure and

Peggy Brickman, *Monitoring Editor*

Submitted October 24, 2016; Revised February 17, 2017; Accepted February 27, 2017

CBE Life Sci Educ June 1, 2017 16:ar35

DOI:10.1187/cbe.16-10-0305

[†]These authors contributed equally to the work.

[‡‡]Deceased.

*Address correspondence to: Rebecca M. Price (beccap@uw.edu).

core body temperature can both be explained by the same concept of a control system (Modell, 2000; Modell *et al.*, 2015). Because students can more readily transfer conceptual understanding across domains (National Research Council, 2000), understanding of core concepts can be used to master new learning in subsequent courses and throughout a student's professional career (Michael *et al.*, 2009).

Homeostasis has been identified as one of the eight core concepts in biology (Michael, 2007). When asked to rank the core principles or "big ideas" in their field, more than 100 physiologists from associate's colleges to research-intensive institutions ranked "homeostasis" and "cell membranes" as the two most important principles for students (Michael and McFarland, 2011; note that associate's colleges is the current categorization of community colleges under the Carnegie classification; Carnegie Classification of Institutions of Higher Education, n.d.). Homeostasis is also one of the core competencies listed in the *Scientific Foundations for Future Physicians* report (M1: the ability to apply knowledge about homeostasis; Association of American Medical Colleges and Howard Hughes Medical Institutes, 2009), and homeostasis is included within the core concept of "systems" in *Vision and Change* (AAAS, 2011).

The concept of homeostasis was first defined by Claude Bernard in 1865 as the ability of a complex organism to maintain its *milieu interieur*, translated as "internal environment," in a fairly steady state in the face of external challenges. Walter Cannon coined the term "homeostasis" to describe this concept in 1926 (Cooper, 2008; Modell *et al.*, 2015). Some argue that "homeostasis is *the* central idea in physiology" (Michael *et al.*, 2009: 13, emphasis in original). Homeostatic mechanisms keep a regulated variable (i.e., a physiological variable that the organism can sense) within a range of values conducive to supporting the life of the organism. To accomplish the task of maintaining a stable internal environment, an elegant interacting system of sensors, integrators with set points, and effectors (or targets) has evolved. Homeostatic mechanisms orchestrate the moment-to-moment responses of an organism to the wide array of its interactions with the world. This orchestration includes small-scale processes such as maintaining human blood pressure while moving from a sitting to a standing position to large-scale responses associated with the cardiovascular response during a major hemorrhage.

What physiologists perceive as elegant looks complex and intricate to students. Students are unsure of which internal environments are maintained, what set points are, whether homeostatic mechanisms are just on/off switches, and what physiological variables are homeostatically regulated. These confusions are just some of the struggles that students have about homeostasis (Modell *et al.*, 2015).

To help address these challenges, our project team has developed a powerful set of tools for teaching homeostasis. We have documented a number of misconceptions (i.e., scientifically inaccurate or incomplete understandings, as in Crowther and Price, 2014; Leonard *et al.*, 2014) regarding homeostasis (Wright *et al.*, 2013, 2015). We have created a simple, yet accurate diagram of the concept of homeostasis that we recommend textbook authors and instructors use to help students visualize this core concept (Modell *et al.*, 2015). We distilled a common vocabulary of terms from 12 undergraduate physiology textbooks to better reflect the way experts communicate on this topic (Modell *et al.*, 2015). We have also developed the Homeostasis Conceptual Framework (HCF), validated with physiology faculty from a broad range of institutions, which describes the critical components and constituent ideas important for undergraduates to understand if they are to correctly apply the concept of homeostasis (McFarland *et al.*, 2016). This suite of tools empowers and guides instructors to help undergraduate students build appropriate mental models of homeostatic regulation in physiological systems.

This paper describes the development and validation of the Homeostasis Concept Inventory (HCI), the next piece of our project. The HCI is a multiple-choice instrument that will allow biology educators to determine how well their teaching has helped their students master the core concept of homeostasis. It can also serve as a diagnostic tool to identify misconceptions concerning homeostasis.

METHODS

We used the HCF (McFarland *et al.*, 2016) as the basis for the HCI. The physiologists in the project team (J.L.M., M.P., W.C., J.M., H.M., A.W.) drafted questions and then embarked on an iterative process of revising them in consultation with a community of physiology faculty and students from geographically and institutionally diverse institutions across the United States, including associate's colleges, primarily undergraduate institutions, regional comprehensive universities, research-intensive universities, and professional schools. Through this process, we wrote three drafts of the HCI before reaching the final version. The drafts are named HCI-Drafts 1, 2, and 3, following the conventions of Price *et al.* (2014) and Newman *et al.* (2016); we acknowledge that the final version presented in this paper will change as understanding of teaching and student learning of homeostasis progresses. Once we compiled the HCI (the version presented in this paper), we used an extensive suite of statistical analyses to find evidence for valid and reliable scores from the test as a whole and to determine whether bias existed for any individual items (as in Martinková *et al.*, 2017b). An overview of the process used to develop the HCI is in Table 1.

HCI-Draft 1

We wrote multiple-choice questions to address the critical components and constituent ideas identified in the HCF (McFarland *et al.*, 2016). In addition to choosing the multiple-choice format, we intentionally constructed a short instrument that students would be able to complete in a reasonable period of time, with a target of approximately 20 minutes. We knew that a short-enough instrument would not be able to assess all of the constituent ideas within each critical component of the HCF, but we thought this trade-off would make the HCI more likely for faculty to use and for students to complete.

We initially wrote at least two multiple-choice questions for each of the five critical components identified in the HCF, with each question having only one correct answer. One question was written as an abstract or theoretical formulation of the idea, in which variables were identified as x or y, while the second question applied to a real-world situation (see Supplemental Table 1). We were not sure whether one type of question would be more challenging than the other, but we anticipated that students would be exposed to both types of questions in their courses.

TABLE 1. Overview of the methods used to generate the HCI[a]

HCI-Draft 1 (19 multiple-choice questions and 1 open-ended)
a. Conducted think-aloud interviews with six students at two BCAS to assess readability, interpretation, and misconceptions
b. Revised to HCI-Draft 1.1
c. Conducted think-aloud interviews with 11 students at an AC on HCI-Draft 1.1
d. Revise to HCI-Draft 2
HCI-Draft 2 (18 multiple-choice questions)
a. Taken by 16 students at an AC
b. Revised to HCI-Draft 2.1
c. HCI-Draft 2.1 distributed in an online survey of 20 physiology faculty members to evaluate accuracy of questions, to assess questions' relevance, and to edit items
d. Revised to HCI-Draft 3
HCI-Draft 3 (20 multiple-choice questions)
a. Taken by 427 students at five institutions (BCAS, AC, MCU, R1, and professional school)
b. Analyzed difficulty and discrimination; no questions were too easy; retained two challenging questions with little ability to discriminate because they tested ideas critical to the concept of homeostasis
c. Conducted think-aloud interviews with seven students at two BCAS to assess readability and interpretation
d. Distributed to faculty at the 2014 Human Anatomy and Physiology Society Conference to evaluate questions
e. Revised two questions by changing the context from blood sodium to blood glucose homeostasis
f. Distributed an online survey to faculty to evaluate the two revised questions; faculty confirmed the questions were accurate.
g. Revised to the HCI
HCI (20 multiple-choice questions)
Steps are enumerated extensively in Tables 2, 3, and 4.

[a]AC, associate's colleges; BCAS, baccalaureate colleges: arts and sciences focus; MCU, master's colleges and universities; R1, doctoral universities–highest research activity.

The questions that are situated in the real world may sometimes be advantageous but at other times disadvantageous in helping students reason about homeostasis or other core concepts. McNeil *et al.* (2009) found that teaching concepts by using examples with concrete objects, such as money, could both help and hurt elementary school students solve math problems. In physics, there is evidence that teaching with concrete representations may be beneficial for students when they address simple problems, but the abstract representations may give students an advantage with more complex problems. Abstract, generic questions contain only relevant relations with minimal information and are therefore not burdened by information-rich, specific details of a physiological example (Kaminski *et al.*, 2013). Perhaps because of this lack of detail, Heckler (2010) reported that students with higher course grades performed better on abstract problems. He also postulated that, when students' prior knowledge disagrees with scientific understanding, questions with a concrete context may trigger application of inaccurate mental models.

Most of the questions in the HCI have four or five answer choices, and the distractors are based on common misconceptions that students hold (Wright *et al.*, 2013, 2015; Modell *et al.*, 2015). However, questions concerning how the concentration of a molecule would change in response to a perturbation to the system necessarily had only three answer choices (increase, decrease, remain constant). We retained all three choices to maintain symmetry, even when students chose one option infrequently.

The end result of this procedure was HCI-Draft 1, composed of 19 multiple-choice questions and 1 open-ended question. We conducted think-aloud interviews with six students from two primarily undergraduate institutions to assess readability, to confirm that students were interpreting the questions as they were intended, and to gain greater insight as to the actual struggles students would have with the concepts we were testing (Pollitt *et al.*, 2008). We learned in these interviews that several students felt that something that was "more or less constant" fluctuated more and was less regulated than something that was "relatively constant." That discovery led us to use the phrase "relatively constant" instead.

We used this feedback to develop a 24-question HCI-Draft 1.1. We used HCI Draft-2.1 to conduct think-aloud interviews, this time with 11 students from two different associate's colleges. The changes based on this feedback led to HCI-Draft 2, an 18-question instrument.

HCI-Draft 2

We administered HCI-Draft 2 to 16 students in an associate's college course on human anatomy and another associate's college course on human anatomy and physiology. We used their responses to remove distractors that were not being chosen. We also removed some questions, added others, and incorporated revisions. The resulting HCI-Draft 2.1 had 19 questions that we distributed through an online survey to 20 faculty experts with whom we consulted regularly throughout this project. These faculty experts teach physiology at research-intensive, regional comprehensive, and primarily undergraduate universities; associate's colleges; and professional schools. The faculty experts were asked to evaluate the accuracy of the questions, to assess each question's importance to their teaching of undergraduate physiology, and to suggest edits. We incorporated this feedback into HCI-Draft 3, a 20-question instrument.

HCI-Draft 3

We administered the HCI-Draft 3 to 427 students at five institutions from different Carnegie classifications (Carnegie

Classification of Institutions of Higher Education, n.d.), and we analyzed the resulting data to determine the difficulty and discrimination of each item (Allen and Yen, 1979), checking to ensure that the instrument captured a range of scores and range of student academic abilities. Two particularly challenging questions (items 7 and 17 in the HCI) did not discriminate well, but we decided to retain them, because they assess concepts that are essential for students to learn. Item 7 reveals the common misconception that the nervous system is always involved in homeostatic regulation, and we wanted faculty to be able to assess the prevalence of this misconception among their students. Item 17 requires students to know how the different components interact in the system that regulates blood pressure. We conducted an additional seven think-aloud interviews with students from two different primarily undergraduate institutions to assess readability and to confirm that students were interpreting the questions on the HCI-Draft 3 as intended.

We distributed HCI-Draft 3 to faculty at a workshop at the 2014 Human Anatomy and Physiology Society Annual Conference. The workshop participants raised concerns about two questions (the precursors to items 11 and 20 on the HCI) that we subsequently addressed by changing the context from blood sodium to blood glucose homeostasis. We asked another 20 faculty members from our group of physiology experts to evaluate the accuracy of these two revised questions, and these faculty members agreed the questions were accurate. The resulting instrument is the HCI, a 20-question multiple-choice instrument.

Homeostasis Concept Inventory

To assess the validity of HCI scores, we recruited a sample of 669 undergraduates from 12 institutions (Tables 2 and 3). These students were enrolled in courses for life science majors, mixed majors, or allied health majors; each course covered homeostasis to some extent. All of the students took the HCI within the last 2 weeks of their courses. Most instructors gave students extra credit for good-faith efforts to complete the HCI. We excluded students who completed the HCI in less than 4 minutes; because none of these respondents scored more than 10 points, we concluded that they rushed through the HCI without considering the questions seriously.

We used a series of smaller student samples (Table 2) to conduct additional statistical tests, including analysis of test–retest and pretest–posttest relationships and to compare graduate students with undergraduates. Graduate students in professional schools served as our upper limit of performance on the HCI, as we determined that the learning goal for our undergraduate curriculum would be to prepare students for professional school.

Our extensive suite of statistical analyses assessed the validity and reliability of the total scores from the HCI (Table 4). We also conducted item-level analyses to relate student ability to each item and to assess whether any of the items are biased. We included both classical test theory analyses and item-response theory (IRT) models to investigate item and test properties (as did Neumann *et al.*, 2011; Jorion *et al.*, 2015; Kalinowski *et al.*, 2016). We have also included structural analyses to test the unidimensionality of the instrument. The analyses were completed in R (R Core Team, 2016; Supplemental Material, R Code) with the libraries ggplot (Wickham, 2009), lme4 (Bates *et al.*, 2015), lmerTest (Kuznetsova *et al.*, 2016), psychometric (Fletcher, 2010), psych (Revelle, 2015), corrplot (Wei and Simko, 2010), ltm (Rizopoulos, 2006), mirt (Chalmers, 2012), WrightMap (Torres Irribarra and Freund, 2014), difNLR

TABLE 2. Types of institutions used in the validation of the HCI

Type[a]	Region	No. of students
Main sample ($N = 669$)		
AC	NW	47
	NW	34
	SE	20
	SW	98
BCAS	NE	48
	SE	38
	MW	16
	MW	21
MCU	SW	68
	SW	76
R1	SW	95
	SE	108
Test–retest ($N = 45$)		
Professional school	NW	45
Graduate student performance ($N = 10$)		
R1	MW	10
Pre/posttesting ($N = 16$)		
AC	NW	16

[a]See Table 1 for definitions of abbreviations.

TABLE 3. Demographic characteristics of students who participated in the large-scale testing of the HCI (main sample of 669, Table 2)

	Category	Count	Percent
Gender	F	405	61
	M	246	37
	NA	18	3
Age (years)	≤24	494	74
	25–29	106	16
	≥30	69	10
Planning to major in the life sciences	No	270	40
	Yes	399	60
Planning to attend professional school	No	190	28
	Yes	479	72
Race/ethnicity	Asian	117	17
	Black	39	6
	Hispanic	85	13
	White	343	51
	Mixed and other	54	8
	Undisclosed	31	5
English as first language	Yes	521	78
	No	148	22
Year in college	Freshman	67	10
	Sophomore	137	20
	Junior	171	26
	Senior	216	32
	Postbaccalaureate	78	12

TABLE 4. The statistical methods used to gather evidence for the validity of the HCI scores

Method	Analytical question
Validity	
Two-sample *t* test	Do graduate students in the life sciences perform better on the HCI than undergraduates?
Pre/posttesting (*t* tests)	Do students perform better on the HCI after receiving instruction about homeostasis? Is this improvement bigger than the improvement of students who did not receive any instruction about homeostasis?
Mixed-effects linear regression	Do students pursuing majors in the life sciences perform better on the HCI than students pursuing other majors? Is this difference significant when controlling for other variables such as gender, ethnicity, institution, and course?
Density plots	Does a range of total scores on the HCI exist for different demographic groups? Can we see a visual difference among the demographic groups? For example, do students pursuing majors in the life sciences perform better on the HCI than students pursuing other majors? Do students from R1 institutions perform better than students from other types of institutions?
Tetrachoric correlation (heat map)	**Do items correlate with each other? Do clusters of items form around similar topics?**
Exploratory factor analysis	**Is the HCI unidimensional?**
Reliability	
Test–retest (Pearson correlation)	Is student performance on the HCI repeatable?
Cronbach's alpha	Is the test internally consistent?
Test item function (TIF)	How reliable is the HCI is for students with different levels of ability?
Item-level analysis	
Estimating item difficulty	Does the HCI have a range of difficulties, as indicated by the percentage of students answering each item correctly?
Estimating item discrimination	Do strong students perform better on harder questions?
Item-person (Wright) map	Does the inventory capture the whole population of students? Do item difficulties correspond to student abilities?
Item characteristic curves	Do items have a range of difficulties, and do they have sufficient discrimination?
Item information function	For which latent abilities do individual items provide the highest information?
DIF analysis	Are the HCI items biased with respect to gender, ethnicity, and English language status?
Abstract and applied questions	
Paired *t* test	Is student performance on abstract questions the same as student performance on applied questions?

(Drabinova *et al.*, 2016), difR (Magis *et al.*, 2015), and ShinyItemAnalysis (Martinková *et al.*, 2017a). Students' names and identification numbers were removed from all data sets before statistical analyses.

Validity of Total Scores. We used a two-sample *t* test to determine whether graduate students in the life sciences scored higher on the HCI than the undergraduates in our main sample of 669 students (Table 3). We also conducted pre/posttesting on a sample of 16 students enrolled in a physiology course that emphasized homeostasis to determine whether students performed better on the HCI after receiving instruction about homeostasis (paired *t* test) and to determine whether this improvement exceeds the improvements of 45 students enrolled in a master's program in nutrition who did not receive explicit instruction about homeostasis during the course that we sampled (two-sample *t* test; this sample of master's students who did not receive instruction about homeostasis was also the same sample used in the test–retest analysis of reliability; see *Reliability*).

We used a mixed-effects linear regression model that accounted for correlated responses of students within classes to determine which demographic variables could be used to predict total score performance on the HCI. In particular, we wanted to know whether students planning to major in the life sciences performed better than students in other majors, because this would indicate that the HCI could discriminate between students who had extensive exposure to homeostasis and those who had more limited exposure. These models were built with our main sample of 669 students. We used Bayesian information criteria (BIC) to select the optimal model (Schwarz, 1978). We used density plots (Hastie *et al.*, 2009) to visualize potential differences among groups of students. Density plots estimate the distribution of total scores as a smooth curve instead of plotting the frequency of exact scores as in histograms. For example, we graphed the range of scores found among students from different types of institutions and enrolled in different kinds of courses.

Reliability. We determined the reliability of the HCI by calculating internal consistency and test–retest reliability with classical test theory. We also present item and test information functions based on IRT models. To determine whether the test was internally consistent, we calculated Cronbach's alpha (Cronbach, 1951) of the main sample of 669 students. We also estimated test–retest reliability, that is, whether the same population of students would have the same performance on the test multiple times (Nunnally and Bernstein, 1994), by calculating the Pearson correlation coefficient and conducting a linear regression on a sample of 45 students enrolled in a nutrition course in a master's program who had no explicit instruction in homeostasis between the first (test) and second (retest) time of taking the HCI. We used IRT models (described in *Item Analysis*) to enumerate item and test information functions (Samejima, 1994). Test information function (TIF) provides an

estimate of reliability that depends on student ability (here, student ability is latent ability that is estimated with an IRT model; see *Item Analysis*).

Item Analysis. In addition to exploring how students performed on the HCI as a whole, we evaluated how they performed on individual items. To assess difficulty, we calculated the percentage of correct responses for each item. For discrimination, we calculated the difference in the percent of correct responses between the upper and lower third of students to assess item discrimination (Allen and Yen, 1979).

To explore how each item performed in more depth, we fitted IRT models to our data (De Ayala, 2008). First, we used the simplest, one-parameter logistic IRT model to generate an *item-person map* (also called a Wright map; e.g., Neumann *et al.*, 2011; Boone, 2016), which compares a histogram of the students' latent ability with the item difficulty. Here, latent ability is an individual's true knowledge—something that a test can only estimate—and difficulty is defined as the ability at which a student has a 50% probability of answering the item correctly.

In addition, we also fitted more complex IRT models that allowed us to explore each item with respect to difficulty and discrimination (two-parameter model) and difficulty, discrimination, and pseudo-guessing (three-parameter model; e.g., Kalinowski *et al.*, 2016). To select the best-fitting model, we used the likelihood ratio test; in this case, the three-parameter model outperformed the one- and two-parameter models. We then used the best-fitting three-parameter logistic IRT model (Livingston, 2006) to plot item characteristic curves and item information functions and to estimate the TIF to assess reliability (see *Reliability*). For each item, fit indices were calculated using the S-X2 statistic (Orlando and Thissen, 2000; Ames and Penfield, 2015) to measure how well an item fits with the estimated IRT model.

Structural Analyses. We performed two different structural analyses. First, we analyzed the correlation structure to explore the relationships among items in an instrument (Jorion *et al.*, 2015). To do this, we used tetrachoric correlations representing dependencies between pairs of items that are scored discretely as either right or wrong. Second, we used exploratory factor analysis to explore the unidimensionality of the HCI. We fitted factor analysis models with one to eight factors and used BIC (Schwarz, 1978) to determine which factor structure had the had the lowest BIC and was therefore optimal. In addition, we checked the model fit with the root-mean-square error of approximation; typically, a value of 0.06 or less indicates that the model has an acceptable fit (Hu and Bentler, 1999).

Checking for Potential Bias. Test items are intended to assess student's understanding of one topic. If additional knowledge unrelated to the content being tested is necessary to answer an item correctly, then the item is potentially unfair, biasing against certain populations (Martinková *et al.*, 2017b). We assessed the items in the HCI to determine whether they were biased with respect to gender, ethnicity, and English language status. We used differential item functioning (DIF) analysis to compare the performance of students from different groups with the same ability on different items (Martinková *et al.*, 2017b). More specifically, we used logistic regression (Zumbo, 1999) to determine whether items performed differently between: men and women; native English speakers and English language learners; and students of different race and ethnicity (six categories: Asian, Black, Hispanic, white, mixed and other, undisclosed). Because we were conducting multiple comparisons (20 for each demographic pairing, because there are 20 questions in the HCI), we used a Benjamini-Hochberg adjusted p value correction to account for multiple comparisons when detecting significance (Benjamini and Hochberg, 1995).

Abstract and Applied Questions. We used a two-sample t test to determine whether students performed differently on the subset of questions that assessed students' understanding of abstract questions and their performance on questions applied to real-world scenarios.

Human Subjects Approval

All procedures were conducted in accordance with approval from the Institutional Review Board at Edmonds Community College (IRB2014-1031).

RESULTS

The HCI (Supplemental Material, Homeostasis Concept Inventory) is a concept inventory with 20 multiple-choice questions that most students can complete within 20 minutes. The results presented below demonstrate that the HCI has been validated with and found reliable for a large group of undergraduates who were diverse with respect to gender, race/ethnicity, English language status, and institution type. The items are of intermediate difficulty and discriminate between high- and low-performing students, and we found no evidence for bias with respect to gender, race/ethnicity, or English language status.

Validity

We compared the scores of the undergraduate students with those of a group of 10 graduate students at a professional school who were studying in a field that required an understanding of the concept of homeostasis. As expected, the mean score of the graduate students (14.50, SD 3.27) was significantly higher than the mean total score of the sample of undergraduate students (12.13, SD 3.65; two-sample t test, $p = 0.024$; Figure 1A). In a pretest–posttest comparison, a group of 16 undergraduate students improved significantly in HCI total score after studying homeostasis (paired t test, $p = 0.010$; Figure 1B). Despite the low sample size, a significant mean improvement of 2.31 points (SD 3.16) was observed between the pretest and posttest. Moreover, this improvement was significantly higher (two-sample t test, $p = 0.048$) than the mean change of 0.82 (SD 2.22) observed in a group of 45 students who were not explicitly taught the concept of homeostasis (Figure 1C). (The group of 45 students who were naïve to homeostasis was also used in the test–retest to calculate reliability.)

The mixed-effects linear regression model indicates that a student's major, year in college, gender, race/ethnicity, and English language status all affect total score on the HCI (Table 5). For example, the HCI assesses understanding of homeostasis for students who are pursuing majors in the life sciences and for students pursuing other majors (Figure 2A), although life science majors tend to show the best performance (Table 5). Keeping all other variables equal, students

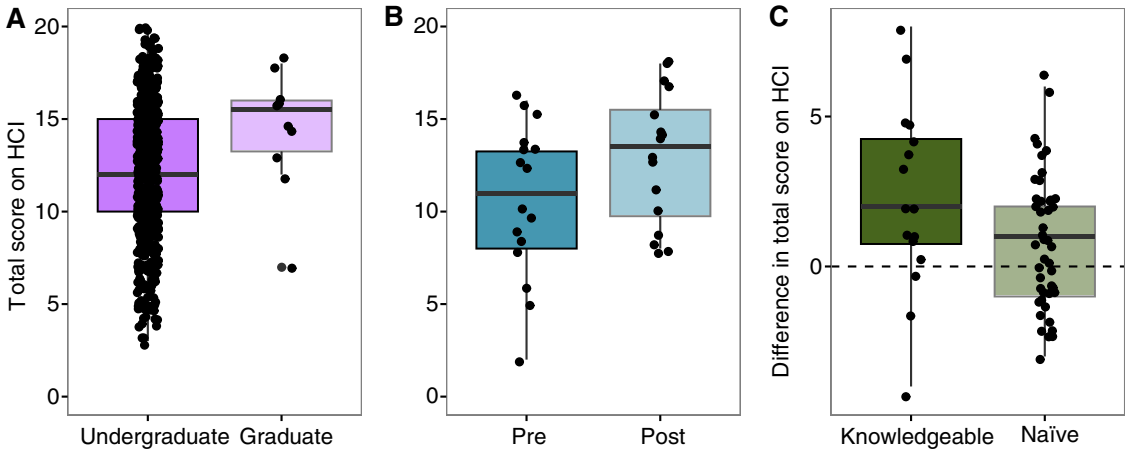

FIGURE 1. Students with different levels of experience perform on the HCI as expected. The horizontal midline in box plots is the median, and the top and bottom of each box represent one quartile from the median. Data beyond the end of whiskers are outliers. (A) Graduate students with more exposure to homeostasis perform better than undergraduates (two-sample *t* test, $p = 0.024$). (B) Undergraduates perform better on the HCI after receiving instruction (paired *t* test, $p = 0.010$). (C) Undergraduates who received instruction about homeostasis had higher gains (measured as the difference in pre–post scores from the sample in B) than master's students from a professional school studying an unrelated life science field who were naïve to the concept (two-sample *t* test, $p = 0.048$). Sample sizes for each comparison are described in Table 2.

who indicated they were pursuing life science majors scored on average 1 point higher than non–science majors; this difference was significant ($p < 0.001$; Table 5). On the other hand, in the final mixed-effects model, the intended audience of the course—whether it was for life science majors, allied health students, or nonmajors—does not predict HCI performance; students enrolled in courses intended for life sciences majors perform slightly better than students enrolled in courses for allied health majors or nonmajors (Figure 2B), but in the model, this difference is captured by effect of the student major and by variability between the courses. Similarly, accounting for the variability between the individual courses, institution type also does not affect performance on the HCI significantly, despite some trends that are observable in the density plot (Figure 2C).

More evidence of the validity of the HCI is the fact that postbaccalaureate students performed significantly higher than freshmen (mean difference 2.29, SD 0.63, $p < 0.001$). However, women performed significantly lower than men (mean difference of 0.77, SD 0.26, p 0.003), Hispanic and Black students performed significantly lower than white students (mean difference of 1.38, SD 0.41 between Hispanic and white, mean difference of 2.28, SD 0.52 between Black and white, $p < 0.001$ for both), and English language learners performed lower (mean difference of 1.53, SD 0.34, $p < 0.001$). This result inspired subsequent analysis to determine whether the HCI is biased with respect to gender, ethnicity, and English language status.

Reliability

Test–retest reliability was assessed for a sample of 45 master's students enrolled in a course that did not teach homeostasis. The Pearson correlation coefficient of the test–retest was 0.77 with a 95% confidence interval of 0.62–0.87, a value considered to be satisfactory for low-stakes exams (Nunnally and Bernstein, 1994). Cronbach's alpha, used to measure internal consistency, was 0.72 with 95% confidence intervals of 0.69 and 0.75, indicating a satisfactory level of internal consistency of the HCI (Nunnally and Bernstein, 1994). We discuss

TABLE 5. Final linear mixed-effects model for total score with the demographic variables ordered in terms of how they impact interpretation of the validity of the HCI (as opposed to Table 3)[a]

Parameter	Model-based estimate
Intercept	12.32 ± 0.62***
Major pursued	
(reference category: Other)	
Life Sciences	1.01 ± 0.269***
Year	
(reference category: Freshman)	
Sophomore	1.00 ± 0.549⁺
Junior	0.01 ± 0.552
Senior	0.26 ± 0.586
Postbaccalaureate	2.29 ± 0.630***
Gender	
(reference category: Male)	
Female	−0.77 ± 0.259**
NA	−1.77 ± 0.820*
Race/ethnicity	
(reference category: White)	
Asian	−0.48 ± 0.386
Black	−2.28 ± 0.523***
Hispanic	−1.38 ± 0.411***
Mixed and other	−0.75 ± 0.461
NA	−2.31 ± 0.631***
English as first language	
(reference category: Yes)	
English as second language	−1.53 ± 0.335***

[a]The differences between school types were not significant after accounting for hierarchical structure with random effects. p values: ⁺<0.1; *<0.05; **<0.01; ***<0.0001.

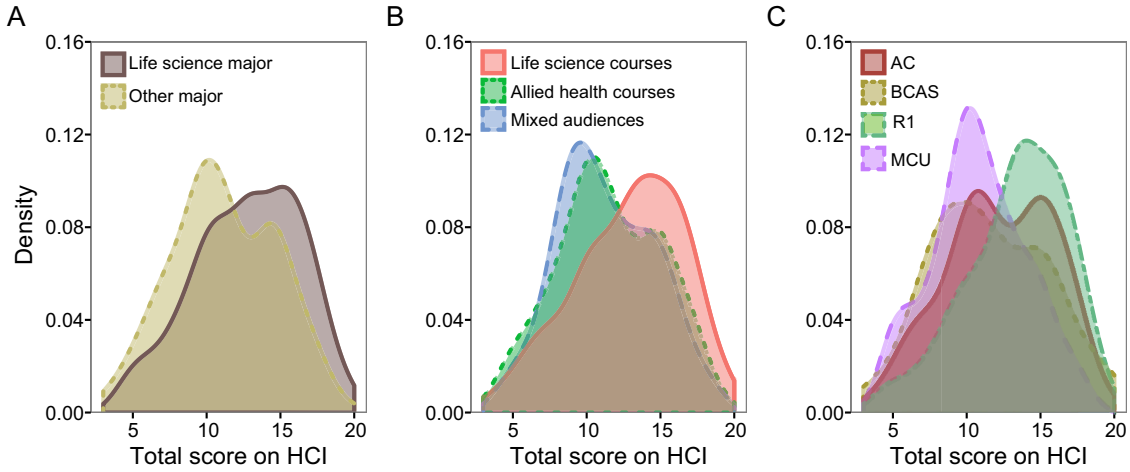

FIGURE 2. Density plots comparing total scores on the HCI for different demographic groups. Density plots are read like histograms, where density is analogous to proportion. Each graph shows a range of scores, indicating that the HCI can assess how students from each of these demographic groups understand the concept of homeostasis. (A) Student major. Students pursuing life science majors scored higher than students pursuing other majors (see Table 5; $p < 0.001$). (B) Course audience. Students enrolled in physiology courses for science majors scored higher than students in courses for allied health students or in courses for nonmajors (but in the final mixed-effects model, this difference is captured by student major and thus not significant). (C) Type of institution. The students in our sample who attended doctoral universities (highest research activity [R1]) tended to perform better than those at master's colleges and universities (MCU), and baccalaureate colleges: arts and sciences focus (BCAS). Students from associate's colleges (AC) show a bimodal distribution, and the higher mode is comparable to performance of students at R1 institutions. The final model accounts for the fact that courses are different, which encompasses the difference among institutions.

our additional analysis of reliability using TIF in the section on *Item Analysis*, as it relies on the results of the IRT analysis.

Item Analysis

The items on the HCI show a wide range of item difficulty and high values of discrimination, indicating that the HCI can be used to assess a broad spectrum of undergraduate physiology students (Figure 3). Low discrimination (slightly lower than 0.2) was observed only on the most difficult item, item 17, which tests misconception about how the control center operates. Item 7 had discrimination slightly higher than 0.2.

IRT models provided more accurate estimates of student abilities and more detailed description of items. The item-person map (Figure 4) shows that the inventory captures the whole population of students, except for a few students in the highest category of ability. However, only a small proportion of students who we tested performed at an ability that high, indicating that this instrument is appropriate for college-level understanding of homeostasis.

We used a likelihood-ratio test to determine that the three-parameter model is the best-fitting IRT model for the HCI, so parameters describing discrimination, difficulty, and guessing are all necessary. The resulting model fits three parameters to each of the 20 items in the HCI (60 parameters total; see Supplemental Table 2; the standard errors for all 60 parameters and model fit indices indicated good model fit). Although guessing might be expected to be significant for a multiple-choice instrument, the guessing parameter was large for only a few of the items (see Supplemental Table 2). Discrimination and difficulty fell within the desired range for all questions, except for item 17.

We also used a TIF to assess the reliability of the HCI. We present these results here, rather than in the *Reliability* section of the *Results*, because TIF makes use of the three-parameter IRT model. In IRT models, student ability is normalized, so

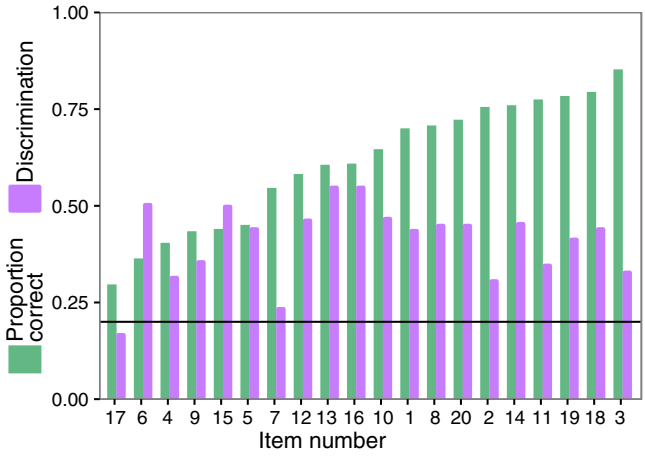

FIGURE 3. Difficulty, which is represented as the proportion of students who answered the item correctly, and discrimination for the 20 questions in the HCI. The items are arranged by percent correct, with the most frequently incorrect on the left, and the most frequently correct on the right. The horizontal line represents a discrimination of 0.2, which is usually considered the minimum discrimination for items to be included in a concept inventory (Nunnally and Bernstein, 1994). However, since item 17 tests a critical misconception about how the body responds to the complete cessation of a signal from the sensor, we felt it was essential to include it in the HCI.

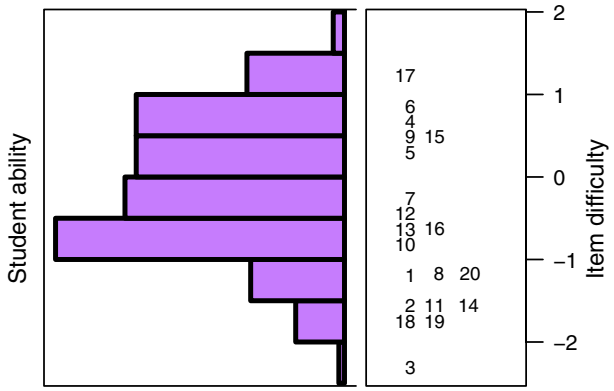

FIGURE 4. Item-person map. The left panel describes the distribution of student abilities, as determined with a one-parameter IRT model; values are arranged from the most able at the top to least able at the bottom. The items in the right panel are organized from the most difficult at the top to the least difficult at the bottom. Here, difficulty is defined as ability level, so a student of this ability has 50% probability of answering the item correctly. For the students in our sample, we found a range of item difficulties in the HCI, from item 3 (frequently correct) to item 17 (frequently incorrect).

that ability is described as SDs from the average (Jorion *et al.*, 2015). The item characteristic curve is a way to summarize the difficulty, discrimination, and guessing for each item (Figure 5A) in order to describe the information provided by individual items (Figure 5B) and the whole test (Figure 5C). For example, item 13, which assesses how students understand the role of an effector, has particularly high discrimination for students of average ability, and therefore provides highest information about these students. The location of the peak of the TIF and the spread of that peak indicate that the HCI is most reliable for students whose abilities range from –1 to 1 (Figure 5C).

Structural Analyses

The tetrachoric correlations heat map (Figure 6) showed that items 7 and 17 have low correlation with other items; they are also items that are difficult and have low discrimination. Clusters of items correspond to different concepts in the HCF (Figure 6, Supplemental Table 1). For example, three items (4, 9, and 20) correlated strongly with one another. These three items address the key concept that the body is constantly working to maintain homeostasis; a common misconception is that the body only tries to establish homeostasis after a perturbation (McFarland *et al.*, 2016). These results indicate that the items are measuring a unidimensional construct of homeostasis, although items 7 and 17 have smaller correlation coefficients.

The exploratory factor analysis supported the heat map findings; that is, that the instrument exhibited unidimensionality. Specifically, the lowest BIC was obtained from the one-factor model (Supplemental Material, R Code), and the root-mean-square error of approximation for the one-factor model was 0.04.

Lack of Bias

Although the mixed-effects linear model uncovered differences in how different demographic groups performed on the total score of the HCI, results of the DIF analysis suggest that no individual item is biased: no item functions differently for students who have the same latent ability but are from different groups (see Supplemental Table 3). There is no DIF with respect to gender, ethnicity, or English language status. These results indicate that the HCI is fair and that the difference between groups in total score is not due to any cultural biases in the wording of the questions.

Abstract and Applied Questions

The mean score of the students on the abstract questions of the HCI (0.67, SD 0.14) was not significantly different from the mean score of the related questions about real-life scenarios (0.56, SD 0.17; paired t test, $p = 0.13$; Figure 7). More items of each type are necessary to determine how abstract and applied questions relate to difficulty.

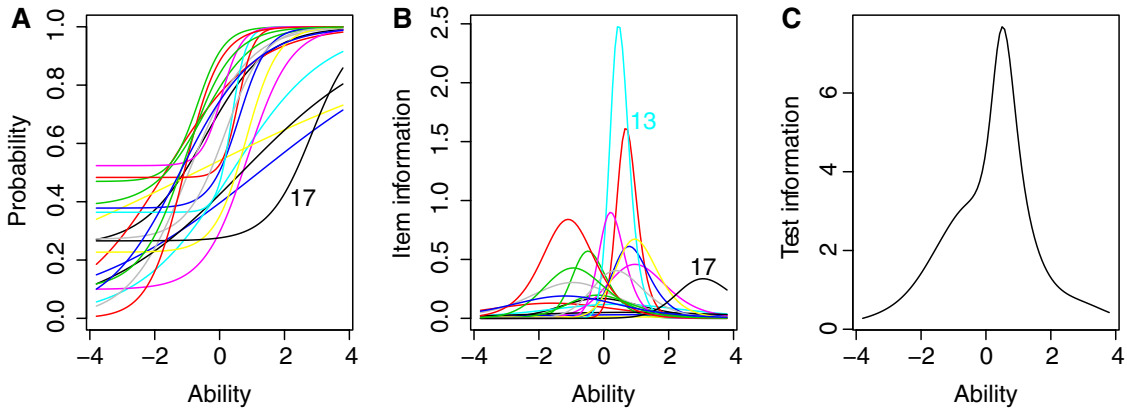

FIGURE 5. Results from the three-parameter IRT model. (A) Item characteristic curves describe the probability of an item being answered correctly by a student of a given ability. Ability is displayed as SDs from average. Item discrimination is represented by the slope; the relatively flat curve for item 17 corresponds to its low discrimination for students of low abilities. (B) Item information curves represent how well the items distinguish between strong and weak students for given ability; note that item 13 discriminates particularly well among students of average ability. (C) The TIF represents the reliability of the test for students of different ability. Based on the peak of this curve, the HCI is most reliable for students whose abilities range from –1 to 1.

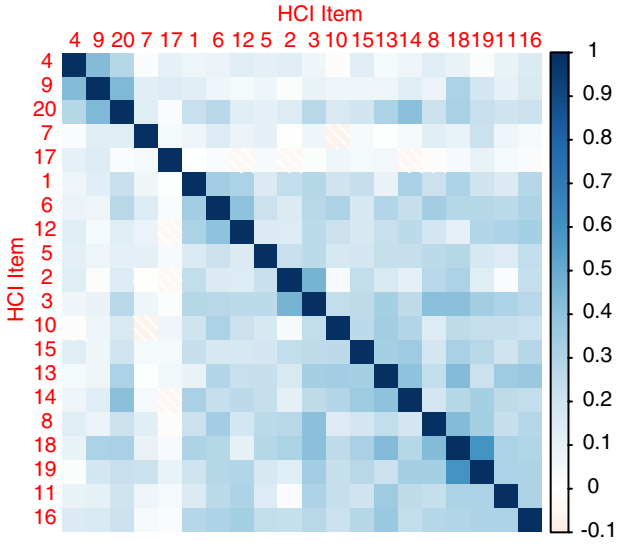

FIGURE 6. Tetrachoric correlation heat map. The items are ordered into clusters based on how correlated they are with each other. Items 4, 9, and 20 form a cluster; items 7 and 17 do not cluster with any other items; and the rest of the items cluster together.

DISCUSSION

The HCI is a 20-question multiple-choice instrument, and we present evidence showing that the total scores in our sample population are valid and reliable for assessing undergraduates' understanding of homeostasis. Questions in the HCI are based on the critical components and constituent ideas identified in the HCF, which the physiologists on the project team previously developed and validated (McFarland *et al.*, 2016). The questions on the HCI were refined through an extensive and iterative process (Table 1). The range of experts and students with whom we worked helped us minimize sampling bias and ensured that the HCI can be used to assess student understanding across institutional and geographic differences. The questions in the HCI address what faculty indicated were the most important critical components and constituent ideas in the HCF.

We have presented evidence of the validity of the HCI on the total scores of a wide range of students across the undergraduate curriculum. A TIF generated from the best-fitting IRT model indicates it is most reliable for students performing in the middle of our student sample. Therefore, the HCI can assist instructors as they help undergraduate students learn how homeostatic mechanisms operate in an integrated manner. Instructors can also use the HCI to identify common misconceptions and errors in students' understanding and to assess the effectiveness of their teaching methods. Future modifications to the HCI may include development of a few more-challenging questions so that the test can better assess the full spectrum of undergraduates.

The questions on the HCI show a range of item difficulty and reasonably high values of discrimination. Only one question, item 17, falls below the generally accepted level of 0.2 for discrimination. We retained this question in the HCI because it tests students' understanding of how the magnitude of the neural signal of the sensory input impacts the size and nature of the error message generated by the integrator. This result points out a major weakness in the students' understanding of how the control center integrates incoming sensory information. Both high- and low-performing students missed this item, indicating that it is a difficult concept for everyone in a class. Nonetheless, this constituent idea is critical for understanding how homeostatic processes work, and it is necessary to assess how well undergraduates can apply their conceptual understanding of homeostatic systems.

When drafting the first set of questions for the HCI, our project team intentionally created two types of questions: those using abstract representation of variables in the stem and those providing a real-world context (Supplemental Table 1). Because students encounter both types of questions when learning homeostasis, we were interested to know whether one representation was more challenging than other (as in Nehm *et al.*, 2012a; Prevost *et al.*, 2013; Weston *et al.*, 2015). For example, items 13 and 14 from the HCI both assess understanding of the role of effectors on regulated variables (Supplemental Material, Homeostasis Concept Inventory); item 13 is abstract and item 14 is an applied example in which the sweat gland is presented as an effector that can change body temperature, the regulated variable. Experts are used to working with this concept in both abstract and applied conditions. Because we did not want to assume that students would be equally capable with both types of questions, we included examples of each. In the HCI, we found no significant difference between question types.

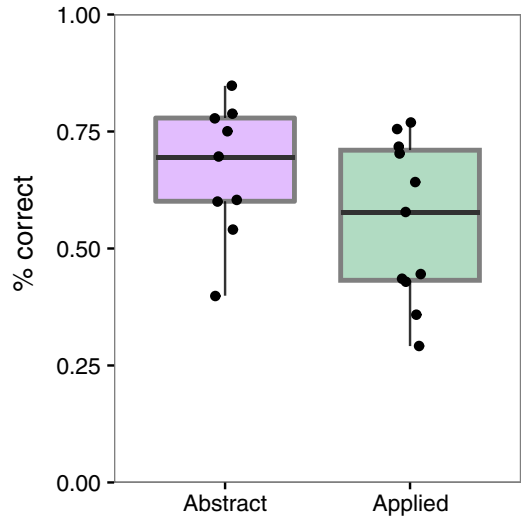

FIGURE 7. Students' performance on abstract questions was indistinguishable from their performance on related questions that applied knowledge to real-world scenarios (paired *t* test, $p = 0.132$). Here, difficulty is represented as the percentage of students who answered the item correctly (% correct), as in Figure 3.

Equity and Diversity in Validation

Equity and diversity are critical issues in undergraduate education. However, unconscious, implicit bias exists in science, technology, engineering, and mathematics (STEM) education. Because concept inventories are developed by the STEM community, it is important that the validation process includes tests to ensure that the individual items are free of bias regarding gender, ethnicity, and English language status. Although women scored lower than men, Black and Hispanic students scored lower than whites, and English language learners scored lower than native

speakers, DIF analysis indicated that no individual item is unfair. In other words, no item functions differently for students who have the same latent ability but are from different demographic groups, and therefore we conclude that the items do not incorporate cultural biases that would impact different constituents' performance on the test. This finding provides a new challenge for discipline-based education researchers more generally: when we develop concept inventories, we need to analyze the potential bias of individual items (perform a DIF analysis, see Martinková *et al.*, 2017b) to explain the observed results.

Men and women with the same latent ability should have the same score on the HCI. However, men and women performed differently on the HCI. Similar discrepancies in academic performance on exams have been observed between males and females taking introductory biology exams (Wright *et al.*, 2016). In their study, Wright *et al.* (2016) found that the genders performed equally on questions of lower cognitive levels (Dirks *et al.*, 2014), but males outperformed women on questions at higher cognitive levels. Social factors such as test anxiety and stereotype threat may explain this difference in test performance. Moderate levels of test anxiety are beneficial to exam performance, but high levels negatively impact performance (Maloney *et al.*, 2014). Excessive worrying occupies more of the individual's working memory, thus limiting cognitive resources to solve exam problems. Some research supports the idea that females have higher test anxiety than males (Stenlund *et al.*, 2017). Stereotype threat is invoked when an individual encounters a problem that societal norms has deemed challenging for the group to which he or she identifies (Steele *et al.*, 2002; Shapiro and Neuberg, 2007). The classic example is females' lower performance than academically matched males on math exams. Homeostasis is a challenging concept and is taught in a STEM course; both of these factors may have been at play while students were taking the HCI (Shapiro and Williams, 2012).

We recommend that students from all institutions of higher learning be routinely included in the development and validation of concept inventories. A thorough suite of statistical tools must also be employed when determining the validity and reliability of the inventory, as it is critical that we ensure the inventory is free of bias against any demographic group. Only when we fully understand the limitations of each inventory will we be able to make the proper inferences from the results.

Limitations and Future Work

The HCI, like any concept inventory, can always be improved. Others have emphasized the ever-changing nature of concept inventories by including version numbers in the names of their instruments (e.g., Price *et al.*, 2014; Newman *et al.*, 2016). Our analysis of the HCI revealed that some items will require attention in the future. Item 17, for example, was not correlated with any of the other items in the instrument, was difficult, and did not discriminate between lower and higher abilities very well. Despite this, we decided to retain the item in this version of the HCI, because it helps us assess how students understand the way different components interact in the system that regulates blood pressure. Additional interviews and questions on this topic should be developed in the future, but these additional steps were beyond the scope of this first HCI.

It should also be noted that items 4, 9, and 20 formed a cluster in the tetrachoric correlation heat map (Figure 6), which could have been due to either their format or their content. The distractors for these questions all included combinations of possible answers, for example, low concentration, high concentration, either, both. This style, sometimes referred to as type K questions, can be misleading to students, and therefore answers may not align to conceptual understanding (Haladyna and Downing, 1989; Libarkin, 2008). We included the questions in this first HCI, because all three items measure the same challenging, yet essential, key concept that homeostasis is functioning all the time. Nonetheless, the fact that these items cluster together in our statistical analysis suggests that they may need future attention as well.

Future iterations of the HCI may also need items to be revised so that each assesses only one part of the HCF (McFarland *et al.*, 2016) at a time. For example, item 17 may be so challenging because it assesses three different ideas about the role of a control center, the way the control center interacts with the sensor, and the way it changes the effectors (Supplemental Table 1). As another example, item 20 includes two different concepts, one pertaining to the role of sensors, and the other pertaining to the fact that the body is constantly working to maintain homeostasis. Of course, the challenge with this kind of revision is that the instrument would become longer; we prioritized building a short instrument that is easy and quick to administer in a class or as a short homework assignment.

Finally, these limitations also speak to some of the challenges of using closed-response tests. The questions are always the same, students can become familiar with them, and their usefulness diminishes over time (Nehm *et al.*, 2012a). Answers can also be readily available, and consequently some authors have chosen not to publish their concept inventory questions (e.g., Deane *et al.*, 2016). Also of concern is the fact that students' open-ended responses can expose more nuanced detail about how students understand challenging and fundamental concepts than can closed-response questions (Prevost *et al.*, 2013). Open-ended responses also help students practice communicating their understanding of challenging scientific concepts in ways that professional scientists do, rather than breaking them into artificially discrete pieces of information (Nehm *et al.*, 2012b). The drawback, of course, is that administering and grading open-ended response questions is time-consuming.

Despite these limitations, concept inventories are valuable tools for helping biology instructors gain a sense of what their students understand and helping instructors tailor their teaching to their student populations (D'Avanzo, 2008). They are also incredibly helpful when instructors and researchers use them to measure what students have learned and to detect concepts that are particularly challenging to teach (D'Avanzo, 2008; Smith and Knight, 2012). On the other hand, concept inventories are constrained by the quality of the procedures used to assess the validity of scores (Jorion *et al.*, 2015; Reeves and Marbach-Ad, 2016). For this reason, we worked with a particularly diverse student body during the development of the HCI (Tables 2 and 3; see also Abraham *et al.*, 2014) and conducted extensive item-level analyses to assess the fairness of these items with our diverse population (Supplemental Table 3; Martinková *et al.*, 2017b).

Implications for Teaching and Learning

The HCI can be used across undergraduate physiology courses and at all types of undergraduate institutions. The instrument

targets the most essential constituent ideas that make up the core concept of homeostasis. Most students can complete the HCI within 20 minutes, so instructors can administer it easily and quickly in their courses. These characteristics make the HCI practical and useful for formative assessment in undergraduate courses.

Each year, 650,000 students take introductory biology, which includes physiology, and 450,000 take anatomy and physiology (S. Beauparlant, Pearson Benjamin Cummings, personal communication). Homeostasis is one of the most important core concepts in physiology (Michael, 2007). We propose a multistep process for instructors to consider when they develop their plan for teaching the concept of homeostasis. First, we recommend that they use the HCF (McFarland *et al.*, 2016) to design interventions and activities to help students build a robust understanding of the process of homeostasis. Second, we recommend that instructors be aware of the misconceptions about homeostasis that cause students to use faulty reasoning that leads to incorrect predictions when explaining perturbations to a physiological system (Modell, 2000; Modell *et al.*, 2015). Familiarity with the common student misconceptions regarding the core concept of homeostasis (Wright *et al.*, 2013, 2015; Modell *et al.*, 2015) will allow instructors to be better prepared to develop in-class activities that generate the deep thinking that students require to move to more expert-like thinking (National Research Council, 2000). Finally, we suggest that the HCI be used in pre/posttesting to provide formative assessment of student learning and for instructors to assess their own effectiveness.

ACKNOWLEDGMENTS

We dedicate this paper to the loving memory of our coauthor, Ann Wright. We thank the many physiology experts and students who have helped us with this project; Peggy Brickman, Jennie Dorman, Adela Drabinova, Elizabeth A. Sanders, and Elli J, Theobald; and two anonymous reviewers. This research was supported by the National Science Foundation (DUE-1043443) and by the Czech Science Foundation (grant number GJ15-15856Y).

REFERENCES

Abraham, J. K., Perez, K. E., & Price, R. M. (2014). The dominance concept inventory: A tool for assessing undergraduate student alternative conceptions about dominance in Mendelian and population genetics. *CBE—Life Sciences Education*, *13*, 349–358.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*, Monterey, CA: Brooks/Cole.

American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education: A call to action*, Washington, DC.

Ames, A. J., & Penfield, R. D. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Management: Issues and Practice*, *34*, 39–48. doi: 10.1111/emip.12067

Association of American Medical Colleges and Howard Hughes Medical Institutes. (2009). *Scientific foundations for future physicians*, Washington, DC.

Bates, D., Maechle, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. doi: 10.18637/jss.v067.i01

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *57*, 289–300.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education*, *15*, rm4. doi: 10.1187/cbe.16-04-0148

Carnegie Classification of Institutions of Higher Education (n.d.). About Carnegie Classification. Retrieved January 26, 2017, from http://carnegieclassifications.iu.edu

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal Statistical Software*, *28*, 1–29.

Cooper, S. J. (2008). From Claude Bernard to Walter Cannon. Emergence of the concept of homeostasis. *Appetite*, *51*, 419–427.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. doi: 10.1007/bf02310555

Crowther, G. J., & Price, R. M. (2014). Re: Misconceptions are "so yesterday!" *CBE—Life Sciences Education*, *13*, 3–5. doi: 10.1187/cbe.13-11-0226

D'Avanzo, C. (2008). Biology concept inventories: Overview, status, and next steps. *BioScience*, *58*, 1079–1085.

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2016). Development of the Statistical Reasoning in Biology Concept Inventory (SRBCI). *CBE—Life Sciences Education*, *15*, ar5. doi: 10.1187/cbe.15-06-0131

De Ayala, R. J. (2008). *The theory and practice of item response theory*, New York: Guilford.

Dirks, C., Wenderoth, M. P., & Withers, M. (2014). *Assessment in the college science classroom*, New York: Freeman.

Drabinova, A., Martinková, P., & Zvara, K. (2016). difNLR: Detection of Dichotomous Differential Item Functioning (DIF) by Non-Linear Regression Function R Package, version 1.0.0. Retrieved May 11, 2017, from https://cran.r-project.org/package=difNLR/

Fletcher, T. D. (2010). psychometric: Applied Psychometric Theory, R Package Version 2.2. Retrieved May 11, 2017, from https://CRAN.R-project.org/package=psychometric

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement Education*, *2*, 37–50. doi: 10.1207/s15324818ame0201_3

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, Inference and prediction*, New York: Springer.

Heckler, A. F. (2010). *Concrete vs. abstract problem formats: A disadvantage of prior knowledge*. Paper presented at 9th International Conference of the Learning Sciences (Chicago, IL).

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Jorion, N., Gane, B. D., James, K., Schroeder, L., DiBello, L. V., & Pellegrino, J. W. (2015). An analytic framework for evaluating the validity of concept inventory claims. *Journal of Engineering Education*, *104*, 454–496. doi: 10.1002/jee.20104

Kalinowski, S. T., Leonard, M. J., & Taper, M. L. (2016). Development and validation of the Conceptual Assessment of Natural Selection (CANS). *CBE—Life Sciences Education*, *15*, ar64. doi: 10.1187/cbe.15-06-0134

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2013). The cost of concreteness: the effect of nonessential information on analogical transfer. *Journal of Experimental Psychology: Applied*, *19*, 14–29.

Knight, J. K., & Wood, W. B. W. (2005). Teaching more by lecturing less. *Cell Biology Education*, *4*, 298–310. doi: 10.1187/05-06-0082

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Tests in Linear Mixed Effects Models, R Package Version 2.0–30. Retrieved May 11, 2017, from https://CRAN.R-project.org/package=lmerTest

Leonard, M. J., Kalinowski, S. T., & Andrews, T. C. (2014). Misconceptions—yesterday, today, and tomorrow. *CBE—Life Sciences Education*, *13*, 179–186. doi: 10.1187/cbe.13-12-0244

Libarkin, J. (2008). *Concept inventories in higher education science*. Paper presented at National Research Council Promising Practices in Undergraduate STEM Education Workshop 2 (Washington, DC).

Livingston, S. A. (2006). Item analysis. In: Downing S. M., & Haladyna T. M. (Eds.), *Handbook of test development* (pp. 421–441). London: Erlbaum.

Magis, D., Beland, S., & Raiche, G. (2015). difR: Collection of Methods to Detect Dichotomous Differential Item Functioning (DIF), R Package, version 4.7. Retrieved May 11, 2017, from https://cran.r-project.org/package=difR/

Maloney, E. A., Sattizahn, J. R., & Beilock, S. L. (2014). Anxiety and cognition. *WIREs Cognitive Science*, *5*, 403–411. doi: 10.1002/wcs.1299

Martinková, P., Drabinova, A., Leder, O., & Houdek, J. (2017a). ShinyItemAnalysis: Test and Item Analysis via Shiny, R Package. Retrieved May 11, 2017, from https://CRAN.R-project.org/package=ShinyItemAnalysis

Martinková, P., Drabinovà, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017b). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*, rm2.

McFarland, J., Wenderoth, M. P., Michael, J., Cliff, W., Wright, A., & Modell, H. (2016). A conceptual framework for homeostasis: development and validation. *Advances in Physiology Education*, *40*, 213–222. doi: 10.1152/advan.00103.2015

McNeil, N. M., Uttal, D. H., Jarvin, L., & Sternberg, R. J. (2009). Should you show me the money? Concrete objects both hurt and help performance on mathematics problems. *Learning and Instruction*, *19*, 171–184.

Michael, J. (2001). In pursuit of meaningful learning. *Advances in Physiology Education*, *25*, 145–158.

Michael, J. (2007). Conceptual assessment in the biological sciences: a National Science Foundation-sponsored workshop. *Advances in Physiology Education*, *31*, 389–391. doi: 10.1152/advan.00047.2007

Michael, J., & McFarland, J. (2011). The core principles ("big ideas") of physiology: results of faculty surveys. *Advances in Physiology Education*, *25*, 336–341. doi: 10.1152/advan.00004.2011

Michael, J. A., & Modell, H. I. (2003). *Active learning in secondary and college science classrooms: a working model for helping the learner to learn*, Mahwah, NJ: Erlbaum.

Michael, J., Modell, H., McFarland, J., & Cliff, W. (2009). The "core principles" of physiology: What should students understand? *Advances in Physiology Education*, *33*, 10–15. doi: 10.1152/advan.90139.2008

Modell, H. I. (2000). How to help students understand physiology? Emphasize general models. *Advances in Physiology Education*, *23*, 101–107.

Modell, H., Cliff, W., Michael, J., McFarland, J., Wenderoth, M. P., & Wright, A. (2015). A physiologist's view of homeostasis. *Advances in Physiology Education*, *39*, 259–266. doi: 10.1152/advan.00107.2015

Momsen, J. L., Long, T. M., Wyse, S. A., & Ebert-May, D. (2010). Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE Life Sciences Education*, *9*, 435–440. doi: 10.1187/cbe.10-01-0001

National Research Council. (2000). *How people learn: brain, mind, experience, and school (expanded ed.)*, Washington, DC: National Academies Press.

National Research Council. (2009). *A New biology for the 21st century: Ensuring the United States leads the coming biology revolution*. Washington, DC: National Academies Press.

National Research Council. (2012). *A framework for K–12 science education*, Washington, DC: National Academies Press.

Nehm, R. H., Beggrow, E. P., Opfer, J. E., & Ha, M. S. (2012a). Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. *American Biology Teacher*, *74*, 92–98.

Nehm, R. H., Ha, M., & Mayfield, E. (2012b). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, *21*, 183–196. doi: 10. 1007/S10956-01 1-9300-9

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. *International Journal of Science Education*, *33*, 1373–1405. doi: 10.1080/09500693.2010.511297

Newman, D. L., Snyder, C. W., Fisk, J. N., & Wright, L. K. (2016). Development of the Central Dogma Concept Inventory (CDCI) assessment tool. *CBE—Life Sciences Education*, *15*, ar9. doi: 10.1187/cbe.15-06-0124

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory*. New York: McGraw-Hill.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64.

Pollitt, A., Ahmed, A., Baird, J.-A., Tognolini, J., & Davidson, M. (2008). *Improving the quality of GCSE assessment*. Coventry, UK: Qualifications and Curriculum Authority.

Prevost, L. B., Knight, J., Smith, M., & Urban-Lurain, M. (2013). Student writing reveals their heterogeneous thinking about the origin of genetic variation in populations. Talk presented at the Annual Meeting of the National Association for Research in Science Teaching, April 6–9, Rio Grande, PR, www.narst.org/annualconference/2013_NARST_Abstracts.pdf

Price, R. M., Andrews, T. C., McElhinny, T. L., Mead, L. S., Abraham, J. K., Thanukos, A., & Perez, K. E. (2014). The Genetic Drift Inventory: A tool for measuring what advanced undergraduates have mastered about genetic drift. *CBE—Life Sciences Education*, *13*, 65–75.

R Core Team (2016). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers. *CBE—Life Sciences Education*, *15*, 1–9. doi: 10.1187/cbe.15-08-0183

Revelle, W. (2015). psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, Retrieved May 11, 2017, from http://CRAN.R-project.org/package=psych

Rizopoulos, D. (2006). ltm: an R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement*, *18*, 229–244.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Shapiro, J. R., & Neuberg, S. L. (2007). From stereotype threat to stereotype threats: Implications of a multi-threat framework for causes, moderators, mediators, consequences, and interventions. *Personality and Social Psychology Review*, *11*, 107–130. doi: 10.1177/1088868306294790

Shapiro, J. R., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles*, *66*, 175–183. doi: 10.1007/s11199-011-0051-0

Smith, M. K., & Knight, J. K. (2012). Using the Genetics Concept Assessment to document persistent conceptual difficulties in undergraduate genetics courses. *Genetics*, *191*, 21–32. doi: 10.1534/genetics.111.137810

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: the psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology*, *34*, 379–440.

Stenlund, T., Eklöf, H., & Lyrén, P. E. (2017). Group differences in test-taking behaviour: An example from a high-stakes testing program. *Assessment in Education Principles, Policy, & Practice*, *24*, 4–20. doi: 10.1080/0969594X.2016.1142935

Torres Irribarra, D., & Freund, R. (2014). Wright Map: IRT Item-Person Map with ConQuest Integration. Retrieved May 11, 2017, from http://github.com/david-ti/wrightmap

Valverde, G. A., & Schmidt, W. H. (1997). Refocusing US math and science education. *Issues in Science and Technology*, *14*, 60–66.

Wei, R., & Simko, V. (2010). corrplot: Visualization of a Correlation Matrix, R Package Version 0.77. Retrieved May 11, 2017, from https://CRAN.R-project.org/package=corrplot

Weston, M., Haudek, K. C., Prevost, L., Urban-Lurain, M., & Merrill, J. (2015). Examining the impact of question surface features on students' answers to constructed-response questions on photosynthesis. *CBE—Life Sciences Education*, *14*, ar19.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer.

Wright, A., McFarland, J., Cliff, W., Michael, J., Modell, H., & Wenderoth, M. P. (2013). Preliminary results on the prevalence of physiology students' homeostatic misconceptions. *FASEB Journal*, *27*, 739.735.

Wright, A., McFarland, J., Wenderoth, M. P., Michael, J., Modell, H., & Cliff, W. (2015). Knowing common misconceptions about homeostasis helps students' learning. *FASEB Journal*, *29*, 541.534.

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE—Life Sciences Education*, *15*, ar23. doi: 10.1187/cbe.15-12-0246

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Application of Bloom's taxonomy debunks the "MCAT myth." *Science*, *319*, 414. doi: 10.1126/science.1147852

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