Adv Physiol Educ 43: 15–27, 2019;
doi:10.1152/advan.00128.2018.

HOW WE TEACH | *Generalizable Education Research*

Phys-MAPS: a programmatic physiology assessment for introductory and advanced undergraduates

Katharine Semsar,[1] Sara Brownell,[2] Brian A. Couch,[3] Alison J. Crowe,[4] Michelle K. Smith,[5] Mindi M. Summers,[6] Christian D. Wright,[2] and Jennifer K. Knight[1]

[1]*Molecular, Cellular, and Developmental Biology, University of Colorado-Boulder, Boulder, Colorado;* [2]*Biology Education Research Laboratory, School of Life Sciences, Arizona State University, Tempe, Arizona;* [3]*School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, Nebraska;* [4]*Department of Biology, University of Washington, Seattle, Washington;* [5]*School of Biology and Ecology, University of Maine, Orono, Maine; and* [6]*Ecology and Evolutionary Biology, University of Calgary, Calgary, Alberta, Canada*

Submitted 11 July 2018; accepted in final form 25 October 2018

Semsar K, Brownell S, Couch BA, Crowe AJ, Smith MK, Summers MM, Wright CD, Knight JK. Phys-MAPS: a programmatic physiology assessment for introductory and advanced undergraduates. *Adv Physiol Educ* 43: 15–27, 2019; doi:10.1152/advan.00128.2018.—We describe the development of a new, freely available, online, programmatic-level assessment tool, Measuring Achievement and Progress in Science in Physiology, or Phys-MAPS (http://cperl.lassp.cornell.edu/bio-maps). Aligned with the conceptual frameworks of Core Principles of Physiology, and Vision and Change Core Concepts, Phys-MAPS can be used to evaluate student learning of core physiology concepts at multiple time points in an undergraduate physiology program, providing a valuable longitudinal tool to gain insight into student thinking and aid in the data-driven reform of physiology curricula. Phys-MAPS questions have a modified multiple true/false design and were developed using an iterative process, including student interviews and physiology expert review to verify scientific accuracy, appropriateness for physiology majors, and clarity. The final version of Phys-MAPS was tested with 2,600 students across 13 universities, has evidence of reliability, and has no significant statement biases. Over 90% of the physiology experts surveyed agreed that each Phys-MAPS statement was scientifically accurate and relevant to a physiology major. When testing each statement for bias, differential item functioning analysis demonstrated only a small effect size (<0.008) of any tested demographic variable. Regarding student performance, Phys-MAPS can also distinguish between lower and upper division students, both across different institutions (average overall scores increase with each level of class standing; two-way ANOVA, $P < 0.001$) and within each of three sample institutions (each ANOVA, $P \leq 0.001$). Furthermore, at the level of individual concepts, only evolution and homeostasis do not demonstrate the typical increase across class standing, suggesting these concepts likely present consistent conceptual challenges for physiology students.

concept assessment; concept inventory; curriculum reform; major; program

INTRODUCTION

Biology instructors and departments are increasingly using data-driven approaches to help improve student engagement, learning, and persistence (1, 29). Central to this approach is defining clear learning goals and using closely aligned assessments to measure student learning in a classroom (19, 58). This same approach can be adopted for curriculum reform, identifying clear outcomes that students should be able to achieve by the time they graduate, aligning the curriculum to these outcomes, and developing assessments that can both measure and make inferences about student learning across a major (4, 16). This approach to curriculum design can serve to help departments better understand how their curriculum is impacting students, which, in turn, can help improve student learning and critical thinking processes (2).

For the field of physiology, curriculum-level student learning goals have been articulated in two separate conceptual frameworks. The first framework, the Core Principles of Physiology (33), is specific to the physiology discipline. The Core Principles framework was informed by over 200 physiologists and ranks 15 physiology concepts from the most to least important to be learned during the undergraduate major. The second relevant conceptual framework, the American Association for the Advancement of Science Vision and Change report, is more general to all of biology, articulating five core concepts that all biology students should master by the time they graduate (3, 4). The development of the Vision and Change framework was informed by over 500 experts in biology education, is supported by several national organizations and overlaps with the Next Generation Science Standards for K–12 education (37). The Vision and Change framework has been further interpreted for what a general biology major should know about major biology subdisciplines, including physiology, in the *BioCore Guide* (9). Together, these two conceptual frameworks provide a strong foundation for defining the most critical concepts that physiology students should master during their undergraduate education.

To date, most concept assessments have focused on measuring student understanding of a single concept or suite of a few specific concepts (15, 24). Furthermore, they are best used to measure changes in student thinking across a single course rather than a curriculum (15, 24). An additional hurdle for the field of physiology is that, while there are concept assessments available for many subdisciplines of biology (e.g., Refs. 5, 15, 24, 44, 45, 48), few cover concepts specifically related to

Address for reprint requests and other correspondence: K. Semsar, Miramontes Arts and Sciences, University of Colorado-Boulder, 347 UCB, Boulder, CO 80309-0347 (e-mail: katharine.semsar@colorado.edu).

Table 1. *Overview of the Phys-MAPS development process and general timeline*

1. Use conceptual frameworks of Vision and Change and Core Principles of Physiology and a literature review of common student difficulties to define the set of concepts to be assessed. (Fall 2014)
2. Conduct open-ended interviews to probe student understanding of these concepts. (Fall 2014)
3. Draft a series of questions stems, incorporating student ideas into multiple-true/false statements (Spring-Fall 2015). Revise statements to likely/unlikely after expert feedback.
4. Iteratively modify questions and statements based on:

- 104 student think-aloud interviews. (Spring 2015–Spring 2016)
- Feedback from 46 physiology experts at 14 institutions regarding the scientific accuracy and clarity of each likely/unlikely statement. (Spring 2015–Spring 2016)
- Results from administering Phys-MAPS to students:
  Pilot 1 (Spring 2015): 318 physiology students at 1 institution.
  Pilot 2 (Fall 2015): 2014 students at 14 institutions.

5. Administer final version of Phys-MAPS to 3455 introductory and advanced students at 14 institutions. (Spring 2016–Fall 2016)
6. Conduct analyses to document overall student performance, question statistics, and instrument reliability. (Fall 2016)

Phys-MAPS, Measuring Achievement and Progress in Science in Physiology.

physiology. Only the osmosis and diffusion diagnostic test (42), diagnostic question clusters (60), and the homeostasis concept inventory (27) contextualize the problems relative to physiology, and each of these is intentionally narrow in scope.

Here we describe the development of a new programmatic-level conceptual assessment: Measuring Achievement and Progress in Science for Physiology, or Phys-MAPS. Phys-MAPS is one of a suite of new programmatic-level assessments referred to collectively as Bio-MAPS. This suite of instruments includes the Molecular Biology Capstone Assessment (12), EcoEvo-MAPS (52), and a general biology assessment (Gen-Bio-MAPS; in press). Each assessment is intended to measure broad-level changes in student thinking when administered at multiple time points during an undergraduate program, including when a student enters the major, after the introductory biology series, and just before graduation. The commonality of all of the Bio-MAPS assessments is that they are aligned with the core concepts of biology. While Bio-MAPS assessments were developed following the general methodology used for other biology concept inventories, they are specifically designed to measure student learning at the scale of a whole curriculum, covering a wider breadth of concepts. We report here on the evidence of Phys-MAPS' validity and reliability and suggest how departments wishing to collect data to make inferences about student conceptual struggles and learning in physiology during their undergraduate major could use Phys-MAPS. Collecting curricular data using Phys-MAPS can help inform data-driven conversations about department-level instructional change.

METHODS

Question Development

To develop the Phys-MAPS assessment, we followed a common approach of iterative question development that incorporates multiple cycles of student interviews, faculty feedback, and large-scale piloting to develop and provide evidence for validation of the assessment (e.g., Refs. 1, 12, 47, 48), as outlined in Table 1. All research activity was approved by the University of Colorado, Boulder Institutional Review Board (protocol no. 15–0283) and/or the Arizona State University Institutional Review Board (STUDY00001058).

Determining content coverage. We designed Phys-MAPS to be aligned to both the Vision and Change (9) and Core Principles of Physiology framework (33). The two conceptual frameworks have substantial overlap (Fig. 1). Some core principles fall within a single vision and change core concept, e.g., the core principle of homeostasis aligns with the vision and change core concept of systems. Meanwhile other core principles, such as cell membrane, span multiple core concepts of vision and change. To identify content for the Phys-MAPS, we focused on the overlap between the five Vision and Change Core Concepts (evolution, transformation of energy and matter, structure/function, information flow, and systems) and six of the seven top-ranked core principles (cell membrane, homeostasis, gradients, structure/function, cell-cell signaling, and interdependence). The only major disagreement between the frameworks centers on evolution. While physiologists ranked evolution as the least important of 15 core principles, it is one of the five primary Vision and Change Core Concepts. To resolve this difference, in the Phys-MAPS, we kept the concept of evolution as it relates specifically to core physiology principles, but dedicated only one scenario and four statements directly to the concept. Finally, when possible, we also

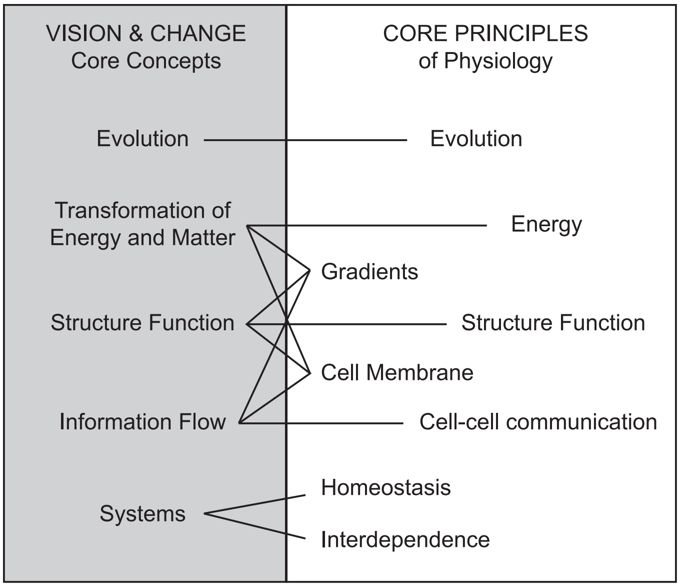

Fig. 1. Overlap of Vision and Change Core Concepts (9) and Core Principles of Physiology (33). All 5 of the Core Concepts from Vision and Change and 8 of the 15 Core Principles of Physiology are represented here. Lines between Core Concepts and Core Principles represent alignment between the concepts. Of the Core Principles listed here [and included on the Phys-MAPS (Measuring Achievement and Progress in Science in Physiology)], their ranking by physiologists from most to least important are as follows: homeostasis (1st), cell membrane (1st), cell-cell communication (3rd), interdependence (4th), (flow down) gradients (5th), energy (6th), structure-function (7th), and evolution (15th).

incorporated published student misconceptions, alternative conceptions, and known student difficulties (14, 22, 32, 34, 37, 38, 42).

Question design. We used student interviews to assist with question design. For each question, students read through a physiological scenario and determined whether a series of statements about the scenario (including predictions, conclusions, and interpretations) were "likely" or "unlikely" to be true. The likely/unlikely ranking is a modified version of the multiple true/false format developed for use in EcoEvo-MAPS (52). With the multiple true/false format, the mixed mental models of students who hold both correct and incorrect ideas simultaneously can be captured (13, 18, 25, 40, 41, 57). We had originally designed Phys-MAPS statements using the multiple true/false format, but, based on feedback from experts and student interviews, we changed to likely/unlikely. As was found for questions in ecology and evolution (52), students and experts found the absolute terms of true and false challenging when making predictions and generalizations, or when transferring their knowledge to a novel scenario. After switching to likely/unlikely, experts felt more comfortable agreeing that a statement's answer was scientifically accurate. Also, students took less time during think-aloud interviews to answer the questions, largely because they stopped trying to remember if there were possible exceptions to the "rule" they were applying. For example, when answering questions in the true/false format, one representative student said, "Well, I think this would be true, but give me a minute, there's always something I seem to forget." With the switch in format to likely/unlikely, students appeared more focused on generalizing concepts rather than spending time trying to remember factual exceptions.

Iterative revision to increase validity. Over the course of developing Phys-MAPS assessment, we interviewed 104 students from two institutions (general student characteristics are reported in Table 2) and surveyed 46 physiology experts from 14 institutions. For student interviews, we followed a think-aloud interview protocol (1). Interviews conducted on the first drafts of questions (interviews, $n = 12$) were largely open ended to help establish student thinking about specific concepts. Interviews conducted on revised versions of questions (interviews, $n = 92$) served to establish that both introductory and advanced students interpreted the figures and the language of the scenarios and statements as intended and to ensure that students' reasoning about each statement matched their answer choices (Table 3).

To determine whether questions and likely/unlikely statements were scientifically accurate, written clearly, and appropriate for physiology majors, the final version of each statement was reviewed by at least five experts, with 12–18 experts reviewing the majority (56 of 68) of final statements (Table 3). Of note, during the iterative development and expert review of questions relating to the concept, homeostasis, we found, as did Modell and colleagues (37), that physiologists around the country used inconsistent terminology. For example, not all experts used the terminology, "regulated variable," nor were all experts comfortable with the use of "homeostatic" as an adjective. This led to extensive rewriting of homeostasis statements to find common language that, in the final statements, was accurately interpreted by a range of experts and students.

Table 2. *Characteristics of the student interviewees*

Characteristic	Student Participants, *n*
Institution	
University of Colorado-Boulder	80
Arizona State University	24
Gender	
Male	37
Female	67
Major	
Physiology	76
Biology	27
Other	1
Class standing	
Lower division	27
Upper division	77

n, No. of students.

Table 3. *Summary of the question reviews on the final versions of Phys-MAPS statements by faculty experts and students*

	No. of Statements with the Given Level of Agreement/Matching		
	>90%	80–90%	<80%
Faculty review: Item is scientifically accurate.	68	0	0
Faculty review: Item is clearly written.	61	7	0
Faculty review: Item is appropriate for a graduating physiology major.	68	0	0
Student review: Students' answers match their reasoning.	64	4	0

For example, for each of the 68 Phys-MAPS (Measuring Achievement and Progress in Science in Physiology) statements, over 90% of faculty who reviewed the final version agreed the statement was scientifically accurate.

Piloting. Before the final administration of Phys-MAPS, we ran two pilots (Table 1). Following the first pilot, we calculated classical descriptive statistics (difficulty and discrimination) to identify statements to eliminate or revise. If a statement had a difficulty score (percent correct) <30% or >90% or a discrimination index <20%, we either dropped or revised the statement, unless it was a known conceptual difficulty with a high difficulty on the pilot [e.g., identifying functions of signaling models in homeostatic pathways (37)]. We then conducted additional interviews on the low discriminating statements to ensure we were confident that students were interpreting these statements correctly.

Following the national second pilot, we used both classical test theory and item response theory (IRT) to categorize statements (psychometric methodology described below). Of the 13 questions and 81 statements in this second pilot, we dropped 1 question along with its 8 likely/unlikely statements because one-half of the statements for that question were poor discriminators. Of the remaining 12 questions and 73 likely/unlikely statements, 60 statements met the discrimination criteria of being moderate or higher discriminators in a three-parameter logistic model (3PL) IRT model (8); however, we dropped two of these statements because they were repetitive with other statements. Of the 13 statements that did not meet the discrimination criteria, we dropped 3, revised 8 based on additional student interviews, and kept 2 low-discriminating but highly difficult statements. These two poorly discriminating statements were both very difficult, even among advanced students. Furthermore, these two statements were both listed by 100% of our experts as something physiology majors should know by the time they graduate. These modifications, plus the addition of two new statements, resulted in a Phys-MAPS version for final piloting that had a total of 12 questions with 70 statements.

Content coverage. Throughout the process, authors J.K.K. and K.S. classified the likely/unlikely statements as addressing one or two concepts for each of the frameworks (Core Principles of Physiology and Vision and Change; Table 4). About one-half of the statements aligned with a single concept, whereas the other one-half aligned with two different concepts. For example, statements that asked for predictions of ion flow across a membrane aligned with both core principles of gradients and cell membrane.

Administration. We administered the final version of Phys-MAPS to over 2,900 students at 13 universities in Spring 2016, and again to an incoming group of first-year students ($n > 500$) at an additional institution in Fall 2016. General demographics of the universities and students who participated are included in Tables 5 and 6.

Table 4. *Alignment of Phys-MAPS statements to Vision and Change Core Concepts and Core Principles of Physiology*

	Questions, no.	Statements, no.	Alignment with Core Concept†
Vision and Change Core Concepts			
Evolution	1	4	W2, W3, W4, W5
Structure function*	10	19	**C4**, **C5**, **C6**, **E1**, **E2**, **E6**, G3, G4, G5, **H1**, H6, **I5**, J6, K5, **V1**, **V5**, W1, W6, **Z5**
Information flow	8	22	**B1**, **B2**, **B3**, **B4**, **B5**, **B6**, E3, F3, F4, F5, G2, I1, I2, I3, I4, **I5**, I6, **V1**, V2, V3, **V5**, **Z5**
Transformation energy and matter	6	21	C1, C2, C3, **C4**, **C5**, **C6**, **E1**, **E2**, **E6**, **H1**, H2, H3, H4, J1, J2, J3, J4, J5, K3, Z3, Z6
Systems	7		**B1**, **B2**, **B3**, **B4, B5**, **B6**, E4, E5, F1, F2, G1, K1, K2, K4, V4, Z1, Z2, Z4
Core Principles of Physiology			
Homeostasis	5	10	E5, F1, F2, G1, K1, K2, **K4,** Z1, Z2, Z4
Cell-cell communication	5	16	F3, F4, F5, G2, I1, I2, I3, I4, **I5**, I6, **V1**, V2, V3, **V4, V5**, **Z5**
Gradients	6	14	C1, C2, C3, E2, **H2**, H3, **H4**, **J1**, **J3, J4**, **J5**, K3, Z3, Z6
Cell membrane	4	13	**C4, C5, C6, E6**, H1, **H2**, **H4**, **H5**, **J1,** J2, **J4**, **J5, J6**
Interdependence	5	11	B1, B2, B3, B4, B5, B6, E4, **J3**, K1, **K4, V4**
Structure function*	9	18	**C4, C5, C6**, E1, **E6**, G3, G4, G5, **H5**, H6, **I5**, **J6**, K5, **V1, V5**, W1, W6, **Z5**
Evolution	1	4	W2, W3, W4, W5
Gene-to-protein	1	1	E3

Questions are referenced by letter (B, C, E, F, G, H, I, J, K, V, W, Z); statements within questions are referenced by number. *This concept is defined slightly differently between conceptual frameworks; therefore, the items aligning within each framework are slightly different. †Statements in bold are aligned with more than one concept/principle.

To conduct this national administration of Phys-MAPS, we recruited faculty members teaching physiology courses in either physiology or general biology departments. Instructors agreed to offer Phys-MAPS to students in the last few weeks of the semester during a time devoid of other major tests or projects. Instructors also agreed to provide a small amount of participation credit for any student taking the online survey (regardless of whether or not they agreed to be part of the study). We introduced students to the study through an e-mail, either from their instructor or from author K.S. The e-mail asked students to give their best effort and told students that their participation would help the department improve its educational program [as recommended by Steedle (51)]. Students took Phys-MAPS online through the Qualtrics platform.

The online Phys-MAPS survey included a consent form, Phys-MAPS questions, an effort survey [student opinion survey (SOS)] (54), and a demographic survey. Each student answered all 12 Phys-MAPS questions. While we randomized the order in which students saw the 12 questions, we did not randomize the order of individual likely/unlikely statements within each question. Because there are groups of likely/unlikely statements that relate to specific subconcepts of the questions, we felt that randomizing the statement order would make it unnecessarily difficult for students, requiring them to jump back and forth among the subconcepts being assessed, rather than being able to think through one subconcept at a time. At the end of the assessment, the students answered the SOS effort questions and a set of demographic questions. The SOS effort scale included five Likert-scale statements relating to student effort when taking the test (54). Demographic questions included the following: class standing (year in college), transfer status, whether Advanced Placement (AP) Biology was taken in high school, number of college biology courses, number of college physiology courses, major, course specialization, grade point average (GPA), sex, underrepresented minority (URM) identity (not White or Asian), whether English was spoken at home growing up, and highest education completed by a student's parents (first-generation status). See Table 7 for a complete list of questions.

Table 5. *Carnegie classifications of piloting institutions for the final Phys-MAPS administration*

Institution Characteristic	Institutions, *n*	Student Participants, *n*
Institution type		
Public	12	2,449
Private	2	133
Institution size		
5–15,000	3	225
15–30,000	2	202
30–50,000	7	1,641
50,000+	2	514
Research activity		
RU/VH	12	2,446
RU/M	1	44
Master's/L	1	92
Department type		
Physiology	5	1,426
Biology	8	1,112
Kinesiology	1	44
Region		
Northeast	1	158
South	4	615
Midwest	5	577
Mountain west	1	444
Southwest	1	424
West coast	2	364

n, No. of institutions or participants. RU/VH, research university/very high research activity; RU/M, research university/medium research activity; Master's/L, master's granting institution/larger program.

Statistical Analysis

We excluded responses from students who had not agreed to join the study or were younger than 18 yr. In addition, we excluded responses from students who completed the survey more quickly than the questions could realistically be read (8 min), or did not take the assessment seriously (had an SOS score <8 out of 25, or did not answer four or more statements). Using these criteria, we removed 855 responses from the initial data. For those remaining responses (n = 2,600), if a student did not answer a question, we marked the question incorrect.

IRT modeling. As standard practice in psychometric analysis, IRT is a highly robust method for identifying item difficulty as it relates to student ability. We used a 3PL IRT to analyze item difficulty, discrimination, and pseudo-guessing, using the software packages R-Studio (46) and MIRT (10). Unlike the 1PL and 2PL models, the 3PL model incorporates an additional parameter for each statement that estimates the "pseudo-guessing" rate for the lowest performing students. Knowing we had statements for which students have known misconceptions and thus do not have 50:50 guess rates on statements, the 3PL model was the most appropriate. After running the first IRT model on the final pilot, we dropped two statements that did not meet the discrimination criteria of at least 0.34 (8). These two statements were also less valued by faculty than other statements: between 80 and

Table 6. *Number of participants and overall students' Phys-MAPS scores for the final Phys-MAPS pilot*

	First-Year	Sophomore	Junior	Senior	Post-Baccalaureate
Students, *n*	366	653	737	748	67
Average score (SD), %	54.7 (7.5)	58.2 (10.5)	60.7 (10.6)	62.2 (11.0)	66.5 (11.1)

Average Phys-MAPS (Measuring Achievement and Progress in Science in Physiology) scores are significantly different from each other at each level of class standing (two-way ANOVA, $P < 0.001$; Tukey post hoc testing, first-year to sophomore $P < 0.0001$, sophomore-junior $P < 0.0001$, junior-senior $P = 0.05$, senior-post-baccalaureate $P = 0.008$).

90% faculty said these statements were relevant, compared with >90% for all other statements. Given both of these factors, we dropped these two items from the final assessment and repeated the 3PL IRT model with the remaining 68 items. All additional statistics used these remaining 68 statements.

Instrument reliability. To estimate instrument reliability (i.e., the consistency with which an assessment measures student performance), we calculated Cronbach's alpha (α), which reflects the internal consistency of student responses by measuring the degree of covariance (ranging between 0 and 1) between all of the items on the test. Higher covariance indicates that high-performing students (as measured by their overall score on Phys-MAPS) outscore low-performing students on most items. We calculated α based on overall statement scores.

Student performance. We first calculated individual student scores by summing the number of correct statements for each student and dividing by the total number of statements. We then calculated statement difficulty as a percentage of correct responses for each statement, and statement discrimination by subtracting the statement difficulty for the bottom one-third of students from the statement difficulty for the top one-third of students. Although IRT modeling also provides statement discrimination and difficulty scores, we provide classical descriptive statistics as well, because the resulting scores for difficulty and discrimination from IRT are not as intuitive as classical descriptive scores. This data presentation strategy has been adopted previously to help make test results more interpretable for the target audience (52, 56). In addition, IRT requires large sample sizes (over 1,500 students) and thus is not always an appropriate analysis for individual institutions that may have fewer students (20).

To further examine differences in student thinking at different time points across a major, we compared all student scores across class standing for both the entire Phys-MAPS assessment and each concept using two-way ANOVAs with Tukey post hoc tests. In addition, for three institutions where we had data from multiple time points in a major, we compared student scores of lower division students (first-year and sophomores) to upper division students (juniors and seniors) using two-way ANOVA with post hoc contrasts, and calculated the corresponding effect sizes with Cohen's *d*. All tests were run using JMP Pro 12 software.

Demographic and effort-level effects. To characterize the student sample and investigate possible effects of demographic variables and motivation on overall student scores, we conducted a linear mixed-model analysis, using institution as a random factor and the following 13 fixed factors: number of biology courses taken, number of physiology classes taken, class standing (first-year, sophomore, junior, senior), self-reported GPA, major (biology, yes/no), physiology specialization, transfer status (yes/no), completion of AP biology (yes/no), sex (male, yes/no), whether English was spoken at home (yes/no), first-generation college status (yes/no), ethnicity (URM, yes/no), and SOS effort scores (JMP Pro 12 software). The model estimates the absolute effect of each variable (unstandardized coefficients) that indicate the average change in student score with each unit of change in the variable (e.g., first year to second year).

To further examine any potential demographic bias on student scores on each individual Phys-MAPS statement, we ran a logistic regression DIF using the difR package in R-Studio (26, 46). We specifically investigated five demographic variables with test items: transfer status (yes/no), sex (male, yes/no), whether English was

Table 7. *Demographic questions following Phys-MAPS questions*

1. Are you 18 yr of age or older? (yes, no)
2. What is your current class standing? (first-year, sophomore, junior, senior, postbaccalaureate, graduate student, other)
3. Are you a transfer student? (yes, no)
  3b. If yes: What other types of institutions have you attended? (select all that apply: 2-yr college or community college, 4-yr college or university, other)
4. Did you take AP biology in high school? (yes, no)
  4b. If yes: What was your score on the AP biology exam? (1, 2, 3, 4, 5, not sure, did not take AP exam)
5. Approximately how many biology-related lecture courses have you taken, including any in which are currently enrolled? (open answer)
6. Approximately how many physiology-related lecture courses have you taken, including any in which are currently enrolled? (open answer)
7. Have you declared, or are you planning to declare, a major in physiology? (yes, no)
  7b. If no: Have you declared, or are you planning to declare, a major in biology? (yes, no)
  7c. If yes: Have you declared, or are you planning to declare, a physiology or physiology-related concentration within your biology major? (yes, no)
8. Please check the subdiscipline(s) of biology in which you have taken the most courses? (molecular/cell biology, physiology, ecology/evolution, no specialization/equal exposure)
9. What is your approximate current overall GPA? [0.0–0.69 (E or F), 0.7–1.69 (D− to D+), 1.7–2.69 (C− to C+), 2.7–3.69 (B− to B+), 3.7–4.00 (A− to A+)]
10. Gender (female, male, other, prefer not to answer)
11. What is your race/ethnicity (select all that apply)? (African American/Black, Asian/Asian American, Caucasian/White, Filipino, Hispanic/Latino, Native American/Alaska Native, Native Hawaiian, Pacific Islander, other, prefer not to answer)
12. Did you speak English at home when you were growing up? (yes, no)
  12b. If no: What language did you speak at home? (open answer)
13. Highest level of education completed by at least one of your parent(s) [did not complete high school, high school/GED, some college (but did not complete college), Associate's degree (2-yr degree), Bachelor's degree, Master's degree, advanced graduate degree (for example, DVM, MD, PhD), not sure]

AP, advanced placement; DVM, Doctor of Veterinary Medicine; GED, general equivalency diploma; GPA, grade point average; MD, Medical Doctor; PhD, Doctor of Philosophy.

spoken at home (yes/no), first-generation college status (yes/no), and race/ethnicity (URM, yes/no). The criterion of statistical significance ($P < 0.05$) was used to flag potentially biased statements, followed by calculating Nagelkerke's R^2 effect size, which classifies effect sizes as negligible, moderate, or large (23, 62).

RESULTS

Although the iterative design of Phys-MAPS included several pilots (Table 1), our results focus on the 2,600 student responses from the final Phys-MAPS administration. The final version of Phys-MAPS has 12 questions with 68 likely/unlikely statements.

Student Performance

The average score on the Phys-MAPS was 59.7% (SD 10.6). Student performance on statements varied in difficulty (percent correct range: 0.17–0.89%) and discrimination scores (range: 0.0–0.47). Both statement difficulty and discrimination also ranged within each conceptual category (Fig. 2). Overall, Phys-MAPS scores were significantly different by class standing, with each progressive step in class standing (from first-years to post-baccalaureates) resulting in significantly higher Phys-MAPS scores (Table 6). When we grouped statements according to the two conceptual frameworks, the majority of the concepts in both the Vision and Change framework and the Core Principles framework followed a similar pattern, with upper division students scoring significantly higher than lower division students (Fig. 3).

Only two concepts varied from this general pattern: evolution and homeostasis. For evolution concept scores, there was an overall significant difference among groups, but only first-year students scored significantly differently (lower) than other class years, and there was no stepwise progression in scores with class standing above sophomores. For homeostasis concept scores, there was an overall significant difference among groups, but only sophomore students scored significantly differently (lower) than other groups (Fig. 3).

For each of the three institutions with physiology majors that assessed students in courses at multiple time points throughout the major, students in the upper-division courses scored significantly higher than students in lower-division courses (Fig. 4). Cohen's *d* effect sizes for upper to lower division scores were as follows: *institution A*, $d = 0.79$; *institution B*, $d = 0.55$; and *institution C*, $d = 0.79$.

IRT Parameter Estimates

Most Phys-MAPS statements had a moderate or high discrimination (Tables 8 and 9). Of the 11 that had a low (but acceptable, >0.34) discrimination value, 7 were difficult or very difficult, 2 were very easy, and 2 were moderately difficult. The overall difficulty of the assessment was skewed toward higher difficulty, with 32 statements being more difficult than moderate and 20 statements being easier than moderate (Tables 8 and 9).

Evidence of Instrument Reliability and Validity

Over 90% of experts agreed that each statement on Phys-MAPS was scientifically accurate and relevant to a physiology major. Over 80% agreed that each statement was written clearly. The Cronbach's α measurement for Phys-MAPS assessment was 0.75.

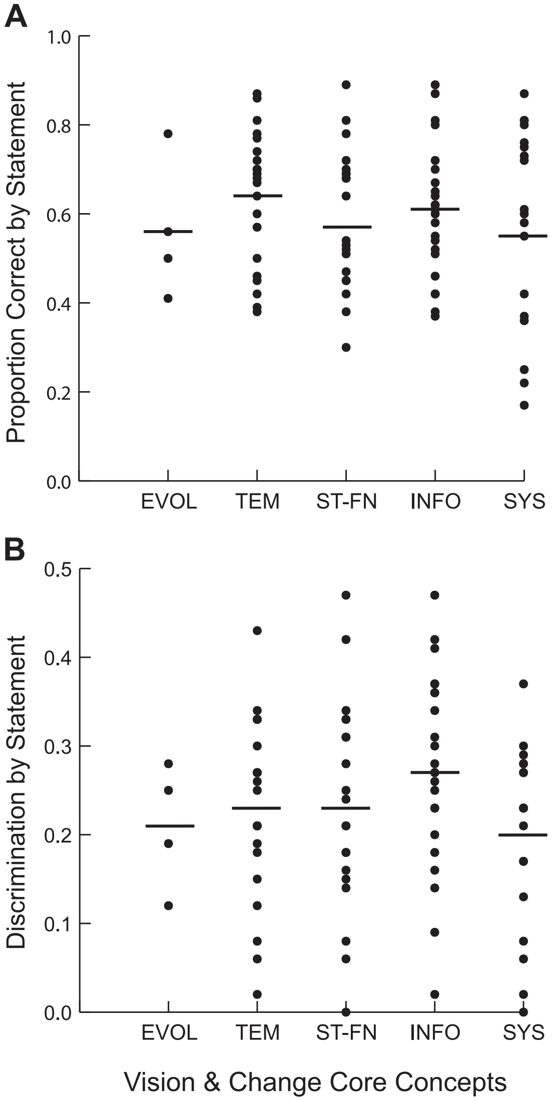

Fig. 2. Overall difficulty and discrimination of Phys-MAPS (Measuring Achievement and Progress in Science in Physiology) statements. Each circle depicts either the difficulty (*A*) or discrimination (*B*) for each likely/unlikely statement on the Phys-MAPS across all respondents, grouped by the Vision and Change Core Concepts. Horizontal lines depict the average difficulty or average discrimination for each Vision and Change Core Concept. EVOL, evolution; TEM, transformation of energy and matter; ST-FN, structure/function; INFO, information flow; SYS, systems.

Demographic and Motivation Effects

A linear mixed model that was used to explore the effects of demographic and motivation variables on overall Phys-MAPS scores was significant ($P < 0.001$) and accounted for 31.2% of the variation in the overall Phys-MAPS scores. At the level of overall Phys-MAPS scores, we found 8 of the 13 fixed variables to be significant ($P < 0.05$): SOS score (a measure of students' effort on the assessment), class standing, completion of AP Biology, number of college physiology courses, GPA, sex, number of college biology courses, and major. Demographic variables that did not significantly affect overall Phys-MAPS scores were as follows: physiol-

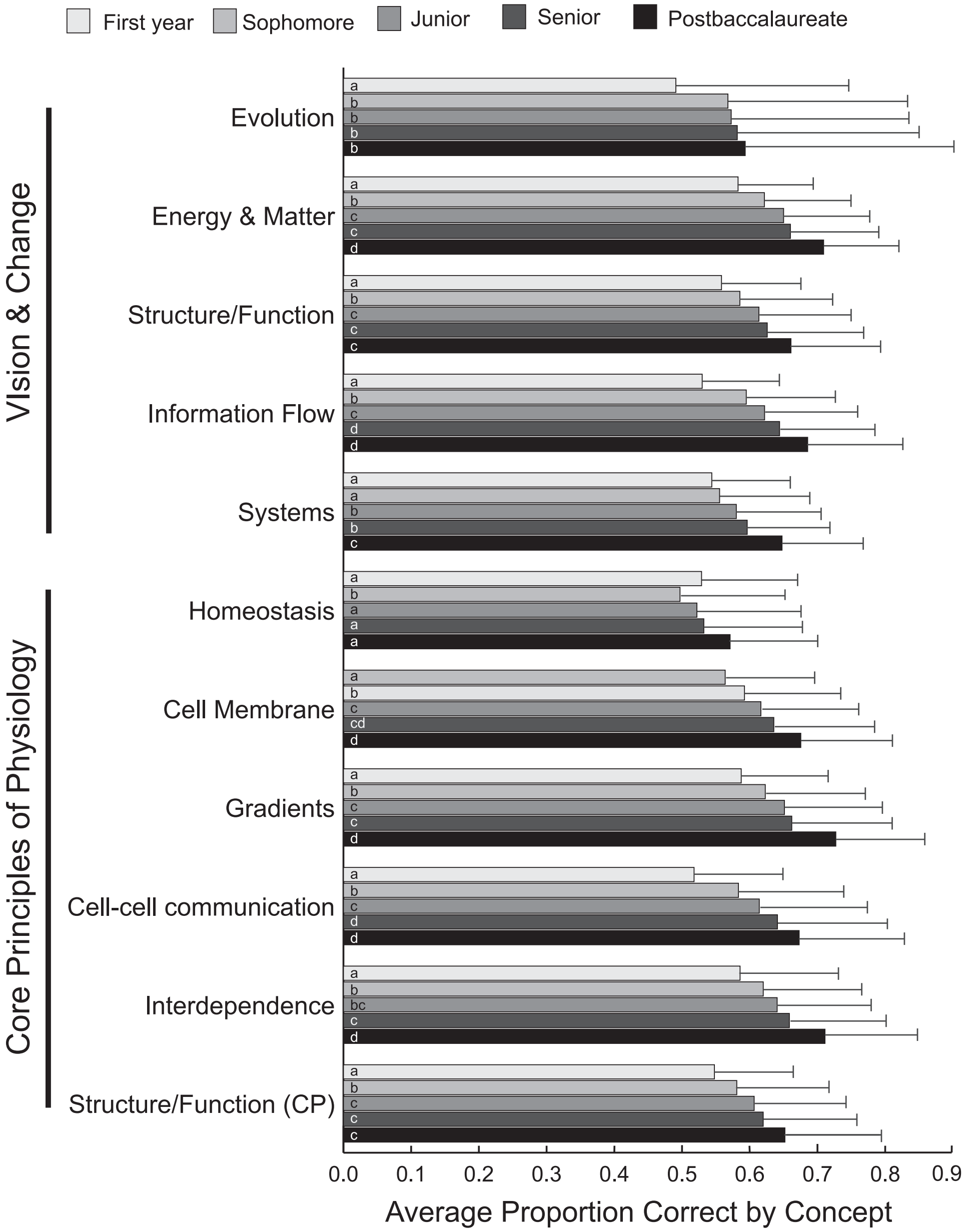

Fig. 3. Average student scores across class standing by concept. For each Vision and Change Core Concept and each Core Principle of Physiology, average student scores differ across class standing (two-way ANOVAs, one for each concept, all $P < 0.001$). Error bars represent standard variation of the averaged proportion correct of each statement in the group. [a–d] Class standing averages that are significantly different (Tukey post hoc testing, threshold: $P < 0.05$).

ogy course specialization ($P = 0.67$), race/ethnicity ($P = 0.39$), transfer status ($P = 0.48$), whether English was spoken at home ($P = 0.11$), and first-generation status ($P = 0.14$). Unstandardized coefficients for the eight significant variables are provided in Table 10.

The unstandardized coefficients indicated the average difference in student scores with each unit of change in the variable (e.g., cumulative GPA of C to B). For the continuous demographic variables of GPA and SOS effort score, students scored 3.8% higher for each letter grade and 0.7% for each point on the SOS effort survey (for example, students who stated they put in full effort, an SOS effort score of 25, had 7% higher Phys-MAPS scores than students who stated they put in a modest effort, with an SOS effort score of 15). For ordinal variables of number of college physiology courses, number of college biology courses, and class standing, unstandardized estimates were reported for each step in the scale. For example, students who had taken one or two physiology courses had an average Phys-MAPS score 2.3% higher than those who had taken zero physiology courses. Meanwhile students who had taken five or six physiology courses had an average Phys-MAPS score 4.0% higher than students who had taken three or four physiology courses. See Table 10 for a list of all of these effects. Finally, for the dichotomous demographic variables: men outscored other students by 1.9%, students who had taken AP Biology outscored other students by 0.7%, and physiology majors outperformed other students by 0.5%.

At the level of individual statements, the DIF analysis flagged 34 of the 68 likely/unlikely statements for one or more of the included demographic variables (transfer student, sex, whether English was spoken at home, first-generation college status, or URM identity). However, Nagelkerke's R^2 effect size for each of these likely/unlikely statements was never >0.008, and thus all were classified as having a negligible effect ($R^2 < 0.034$).

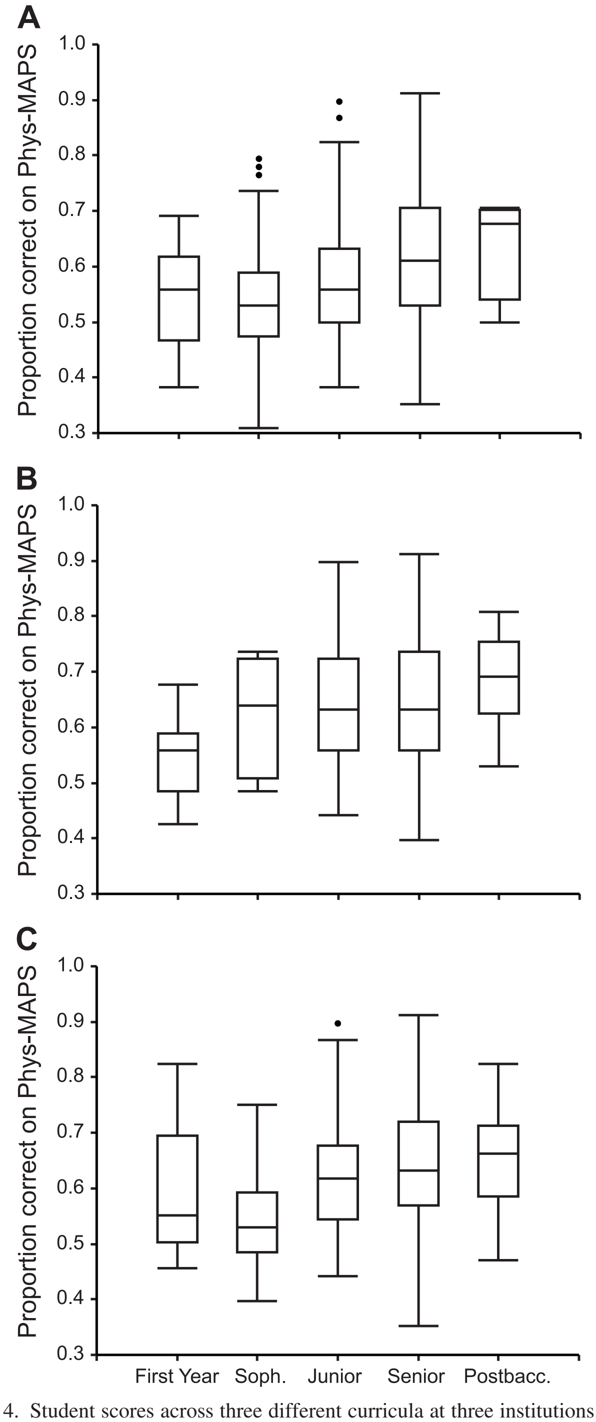

Fig. 4. Student scores across three different curricula at three institutions. For all three institutions (*A*, *B*, and *C*), average student scores on the Phys-MAPS (Measuring Achievement and Progress in Science in Physiology; whiskers on box plots) are higher for upper division students (pooled juniors and seniors) than for lower division students (pooled first-years and sophomores). *Institution A*: $F = 15.1$, df = 4, $P < 0.0001$, contrast $P = 0.001$. *Institution B*: $F = 4.64$, df = 4, $P = 0.001$, contrast $P = 0.01$. *Institution C*: $F = 10.0$, df = 4, $P < 0.0001$, contrast $P < 0.0001$.

DISCUSSION

The Phys-MAPS is a programmatic assessment tool that aligns with the nationally recognized conceptual frameworks for undergraduate physiology curricula, Vision and Change Core Concepts and Core Principles of Physiology. With the evidence of validity and reliability presented here, the Phys-MAPS can be used by departments to better understand population differences in student thinking at different time points in a physiology curriculum and make data-driven choices about curriculum development.

Student Performance

Phys-MAPS scores differed significantly across undergraduate class standing, with lower-division students scoring significantly lower than upper-division students (Table 6). In addition, mixed-model analysis showed overall scores increased across class standing, even when other variables are held constant (Table 10). While we cannot rule out the possibility that there was a disproportionate student selection bias of who took the survey at each level, or that the results were affected by students dropping out of physiology majors, these data nevertheless suggested that Phys-MAPS can distinguish differences among populations at different time points in a major.

When we partition Phys-MAPS scores by the concepts delineated in both the Vision and Change and Core Principles frameworks, nearly all concept scores showed a similar pattern, with post-baccalaureate students scoring the highest, and lower division students scoring the lowest. However, we did not see this general pattern for two concepts: evolution and homeostasis. For evolution, students at all levels scored similarly. While this performance is based on only four evolution statements, the statements ranged in their difficulty and align with major subconcepts of evolution on other Bio-MAPS assessments. This lack of difference in students' understanding of evolution between sophomores to seniors is particularly interesting, given how expert physiologists rank evolution as the least important physiology concept (33). In line with this, physiology programs generally do not emphasize evolution in their curricula (55). As approximately two-thirds of the student population in the final Phys-MAPS administration were in physiology-focused majors, it, therefore, may not be surprising that advanced students did not have a more advanced understanding of this concept. The improvement that is seen between first-years and sophomores may be due to introductory biology instruction before the specialization to physiology majors and concentrations. Arguably, this concept then served to act as a control for the sensitivity of Phys-MAPS, suggesting that the higher scores on other concepts were not simply because students were more advanced in their ability to read and answer any of the questions or that lower-performing students had dropped the major, but rather that changes in student performance were reflective of students having more advanced understanding of concepts most relevant to the field of physiology. Unfortunately, it also demonstrated that physiology students struggle with evolution concepts, even as graduating seniors.

The second exception to the pattern of higher scores among advanced students was the concept of homeostasis. In this case, sophomores had significantly lower scores than any other undergraduate class, and overall this concept appeared more challenging than any other. This finding warrants further research, especially as homeostasis is regarded as a central concept to the field of physiology (33, 37). Many of the homeostasis-related Phys-MAPS statements were representa-

Table 8. *Summary of three-parameter logistic model item response theory analysis on final Phys-MAPS administration*

Difficulty	Very easy (<−3)	Easy (−3 to −1)	Moderate (−1 to 1)	Difficult (1–3)	Very difficult (>3)
No. of statements	9	11	18	24	8
Discrimination	Very low (0–0.34)	Low (0.35–0.64)	Moderate (0.65–1.34)	High (1.35–1.69)	Very high (>1.69)
No. of statements	0	11*	28	14	15

Ranges for both difficulty and discrimination are described in Baker (8). *Seven of these are very difficult/difficult questions. All were rated as appropriate for physiology majors by faculty and interpreted accurately by students in interviews.

tive of the "sticky points," i.e., notably strong misconceptions (37, 49). The Phys-MAPS homeostasis "sticky points" are similar to the ones Modell and colleagues have previously described, such as, recognizing that not all negative feedback mechanisms are homeostatic, determining what physiological variables are and are not regulated through homeostatic mechanisms, and distinguishing among sensor and effector functions (28, 37). It is possible that students struggled with replacing their incorrect ideas, and that it would take more than just 1 or 2 yr of instruction to do so.

In addition to the difficult concept of homeostasis, Phys-MAPS was a challenging assessment overall (average score of 59.7%). However, given that all of the questions on Phys-MAPS were deemed "relevant to graduating seniors" by a consensus of physiology experts, the questions should serve to identify the most difficult elements of physiology. Importantly, each concept contained a set of statements with a range of difficulty levels, allowing Phys-MAPS to assess both easier and more difficult elements within each concept. Furthermore, as our goal was to align the assessment with concepts physiologists valued as important to the curriculum, we chose to retain the more difficult statements, even though some did not yet discriminate well due to their high difficulty level.

Advantages of Using Phys-MAPS for Program Assessment

Three design features of Phys-MAPS warrant more in-depth discussion as to how they are able to contribute meaningful data on student thinking that departments can use to redesign and build their physiology curricula: conceptual and application-level focus, modified multiple true/false design, and transparency of data collection.

Conceptual, application-level focus. We deliberately designed Phys-MAPS to have students apply basic concepts to

Table 9. *Statement discrimination, difficulty, and pseudo-guess rate values for the final item response theory analysis*

Question/Statement	a	b	g	Question/Statement	a	b	g
B1	1.534	−0.085	0.577	I1	1.933	0.168	0.365
B2	1.864	−0.228	0.566	I2	2.375	0.734	0.582
B3	1.716	0.927	0.396	I3	1.807	0.737	0.439
B4	0.581	7.828	0.36	I4	2.902	0.906	0.464
B5	1.49	1.505	0.301	I5	3.073	0.687	0.339
B6	2.023	−0.65	0.582	I6	2.398	1.121	0.293
C1	1	−1.373	0.471	J1	0.747	−1.33	0.526
C2	1.662	−0.231	0.324	J2	0.469	2.671	0.437
C3	1.148	1.251	0.475	J3	0.887	−0.146	0.444
C4	0.746	−1.956	0.118	J4	1.602	1.686	0.306
C5	1.255	−1.128	0.135	J5	0.743	5.32	0.488
C6	1.635	1.299	0.325	J6	0.573	2.189	0.531
E1	0.573	1.568	0.536	K1	0.611	0.609	0.341
E2	1.791	1.321	0.293	K2	0.492	−1.213	0.262
E3	2.118	0.925	0.5	K3	1.18	0.46	0.449
E4	0.56	5.914	0.587	K4	0.54	9.058	0.353
E5	0.362	0.159	0.46	K5	1.077	4.427	0.289
E6	0.776	4.631	0.36	V1	0.969	−1.809	0.393
F1	1.945	2.087	0.165	V2	1.035	1.628	0.326
F2	1.94	0.455	0.599	V3	1.809	0.222	0.415
F3	0.511	7.605	0.584	V4	0.814	−1.276	0.158
F4	1.793	1.861	0.317	V5	0.879	1.102	0.301
F5	1.344	1.064	0.488	W1	1.636	0.803	0.251
G1	1.123	2.616	0.187	W2	1.194	2.658	0.533
G2	1.069	0.868	0.588	W3	1.558	1.928	0.34
G3	0.91	0.609	0.489	W4	1.636	1.978	0.447
G4	1.317	0.68	0.511	W5	1.011	−0.863	0.335
G5	0.931	−0.563	0.244	W6	0.417	−1.074	0.289
H1	1.628	2.175	0.667	Z1	1.089	3.634	0.141
H2	1.352	1.426	0.551	Z2	0.843	−1.374	0.306
H3	1.211	0.518	0.517	Z3	0.927	−0.311	0.358
H4	1.447	1.717	0.379	Z4	0.894	0.959	0.379
H5	1.54	1.37	0.426	Z5	1.35	1.961	0.482
H6	1.487	1.839	0.377	Z6	0.902	−0.939	0.292

a, Statement discrimination; b, difficulty; g, pseudo-guess rate. Questions are referenced by letter; statements within question are referenced by number.

Table 10. *Estimated coefficients for statistically significant variables from a linear mixed-model analysis of demographic and motivation variables on overall Phys-MAPS scores*

	Unstandardized Coefficient		
Fixed Factor	Estimate	SE	*P* value
GPA	0.038	0.003	**<0.0001**
Sex (male, yes/no)	0.019	0.002	**<0.0001**
AP Biology (yes/no)	0.007	0.001	**0.0001**
SOS score	0.007	0.001	**<0.0001**
Physiology major (yes/no)	0.005	0.002	**0.05**
No. of physiology courses			
(1, 2)–(0) classes	0.023	0.008	**0.005**
(3, 4)–(1, 2) classes	0.013	0.006	**0.042**
(5, 6)–(3, 4) classes	0.040	0.009	**<0.001**
(7, 8)–(5, 6) classes	−0.003	0.012	0.778
(9+)–(7, 8) classes	−0.007	0.019	0.646
No. of biology courses			
(1, 2)–(0) classes	0.033	0.012	**0.008**
(3, 4)–(1, 2) classes	0.010	0.005	**0.038**
(5, 6)–(3, 4) classes	0.005	0.006	0.381
(7, 8)–(5, 6) classes	0.015	0.008	0.082
(9+)–(7, 8) classes	−0.005	0.009	0.584
Class standing			
Sophomores-first-years	0.008	0.010	0.455
Juniors-sophomores	0.011	0.005	**0.040**
Seniors-juniors	0.006	0.005	0.262
Post-baccalaureate-seniors	0.032	0.012	**0.007**

Dependent variable = percent Phys-MAPS (Measuring Achievement and Progress in Science in Physiology) score. Random factor = Institution. AP, Advanced Placement; GPA, grade point average; SOS, student opinion survey. Values in bold are statistically significant.

relatively novel or hypothetical scenarios. While concept assessments are generally written at Bloom's level of comprehension/application (6, 24), program-level assessments, such as the Human Anatomy and Physiology Society (HAPS) Comprehensive Exam (61) and the ETS Major Field Test for Biology, often assess students at the level of recall and understanding. On Phys-MAPS, each question provided students with the relevant facts necessary for students to apply their knowledge of general concepts to the specific scenarios. To avoid student familiarity with particular systems, we designed Phys-MAPS to use primarily novel and/or invented scenarios. For example, one of the original scenarios described the role of angiotensin and aldosterone in blood pressure regulation, a commonly used example when teaching homeostasis. Students commonly used strategies to recall what they knew, rather than using information in the scenario to answer the question. Accordingly, we rewrote that question into a scenario describing mosquito fluid balance and, consequently, saw students in interviews displaying application and analysis-level cognitive skills when answering the question rather than relying primarily on memory. Similar to using the likely/unlikely-to-be-true format, using clearly novel contexts shifted students' mindsets from factual mindsets to more conceptual mindsets.

Modified multiple true/false format. Similar to EcoEvo-MAPS (51), Phys-MAPS utilizes a modified multiple true/false format, using "likely/unlikely" instead of "true/false." In contrast to multiple choice, the multiple true/false format may better reflect how students answer open-ended questions by revealing that students have both correct and incorrect ideas about certain concepts (21, 25, 40, 41, 57, 59). In line with constructivist models of learning, students often integrate correct scientific ideas into incorrect mental models, thus creating heterogeneous mental models that simultaneously contain both accurate scientific and inaccurate ideas (40, 41). Several suites of statements on Phys-MAPS demonstrated these heterogeneous mental models in student thinking. For example, in one of our questions, students were asked to predict whether a protein can cross a membrane under different circumstances. While most seniors stated that the protein cannot freely cross the membrane, one-half of seniors also said that the protein can likely move through an open ion channel. Thus, whereas seniors often recognized that a protein needs some sort of transport, one-half remained unclear as to what that transport mechanism is or why it is needed. By using the modified multiple true/false format, Phys-MAPS can provide useful information about how students build their knowledge, starting with where students' knowledge base is entering a physiology program, in what concepts students develop proficiency during general biology courses, and what inaccurate/incomplete concepts students still hold on graduation.

One potential drawback of this format is the assumption that each statement has a 50% theoretical guess rate. If one assumes this, it has the potential to limit the range of overall scores. However, it is important to keep in mind three other factors when interpreting Phys-MAPS data. First, while the guess rate is theoretically higher than the multiple-choice format, this is offset by students being able to answer more individual items in the same amount of time, generally resulting in higher reliability overall (18). Second, many individual Phys-MAPS statement difficulty scores for lower division students include scores >0.9 and <0.2. These extremes would be difficult to achieve by chance (12) and indicate that the Phys-MAPS captures student thinking beyond mere guessing, even for introductory-level students. The third factor to keep in mind is that many of these individual statements represent known conceptual difficulties that have also been documented elsewhere. As many of these conceptual difficulties (e.g., homeostasis) persist to senior year (e.g., one statement of a known homeostasis misconception has only 17% of seniors who can answer it correctly), these statements influence overall scores in individual concept categories. Thus, for the richness of information about student thinking, it will be important to look at the individual statements in addition to overall scores.

Transparency. In the context of other program assessments, Phys-MAPS offers a unique level of transparency to instructors administering the assessment. Unlike other program-level assessments (e.g., HAPS Comprehensive Exam) that do not allow users to see the assessment items and only offer a summarized score report, instructors who administer Phys-MAPS will be able to know specifically what is being asked of students and what scores are for every statement. As a major goal of all Bio-MAPS assessments, to aid in a data-driven reform of biology curricula, we wanted to provide as much specific information about student thinking as possible to instructors and departments.

In addition to the three features mentioned above, we also examined Phys-MAPS for evidence of biases. Although we took care to look for statement bias during the design process, made changes accordingly when biases were found, and have no evidence of bias for individual statements regarding race/ethnicity, first-generation status, transfer status, whether English was spoken at home, or sex, we do have evidence of a gender bias for the Phys-MAPS assessment overall, where men perform better than other students. This gender effect is similar to those in other studies and may be due to the closed-response format of the test (e.g., Refs. 17, 50).

Limitations

Two particular elements of Phys-MAPS design introduce limitations, namely its complexity and breadth. First, the complexity of the physiological systems in Phys-MAPS questions requires constraining them such that there are clear, correct answers. This makes each one of Phys-MAPS questions relatively complex to read. This complexity makes the length of time required to complete the assessment relatively longer than for other concept assessments: generally 30–40 min. As student engagement in lengthy assessments has been found to wane over time (54) and the effort students put into the Phys-MAPS correlated with their scores (Table 7), users are encouraged to think of incentives that will encourage students to put in serious effort. In addition, monitoring both the length of time it takes students to complete the assessment and students' effort on the assessment can provide data on whether students took the assessment seriously (51).

Second, we necessarily had to sacrifice some depth for breadth to assess all major concepts in the Vision and Change and Core Principles frameworks. When possible, we attempted to use published work to align Phys-MAPS with subconcepts that are most valued by physiologists. Three of the top-ranked concepts in the Core Principles framework (gradients, cell-cell communication, homeostasis) had been previously unpacked through rigorous faculty feedback on the content and relative importance of the various subconcepts (28, 33, 35). Thus, while Phys-MAPS may not test all subconcepts, we believe it assesses what is considered to be most important by practitioners. To more fully understand why students may struggle with any singular concept, it will be necessary to follow up Phys-MAPS administration with other assessments that address single concepts.

Recommended Use

Undergraduate programmatic assessment. As physiology programs move toward defining and assessing what students should and do learn in their programs (55), Phys-MAPS can be used to measure students' ability to integrate and apply their knowledge on the core concepts of physiology across time in a major/program. If the assessment is administered each year to seniors, the outcomes can help provide a longitudinal measure for accreditation agencies, deans, department chairs, and faculty to understand the level of expertise of graduating students. As there is an increase in demand for departments to have such programmatic assessments (16, 30), tools such as Phys-MAPS will likely become increasingly important. Furthermore, assessment outcomes could help education researchers answer broad questions, such as how departments can best structure an undergraduate curriculum to maximize student learning of challenging concepts.

Medical and other health programs may also find this assessment useful to probe first-year students' conceptual understanding. Whereas Phys-MAPS is not solely based in human physiology, it is highly conceptually relevant to the field, and first-year medical students often have naive conceptions of the most difficult concepts on Phys-MAPS, such as homeostasis (e.g., Ref. 7).

Table 11. *Phys-MAPS recommendations for administration and student recruitment*

1. Identify course and/or time points to administer Phys-MAPS. For example, plan for different cohorts of students to take the assessment when they begin their introductory courses, at the end of their introductory course series, and upon graduation.
2. Contact the corresponding author for the freely available web-based assessment and automatic scoring template.
3. When administering the survey we recommend:
a. Using the online Qualtrics survey platform.
b. Giving students 1 wk to complete the survey.
c. Awarding low-stakes incentives for completion (e.g., participation points if associated with a course).
d. Including the SOS effort survey (54) along side demographic variables.
4. Input student responses into automatic-scoring template provided by the corresponding author. For each administration, you will receive:
a. The mean, median, and range of student scores for the assessment overall and for each of Vision and Change Core Concept and each of the Core Principles of Physiology.
b. The percent correct for each statement on the assessment.
5. Identify concepts that students understand and struggle with at your institution. Identify specific concepts and/or conceptual difficulties for targeted instruction and curriculum redesign. Consult the education literature for deeper understanding of student thinking, targeted concept inventories, and evidence-based teaching strategies. (e.g., Refs. 31, 36).

Phys-MAPS, Measuring Achievement and Progress in Science in Physiology; SOS, student opinion survey.

Administration format. Our recommendation is that Phys-MAPS be administered pre- and postprogram (rather than pre- and postcourse), online, and with adequate student incentives (Table 11). For the purposes of program assessment, we suggest assessing students at time points at which they will have similar curricular experiences. We also recommend an online format for the ease of delivering the assessment to students outside of individual courses. Previous exploration of the Molecular Biology Capstone Assessment found no difference between in-class and at-home performance (11). When using an online format, we recommend that questions be randomized, but the order of statements within each question is kept constant, as this is how the assessment was validated.

To provide online access and administration of the Phys-MAPS, we have established a portal system (located at http://cperl.lassp.cornell.edu/bio-maps) through which any instructor wishing to administer the Phys-MAPS can register his/her course and provide students the opportunity to take the assessment online. (Students will take the assessment via a unique Qualtrics link generated by the portal, and the instructor will receive automated score reports after the administration has closed.) In addition, users conducting research can obtain more information about access to data by contacting the administrators of the portal system.

Finally, we encourage finding appropriate incentives that result in high-quality data without sacrificing assessment integrity. If the assessment were to become high stakes, then students might be encouraged to post questions on the internet and use outside resources to answer questions, which will result in scores that do not accurately represent a student's understanding of physiology. However, if there are no incentives for students, one is likely to get low participation and/or high guess rates.

Once a program has measured student performance on Phys-MAPS, we suggest using more tailored concept inventories to follow up at the course level. For example, Phys-MAPS data suggest that students at all levels struggle with the concept of homeostasis. To look more closely at student thinking on this concept, faculty could administer the newly developed homeostasis concept inventory (27) on a course-by-course basis and then design curricular materials to improve student learning. While not all concepts covered by Phys-MAPS are addressed in more detail by a course-level concept assessment, authors have developed conceptual frameworks for many of the Core Principles, which can help identify places to start identifying changes to instructional practices (28, 36, 37).

Availability. Phys-MAPS, along with all of the Bio-MAPS assessments (GenBio-MAPS, EcoEvo-MAPS, Molecular and Cellular Capstone Assessment), are freely available by contacting the corresponding author and/or visiting the website mentioned above. When contacting the corresponding author, we will provide a PDF copy of Phys-MAPS and the link to the online portal administration for Phys-MAPS.

ACKNOWLEDGMENTS

We thank all of the faculty members who provided feedback on the Phys-MAPS assessment; faculty members who piloted Phys-MAPS, students who participated in the interviews and pilots; Dr. Yi Zheng and Dr. Jennifer Avena for advice on statistical analyses; Dr. Leif Saul for original artwork in question figures; Roshanne Ebrahimian and Giovanni Venzor Melendez for help with manuscript preparation; and the 19 Institutional Review Board committees who considered our requests to pilot.

Present address of M. K. Smith: Dept. of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY 14853.

GRANTS

This work was supported by a National Science Foundation Transforming Undergraduate Education in Science, Technology, Engineering and Mathematics award (TUES) program to [DUE-1322364 (Colorado), DUE-1322556 (Maine), and DUE-1323010 (Washington).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

K.S. and C.D.W. performed experiments; K.S., B.A.C., M.K.S., M.M.S., and J.K.K. analyzed data; K.S., S.E.B., B.A.C., A.J.C., M.K.S., M.M.S., C.D.W., and J.K.K. interpreted results of experiments; K.S. prepared figures; K.S. drafted manuscript; K.S., S.E.B., B.A.C., A.J.C., M.K.S., M.M.S., C.D.W., and J.K.K. edited and revised manuscript; K.S., S.E.B., B.A.C., A.J.C., M.K.S., M.M.S., C.D.W., and J.K.K. approved final version of manuscript.

REFERENCES

1. **Adams WK, Wieman CE.** Development and validation of instruments to measure learning of expert-like thinking. *Int J Sci Educ* 33: 1289–1312, 2011. doi:10.1080/09500693.2010.512369.
2. **Auerbach AJ, Schussler EE.** Curriculum alignment with vision and change improves student scientific literacy. *CBE Life Sci Educ* 16: ar29, 2017. doi:10.1187/cbe.16-04-0160.
3. **American Association for the Advancement of Science.** *Vision and Change in Undergraduate Biology Education: A Call to Action: Final Report of a National Conference*. Washington, DC: AAAS, 2011.
4. **American Association for the Advancement of Science.** *Vision and Change in Undergraduate Biology Education: Chronicling Change, Inspiring the Future*. Washington, DC: AAAS, 2015.
5. **Anderson DL, Fisher KM, Norman JG.** Development and evaluation of the conceptual inventory of natural selection. *J Res Sci Teach* 39: 952–978, 2002. doi:10.1002/tea.10053.
6. **Anderson LW, Krathwohl DR, Bloom BS (Editors).** *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. San Francisco, CA: Addison Wesley Longman, 2001.
7. **Badenhorst E, Mamede S, Abrahams A, Bugarith K, Friedling J, Gunston G, Kelly-Laubscher R, Schmidt HG.** First-year medical students' naïve beliefs about respiratory physiology. *Adv Physiol Educ* 40: 342–348, 2016. doi:10.1152/advan.00193.2015.
8. **Baker FB.** *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker, 1992.
9. **Brownell SE, Freeman S, Wenderoth MP, Crowe AJ.** BioCore Guide: a tool for interpreting the core concepts of vision and change for biology majors. *CBE Life Sci Educ* 13: 200–211, 2014. doi:10.1187/cbe.13-12-0233.
10. **Chalmers RP.** Mirt: a multidimensional item response theory package for the R environment. *J Stat Softw* 48: 1–29, 2012. doi:10.18637/jss.v048.i06.
11. **Couch BA, Knight JK.** A comparison of two low-stakes methods for administering a program-level biology concept assessment. *J Microbiol Biol Educ* 16: 178–185, 2015. doi:10.1128/jmbe.v16i2.953.
12. **Couch BA, Wood WB, Knight JK.** The molecular biology capstone assessment: a concept assessment for upper-division molecular biology students. *CBE Life Sci Educ* 14: ar10, 2015. doi:10.1187/cbe.14-04-0071.
13. **Couch BA, Hubbard JK, Brassil CE.** Multiple-true-false questions reveal limitations of the multiple-choice format for detecting students with mixed and partial conceptions. *Bioscience* 68: 455–463, 2018. doi:10.1093/biosci/biy037.
14. **Cliff WH.** Case study analysis and the remediation of misconceptions about respiratory physiology. *Adv Physiol Educ* 30: 215–223, 2006. doi:10.1152/advan.00002.2006.
15. **D'Avanzo C.** Biology concept inventories: overview, status, and next steps. *Bioscience* 58: 1079–1085, 2008. doi:10.1641/B581111.
16. **Dirks C, Knight JK.** Measuring College Learning in Biology. In: *Improving Quality in American Higher Education: Learning Outcomes and Assessments for the 21st Century*, edited by Arum R, Roksa J, Cook A. San Francisco, CA: Jossey-Bass, 2016, p. 225–260.
17. **Eddy SL, Brownell SE, Wenderoth MP.** Gender gaps in achievement and participation in multiple introductory biology classrooms. *CBE Life Sci Educ* 13: 478–492, 2014. doi:10.1187/cbe.13-10-0204.
18. **Frisbie DA, Sweeney DC.** The relative merits of multiple true-false achievement tests. *J Educ Meas* 19: 29–35, 1982. doi:10.1111/j.1745-3984.1982.tb00112.x.
19. **Handelsman J, Miller S, Pfund C.** *Scientific Teaching*. London: Macmillan, 2007.
20. **Hambleton RK, Swaminathan H, Rogers HJ.** *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage, 1991.
21. **Hubbard JK, Potts MA, Couch BA.** How question types reveal student thinking: An experimental comparison of multiple-true-false and free-response formats. *CBE Life Sci Educ* 16: ar26, 2017. doi:10.1187/cbe.16-12-0339.
22. **Jittivadhna K, Ruenwongsa P, Panijpan B.** Hand-held model of a sarcomere to illustrate the sliding filament mechanism in muscle contraction. *Adv Physiol Educ* 33: 297–301, 2009. doi:10.1152/advan.00036.2009.
23. **Jodoin MG, Gierl MJ.** Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Appl Meas Educ* 14: 329–349, 2001. doi:10.1207/S15324818AME1404_2.
24. **Knight JK.** Biology concept assessment tools: design and use. *Microbiol Aust* 31: 5–8, 2010.
25. **Kubinger KD, Gottschall CH.** Item difficulty of multiple choice tests dependent on different item response formats–an experiment in fundamental research on psychological assessment. *Psychol Sci* 49: 361, 2007.
26. **Martinková P, Drabinová A, Liaw Y-L, Sanders EA, McFarland JL, Price RM.** Checking equity: why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE Life Sci Educ* 16: rm2, 2017. doi:10.1187/cbe.16-10-0307.
27. **McFarland JL, Price RM, Wenderoth MP, Martinková P, Cliff W, Michael J, Modell H, Wright A.** Development and validation of the homeostasis concept inventory. *CBE Life Sci Educ* 16: ar35, 2017. doi:10.1187/cbe.16-10-0305.
28. **McFarland J, Wenderoth MP, Michael J, Cliff W, Wright A, Modell H.** A conceptual framework for homeostasis: development and validation. *Adv Physiol Educ* 40: 213–222, 2016. doi:10.1152/advan.00103.2015.

29. **McLaughlin J, Metz A.** Vision & change: why it matters. *Am Biol Teach* 78: 456–462, 2016. doi:10.1525/abt.2016.78.6.456.
30. **Middaugh MF, Nelson D, Damminger JK.** *Planning and Assessment in Higher Education: Demonstrating Institutional Effectiveness*. San Francisco, CA: Jossey-Bass, 2011.
31. **Michael J.** What makes physiology hard for students to learn? Results of a faculty survey. *Adv Physiol Educ* 31: 34–40, 2007. doi:10.1152/advan.00057.2006.
32. **Michael JA, Wenderoth MP, Modell HI, Cliff W, Horwitz B, McHale P, Richardson D, Silverthorn D, Williams S, Whitescarver S.** Undergraduates' understanding of cardiovascular phenomena. *Adv Physiol Educ* 26: 72–84, 2002. doi:10.1152/advan.00002.2002.
33. **Michael J, McFarland J.** The core principles ("big ideas") of physiology: results of faculty surveys. *Adv Physiol Educ* 35: 336–341, 2011. doi:10.1152/advan.00004.2011.
34. **Michael JA, Richardson D, Rovick A, Modell H, Bruce D, Horwitz B, Hudson M, Silverthorn D, Whitescarver S, Williams S.** Undergraduate students' misconceptions about respiratory physiology. *Adv Physiol Educ* 277: S127–S135, 1999. doi:10.1152/advances.1999.277.6.S127.
35. **Michael J, Martinkova P, McFarland J, Wright A, Cliff W, Modell H, Wenderoth MP.** Validating a conceptual framework for the core concept of "cell-cell communication". *Adv Physiol Educ* 41: 260–265, 2017. doi:10.1152/advan.00100.2016.
36. **Michael J, Cliff W, McFarland J, Modell H, Wright A.** *The Core Concepts of Physiology: A New Paradigm for Teaching Physiology*. New York: Springer, 2017. doi:10.1007/978-1-4939-6909-8.
37. **Modell H, Cliff W, Michael J, McFarland J, Wenderoth MP, Wright A.** A physiologist's view of homeostasis. *Adv Physiol Educ* 39: 259–266, 2015. doi:10.1152/advan.00107.2015.
38. **Montagna E, de Azevedo AMS, Romano C, Ranvaud R.** What is transmitted in "synaptic transmission"? *Adv Physiol Educ* 34: 115–116, 2010. doi:10.1152/advan.00006.2010.
39. **Next Generation Science Standards Lead States.** *Next Generation Science Standards: for States, by States*. Washington, DC: National Academies, 2013.
40. **Nehm RH, Reilly L.** Biology majors' knowledge and misconceptions of natural selection. *Bioscience* 57: 263–272, 2007. doi:10.1641/B570311.
41. **Nehm RH, Schonfeld IS.** Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. *J Res Sci Teach* 45: 1131–1160, 2008. doi:10.1002/tea.20251.
42. **Odom AL, Barrow LH.** Development and application of a two-tier diagnostic test measuring college biology students' understanding of diffusion and osmosis after a course of instruction. *J Res Sci Teach* 32: 45–61, 1995. doi:10.1002/tea.3660320106.
43. **Pelaez NJ, Boyd DD, Rojas JB, Hoover MA.** Prevalence of blood circulation misconceptions among prospective elementary teachers. *Adv Physiol Educ* 29: 172–181, 2005. doi:10.1152/advan.00022.2004.
44. **Price RM, Andrews TC, McElhinny TL, Mead LS, Abraham JK, Thanukos A, Perez KE.** The genetic drift inventory: a tool for measuring what advanced undergraduates have mastered about genetic drift. *CBE Life Sci Educ* 13: 65–75, 2014. doi:10.1187/cbe.13-08-0159.
45. **Perez KE, Hiatt A, Davis GK, Trujillo C, French DP, Terry M, Price RM.** The EvoDevoCI: a concept inventory for gauging students' understanding of evolutionary developmental biology. *CBE Life Sci Educ* 12: 665–675, 2013. doi:10.1187/cbe.13-04-0079.
46. **R Studio Team.** *RStudio: Integrated Development for R* (Online). Boston, MA: RStudio. https://www.rstudio.com/ [1 Jan 2015].
47. **Shi J, Wood WB, Martin JM, Guild NA, Vicens Q, Knight JK.** A diagnostic assessment for introductory molecular and cell biology. *CBE Life Sci Educ* 9: 453–461, 2010. doi:10.1187/cbe.10-04-0055.
48. **Smith MK, Wood WB, Knight JK.** The Genetics Concept Assessment: a new concept inventory for gauging student understanding of genetics. *CBE Life Sci Educ* 7: 422–430, 2008. doi:10.1187/cbe.08-08-0045.
49. **Smith MK, Knight JK.** Using the Genetics Concept Assessment to document persistent conceptual difficulties in undergraduate genetics courses. *Genetics* 191: 21–32, 2012. doi:10.1534/genetics.111.137810.
50. **Stanger-Hall KF.** Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. *CBE Life Sci Educ* 11: 294–306, 2012. doi:10.1187/cbe.11-11-0100.
51. **Steedle JT.** Motivation filtering on a multi-institution assessment of general college outcomes. *Appl Meas Educ* 27: 58–76, 2014. doi:10.1080/08957347.2013.853072.
52. **Summers MM, Couch BA, Knight JK, Brownell SE, Crowe AJ, Semsar K, Wright CD, Smith MK.** EcoEvo-MAPS: an ecology and evolution assessment for introductory through advanced undergraduates. *CBE Life Sci Educ* 17: ar18, 2018. doi:10.1187/cbe.17-02-0037.
54. **Thelk AD, Dundre DL, Horst SJ, Finney SJ.** Motivation matters: using the student opinion scale to make valid inferences about student performance. *J Gen Educ* 58: 129–151, 2009. doi:10.1353/jge.0.0047.
55. **VanRyn VS, Poteracki JM, Wehrwein EA.** Physiology undergraduate degree requirements in the U.S. *Adv Physiol Educ* 41: 572–577, 2017. doi:10.1152/advan.00104.2016.
56. **Vincent-Ruz P, Schunn CD.** The increasingly important role of science competency beliefs for science learning in girls. *J Res Sci Teach* 54: 790–822, 2017. doi:10.1002/tea.21387.
57. **Vosniadou S.** Bridging culture with cognition: a commentary on "culturing conceptions: from first principles". *Cult Stud Sci Educ* 3: 277–282, 2008. doi:10.1007/s11422-008-9098-9.
58. **Wiggins GP, McTighe J.** *Understanding by Design* (2nd Ed.). Alexandria, VA: Association for Supervision and Curriculum Development, 2005.
59. **Wilcox BR, Pollock SJ.** Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics. *Phys Rev Spec Top Phys Educ Res* 10: 020124, 2014. doi:10.1103/PhysRevST-PER.10.020124.
60. **Wilson CD, Anderson CW, Heidemann M, Merrill JE, Merritt BW, Richmond G, Sibley DF, Parker JM.** Assessing students' ability to trace matter in dynamic systems in cell biology. *CBE Life Sci Educ* 5: 323–331, 2006. doi:10.1187/cbe.06-02-0142.
61. **Witt E; Witt Measurement Consulting and the HAPS Testing Task Force.** *Psychometric Evaluation of the HAPS Anatomy & Physiology Comprehensive Exam*. LaGrange, GA: HAPS, 2017.
62. **Zumbo BD, Thomas DR.** *A Measure of Effect Size for a Model-Based Approach for Studying DIF*. Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science, 1997.