*Adv Physiol Educ* 43: 211–220, 2019;
doi:10.1152/advan.00209.2018.

**HOW WE TEACH** | *Generalizable Education Research*

# A new assessment to monitor student performance in introductory neurophysiology: Electrochemical Gradients Assessment Device

**Jack A. Cerchiara,[1] Kerry J. Kim,[2] Eli Meir,[2] Mary Pat Wenderoth,[1] and Jennifer H. Doherty[1]**
[1]*Department of Biology, University of Washington, Seattle, Washington; and* [2]*SimBiotic Software, Missoula, Montana*

Submitted 11 December 2018; accepted in final form 9 April 2019

**Cerchiara JA, Kim KJ, Meir E, Wenderoth MP, Doherty JH.** A new assessment to monitor student performance in introductory neurophysiology: Electrochemical Gradients Assessment Device. *Adv Physiol Educ* 43: 211–220, 2019; doi:10.1152/advan.00209.2018.—The basis for understanding neurophysiology is understanding ion movement across cell membranes. Students in introductory courses recognize ion concentration gradients as a driving force for ion movement but struggle to simultaneously account for electrical charge gradients. We developed a 17-multiple-choice item assessment of students' understanding of electrochemical gradients and resistance in neurophysiology, the Electrochemical Gradients Assessment Device (EGAD). We investigated the internal evidence validity of the assessment by analyzing item characteristic curves of score probability and student ability for each question, and a Wright map of student scores and ability. We used linear mixed-effect regression to test student performance and ability. Our assessment discriminated students with average ability (weighted likelihood estimate: $-2$ to 1.5 $\Theta$); however, it was not as effective at discriminating students at the highest ability (weighted likelihood estimate: $>2$ $\Theta$). We determined the assessment could capture changes in both assessment scores (model $r^2 = 0.51$, $P < 0.001$, $n = 444$) and ability estimates (model $r^2 = 0.47$, $P < 0.001$, $n = 444$) after a simulation-based laboratory and course instruction for 222 students. Differential item function analysis determined that each item on the assessment performed equitably for all students, regardless of gender, race/ethnicity, or economic status. Overall, we found that men scored higher ($r^2 = 0.51$, $P = 0.014$, $n = 444$) and had higher ability scores ($P = 0.003$) on the EGAD assessment. Caucasian students of both genders were positively correlated with score ($r^2 = 0.51$, $P < 0.001$, $n = 444$) and ability ($r^2 = 0.47$, $P < 0.001$, $n = 444$). Based on the evidence gathered through our analyses, the scores obtained from the EGAD can distinguish between levels of content knowledge on neurophysiology principles for students in introductory physiology courses.

assessment; college biology; electrophysiology; flux; physiology

## INTRODUCTION

Neurophysiology is a challenging topic for students to learn (17, 29, 42). The basis for mastering neurophysiology is understanding how ions move across cell membranes and the impact ion movement has on membrane potential. Understanding the movement of ions means being able to integrate both the chemical and electrical driving forces, as well as account for resistance to ion movement. While students recognize ion chemical gradients as a driving force for ion movement, they struggle to simultaneously account for electrical forces acting on the ion (42). This struggle is exemplified in one study that found that, even after a semester of instruction, students entering upper level courses had only a superficial understanding of membrane potential (42). Students also have trouble explaining how ion movement impacts the specific cell functions of generation and propagation of graded and action potentials and synaptic signaling (17).

Instructors must go beyond simple instruction to help students learn and gain mastery of these neurophysiology concepts. Rather, faculty need to provide students with opportunities to explore and practice working with these difficult concepts (16, 30, 32–35). Many instructors have taken on this challenge by creating a wide variety of active-learning instructional materials to teach neurophysiology (8, 9, 24). These innovative tools may indeed help students gain a deeper understanding of ion movement and membrane potentials. However, instructors need a way to empirically test the efficacy of each of these tools in their own classrooms.

One way to assess the efficacy of innovative teaching methods is to use validated assessments published in the primary literature. In the field of physics, the Force Concept Inventory (19) has been used for over 20 yr to show that active learning and physics tutorials worked better than traditional lecture to improve student understanding of the concept of force in physics. In biology, similar concept inventories have been developed for the study of evolution, particularly the Concept Inventory for Natural Selection, as well as the concept of genetic drift in the Genetic Drift Concept Inventory (38, 39). In the field of physiology, there are currently two validated assessments, the Osmosis and Diffusion Concept Assessment (11) and the Homeostasis Concept Inventory (28). In neurophysiology, however, there are currently no tools of which we are aware to assess student learning and the efficacy of learning interventions.

Concept inventories are usually in a multiple-choice format. When creating multiple-choice questions, it is important to realize the limitations of this format. Previous research has shown that students often can pick a correct answer without having the correct underlying reasoning (11, 14, 18). This results in sending a "false positive" message to both the student and the instructor, as a correct answer implies that the student understands why the answer is correct. To guard against this misinterpretation of student understanding, it can be helpful to use a follow-up constructed response question, asking the students to state their reasoning for their previous answer (22). While constructed response items are ideal, they are also labor

Address for reprint requests and other correspondence: J. A. Cerchiara, Dept. of Biology, University of Washington, Hitchcock 430, Box 351800, Seattle, WA 98195-1800 (e-mail: jackc44@uw.edu).

intensive to score. Multiple true/false formats have shown promise in elucidating student thinking (20, 40). In multiple true/false questions, rather than selecting the "most correct" answer choice as in traditional multiple choice, students will evaluate each potential answer choice and determine whether it is true or false (20, 40). As an alternative, some concept inventories use pairs of questions in two-tiered, multiple-choice format (11, 14, 31). The first question in each pair often asks "what" will happen in the scenario presented in the stem. The follow-up item asks the "how'" question, where students are asked to choose the reason for their prediction. This two-tiered design assesses student understanding of both "what" will happen and "how" it happens.

The current best practices that govern the development of learning assessments suggest five types of validity evidence be used to construct a validity argument [American Educational Research Association/American Psychological Association/National Council on Measurement in Education (AERA/APA/NCME) Standards] (2). Three of these types of evidence are most relevant to assessments in the college classroom: *1*) that faculty agree the questions align with appropriate learning objectives; *2*) that students understand and can provide answers to the questions asked; and *3*) the assessment differentiates students based on their performance, and it performs equitably for all students. The first type is "evidence based on test content." This evidence encompasses logical analysis and expert judgment of the alignment of the assessment questions to the technical content of the discipline. The second type is "evidence based on response process." This evidence is collected by interviewing students and is meant to answer the questions: *1*) Is there evidence that students interpret the questions in the same way that the writers of the question intended? and *2*) Do the questions elicit the intended range of student responses? The third type is "evidence based on internal structure" of the assessment. This evidence is collected using a set of statistical analyses to determine whether the internal components of the assessment are free of bias and measure variation in student ability.

A variety of statistical tests can be used to collect evidence for internal structure. We use item response theory (IRT) models, specifically Rasch-family models (15). The use of such models allows us to empirically evaluate the relationship between student performance on assessment items, and the overall student ability distribution, as well as to evaluate the fit of the model to the data at the individual item and student level. Student ability, in IRT modeling, is not a measure of inherent academic ability or potential achievement, but an estimate of their ability to comprehend the content covered in the assessment. Tests using IRT models to confirm the assessment can differentiate students with a wide range of abilities, including, but not limited to, the following: item-person (Wright) mapping, which visualizes assessment item difficulty and student ability on the same scale; item characteristic curves (ICCs), which visualize how well assessment items differentiate students by ability; and test item function (TIF), which visualizes what proportion of the student population the assessment is most informative. The test used to determine whether the assessment is equitable across all student subgroups is a differential item function (DIF) analysis, which determines whether any individual test item is biased against any particular group of students. For a primer on assessment development and validation evidence written for faculty, see Bass et al. (5). A good overview on how the statistical tests are used in assessment creation and how validation evidence is evaluated is presented in Gómez-Benito et al. (15) and Huggins-Manley (21).

In this study, we present how we developed and provide validity evidence for the Electrochemical Gradient Assessment Device (EGAD) for students' understanding of ion movement and membrane potentials at the introductory college level. We provide the full EGAD (Supplemental Data are available online at dx.doi.org/10.17504/protocols.io.z4ff8tn) as another instrument to add to the growing number of conceptual assessments in biology. The EGAD can be used to assess instructional effectiveness at both the curricular and programmatic levels (46).

## METHODS

We developed the EGAD for introductory electrophysiology through multiple rounds of revision and using the AERA/APA/NCME guidelines for collecting validity evidence (2). Two physiologists on the project team (J.H.D., M.P.W.) initially drafted learning objectives and assessment questions. We then proceeded with an iterative process of question revision and validity evidence collection in consultation with a community of physiologists, education researchers, and students (Table 1). We used statistical analysis to collect validity

Table 1. *Iterative development of EGAD*

*Iteration 1*: 23 multiple-choice and 23 constructed response questions
- Drafted initial set of questions (J.H.D., M.P.W.)
- Collected student answers via online assessment (550 students)
- Revised questions

*Iteration 2*: 16 multiple-choice and 2 constructed response questions
- Content validity analysis. Internal expert review: two physiology education researchers (J.H.D., M.P.W.), two discipline-based education researchers (E.M., D.P.), one neurobiologist (K.J.K.)
- Content validity analysis (external expert review): four anonymous physiology faculty
- Collected student answers via online assessment (100 students)
- Revised questions

*Iteration 3*: 16 multiple-choice and 2 constructed response questions
- Response processes analysis: think-aloud interviews (9 students)
- Revised questions

Final EGAD: 17 multiple-choice questions
- Collected student answers via online assessment (227 students)
- Internal structure analysis

EGAD, Electrochemical Gradients Assessment Device.

evidence for internal structure. Through this process, we created and tested four versions of the EGAD. Below, we describe in greater detail the steps taken to develop the EGAD and then report on the final version.

*Iteration 1.* Before assessment development, we established learning objectives for understanding of ion movement and membrane potentials at the introductory college level (Table 2). In general, these learning objectives focused on a student's mastery of the concept of ion flux (the relationship between concentration gradient, electrical driving force, and resistance) and the impact on membrane potential. We wrote these learning objectives from our own classroom experience teaching undergraduate physiology. To confirm these learning objectives were appropriate for undergraduate physiology, we conducted interviews with 10 physiology faculty across a diversity of institution types. All 10 faculty concluded that these learning objectives were aligned with their in-course instruction and are attainable goals for physiology undergraduate students.

Initially, we wrote 23 questions. Each assessment question was written as both a multiple-choice and a constructed response. We gave questions in the multiple-choice format to test the distractors we created from our own classroom experience. We gave questions in the constructed response format to require students to express their reasoning in their own words. This allowed us to incorporate this student wording on future item distractors.

Of the 23 questions, 9 pairs (18 questions) were in the two-tier format. The five remaining questions were stand-alone, multiple-choice questions, as those topics were not conducive to the "what" and "how" format (i.e., we had to tell the students the "what" in order for them to choose the "how"). As it is fundamental that students understand diffusion to understand ion movement, we modified a pair of questions from the Osmosis and Diffusion Concept Assessment (15, 21) to align with learning *objective 1*.

One example of a two-tiered assessment item from the final EGAD is shown in Fig. 1. This item assessed students' understanding of changes in membrane potential observed in the action potential recording. Students were provided an illustration of a standard action potential and a schematic diagram of an axon identifying membrane ion channels. The initial question asked "what" was responsible for the changes in membrane potential. The second question asked the "how," which was designed to uncover the students' reasoning that supported their answer to the "what" question.

To keep the assessment to ~20 min, we distributed the 23 questions randomly across four assessment forms. Each form contained a subset of 13 or 14 questions. Each subset consisted of a mix of the multiple-choice and the constructed response versions of unique items. For example, for the question shown in Fig. 1, some students got the multiple-choice format, whereas others got the constructed response format. We administered the forms online to 550 students in a large-enrollment Introductory Plant and Animal Physiology course. The assessment was administered at two time points, at the beginning and end of the week's instruction, to capture novice and more sophisticated student reasoning.

Based on the results from this first assessment round, we dropped five questions to shorten the assessment. We dropped four questions because results showed nearly 100% of students were answering them correctly, and we already had more challenging questions that addressed the same topic. We dropped one question because, upon review, it was not aligned with our learning objectives. Furthermore, we used the constructed response answers to refine the incorrect answers to better model students' alternative reasoning and misconceptions. As a result of these changes, the second iteration of the assessment had a total of 18 questions: 2 constructed response questions and 16 multiple-choice questions. The 16 multiple-choice questions consisted of 6 pairs of two-tiered questions and 4 single questions. We kept two questions in the constructed response format because we observed that, in the multiple-choice format, students of all abilities could recognize the correct answer, but only high-performing students could answer in the constructed response format correctly, and we wished to capture this high-level reasoning.

*Iteration 2.* First, we used internal expert review (researchers involved with the project) to collect evidence on test content (2). These internal experts, the authors of the paper [two physiology education researchers (M.P.W., J.H.D.), one discipline-based education researchers (E.M.), and one neurobiologist (K.K.)] and an additional discipline-based education researcher (D.P.) reviewed each item for wording, content accuracy, alignment with learning objectives, student misconceptions, and correctness of answer choices. Experts were given the complete assessment and asked for comments. These included tracked changes of the wording, additional answer choices, and open comments on content and alignment to learning objectives.

We next asked four faculty who teach introductory neurobiology and who were external to the project and anonymous to the authors (except E.M.) to review the assessment. Four faculty scored scientific accuracy, clarity, and relevance to learning objectives on a 5-point Likert scale. Three provided tracked-changes wording and open comments. These results allowed us to confirm items were accurate, clear, and relevant to the stated learning objectives. These external expert faculty also provided minor changes in wording.

Based on these reviews, we again revised the wording of some questions, including changing the format of one item to a "choose all that apply" question, as more than one answer was correct. We administered the revised set of questions at two time points to 100 students in a new offering of the Introductory Plant and Animal Physiology course. The results from administration of *iteration 2* of the assessment informed the final revision.

*Iteration 3.* We collected evidence based on response processes (2) by conducting think-aloud interviews on these revised items with nine students from both lower and upper division biology courses. In these interviews, students were asked to answer the assessment questions

Table 2. *Learning objectives and alignment of the questions of the final version of EGAD*

| Learning Objective | Assessment Item |
|---|---|
| *1*. Explain that diffusion is the net movement of particles from a region of high concentration to a region of low concentration, based on the random motion of particles. | Q1/Q2 |
| *2*. Predict and explain ion movement in response to diffusion and electric forces, including application of the concept of the equilibrium potential (membrane potential at which diffusion and electric forces are in balanced opposition, resulting in no net flow of ions). | Q3/Q4, Q5/Q6, Q7/Q8, Q9/Q10 |
| *3*. Predict and explain changes in membrane potential based on membrane conductance to $Na^+$ and $K^+$ and ion movement. | Q9/Q10, Q13/Q14 |
| *4*. Predict and explain changes in membrane potential based on concentrations of $Na^+$ and $K^+$ and ion movement. | Q11/Q12 |
| *5*. Explain that transmission between neurons usually uses chemical neurotransmitter. | Q15 |
| *6*. Explain that stimulus intensity determines action potential firing rate. | Q16 |
| *7*. Explain that neurons have stimulus thresholds for producing action potentials. | Q17 |

Q1–Q17, *questions 1–17*.

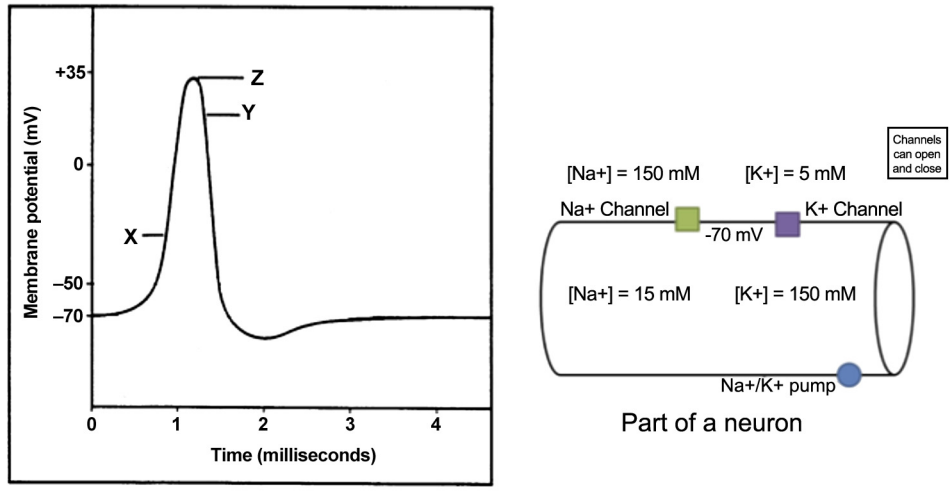


Fig. 1. Two-tiered assessment item on Electrochemical Gradients Assessment Device (EGAD). The initial "what" question asked was: What was responsible for the changes in membrane potential? The item's second question assessed student's reasoning.

*What causes the membrane potential to be more positive than -70 mV at the point labeled X?*

a. Na+ and K+ channels are open and Na+ and K+ ions are rushing into the neuron.
b. Na+ channels are open and Na+ ions are rushing into the neuron.
c. K+ channels are open and K+ ions are rushing into the neuron.

*Those ions are moving into the neuron because*

a. by coming in they'll cause the membrane potential to become more positive.
b. the neuron needs to generate a more positive membrane potential.
c. of electric forces.
d. of the concentration gradient.
e. of the concentration gradient and the neuron needs to generate a more positive membrane potential.
f. of electric forces and the concentration gradient.

while they explained their reasoning about their answer choices. These interviews confirmed that the students were interpreting each question the way the question had been intended and did not omit necessary details. However, in a few cases, students reasoned in a way that was not captured by our answer choices. For example, in questions where students were asked to account for charges (+ or −) on ions [for example, *question 6* (Q6)], students told us they wanted to combine multiple-choice answers that described concentration gradient and ions "needing" or "wanting" to move to affect membrane potential. In these cases, we added an additional distractor that aligned with the reasoning presented in the interview. When additional distractors were developed, we confirmed symmetry in the choices. For example, in Q4, students tended to reason with an understanding of concentration gradient; however, they were also attracted to answer choices that invoked electrical gradient reasoning. For this reason, we added two additional distractors: one that identified only electrical gradient, and another that identified a combination of electrical and concentration gradient. The think-alouds with students also confirmed that the questions returned answers of variable sophistication, and that the assessment could capture the highest-level reasoning possible from this population.

Following student think-alouds, we conducted our final revision of the assessment. We removed the two constructed response questions for this final version. We thought the reasoning that the questions elicited was valuable, but the answers were too laborious to grade. We also split the "choose all that apply" question into two questions (Q13, Q14), as it was easier to grade in that format.

After these revisions, we now have a 17-item EGAD. Of the 17 multiple-choice items, there are six sets (12 total) of two-tiered questions, with Q1 of each pair asking the "what" and Q2 asking the "how" of the electrophysiology situation. Five questions are stand-alone multiple choice.

*Final EGAD.* We tested the final iteration of the 17-item EGAD in a large-enrollment, Introductory Level Animal and Plant Physiology

course with 257 students at the University of Washington, a large, R1 university. The course consisted of a daily lecture hour, taught with evidence-based teaching methods (12, 13), as well as one weekly laboratory session. Laboratories ran Tuesday through Thursday with 24 students per 3-h laboratory and two laboratory sections run concurrently. During the neurophysiology unit, students began assessment-related content on Monday. Beginning on Tuesday, students attended one weekly laboratory section. During this laboratory, students completed a computer simulation-based laboratory from SimBiotic Software called Action Potentials Extended (24), where they worked through interactive activities that support understanding of membrane potentials. The Action Potentials Module is available from SimBiotic Software at http://simbio.com. Students also worked in groups of four at white boards to answer instructor-generated mechanistic questions about the simulation, and groups reported out results to the larger laboratory section.

The EGAD assessment was identical at each time point and was administered online. The first assessment was completed on Sunday, 1 day before the start of the neurophysiology and action potential unit (lecture and laboratory). Students took the second assessment the evening after they completed their laboratory for that unit. To protect class time, we designed this survey to be taken online. We realize all online surveys are vulnerable to cheating, so we provided no incentive to cheat (i.e., only giving participation points, not points for correctness). Students received credit for completing the EGAD, but questions were not graded for correctness. Given that student laboratories occurred Tuesday through Thursday, students may have attended two to four lecture sessions before the second assessment. For this reason, the number of in-class sessions was included in the data analysis.

Reliability is the degree to which an assessment produces stable and consistent results. To determine reliability for the EGAD we used TIF, which determines whether the information gained from the assessment is consistent across student abilities, and Cronbach's $\alpha$, which determines whether the assessment is internally consistent. To collect evidence for internal structure validity, we used IRT and DIF analyses (27).

To determine whether the assessment captures the whole population of students and whether item difficulty corresponded to student abilities, we used IRT (4, 27). IRT estimates student ability and item difficulty and allows us to compare these two variables on the same scale. Student ability is an estimate derived from the likelihood function of the IRT model presented on a logit scale, or more generally, the likelihood of a student to correctly answer an item based upon his/her performance on all items on the assessment and the performance of the population. We used the weighted likelihood estimates (WLE) to estimate student ability. It is important to note that the ability estimate (WLE) does not measure inherent academic ability or potential achievement, but performance on this assessment, given the difficulty of the questions. Item difficulty is a measure of the probability of getting the question correct in relation to the probability of getting other questions correct and student performance on the assessment as a whole.

We fit a four-parameter IRT model (23) that allowed us to account for the potential that students could guess correct answers on the assessment and, therefore, score correct answers, which could artificially inflate ability estimates. We used these data to generate an item-person map [also called a Wright map (6, 36)]. The Wright map correlates a histogram of the students' WLE with the item difficulty. Difficulty of an item is defined as the 50% probability of a student answering the item correctly. We analyzed ICCs of score probability and student ability for each question to determine whether the questions have a range of difficulties and have sufficient discrimination.

It is vital that our assessment performs equitably for all students and that individual items are free of unintentional bias. We conducted a DIF analysis to determine whether any items were disproportionately more challenging to particular student demographic groups. We obtained students' college grade point average (GPA) and demographic data from the registrar. Demographic data included binary gender (female = 64%), participation in the University's Educational Opportunity Program (EOP; i.e., students identified as economically or educationally disadvantaged, 21%), and whether the student was from a race/ethnicity that is an underrepresented minority in science (URM) (i.e., African-American, Hispanic, Native American, or Pacific Islander, 31%). We conducted DIF analysis on these three binary groups (gender, EOP/non-EOP, and URM/non-URM) to determine whether there were significant differences in performance on individual items of the assessment.

We investigated whether the assessment, as a whole, performed equitably for all students, and whether it could capture differences in the ability between populations of students. We will refer to these as two different time points rather than pre- or postinstruction, as we did not use a control to test for the impact of instruction on learning. We removed individuals who did not take both assessments, leaving 444 responses from 222 students. To conduct this exploration, we used a mixed-effects linear regression model with student WLE as the response variable. As fixed effects, we included students' GPA, gender, EOP status, URM status, timing of the assessment, and the number of days of lecture between the two time points of the assessment, as well as the two-way interaction terms among all of these variables. We included student number as a random effect in the model, as the same students took the assessment at both time points. We conducted backward simplification using the "step" function in R with a maximum of 1,000 iterations (R version 3.1.3). The "step" package simplifies models based on Akaike information criteria (3). The function is conservative by nature, retaining $P$ values as high as 0.33. We selected the best fit model based on Akaike information criterion (AIC). For all statistical tests, we used R Statistical software (R Foundation for Statistical Computing: Development Core Team, version 3.1.3).

*Human subjects approval.* All procedures were conducted in accordance with approval from the Institutional Review Board at The University of Washington (no. HSD 00001316).

## RESULTS

Below we present validity evidence we collected for internal structure showing the EGAD is reliable, can measure variation in student ability, and is free from bias in undergraduate physiology students.

*Reliability.* We determined student ability estimate by calculating WLEs through IRT analysis. WLEs are estimated on a logit scale, which is the likelihood of a student to score correctly an item based on his/her performance on all items on the assessment and the performance of the population. Logit scale is the log-odds, or a logarithm of the odds (probability) of correctly answering an item. The TIF determined that our assessment-discriminated student scores with those of average ability ($-2$ to 1.5), however, not with students of the highest ability ($> 2$), suggesting some students had more sophisticated knowledge than was measured by the current assessment (Fig. 2). Cronbach's $\alpha$ was 0.76, with a 95% confidence interval of 0.73 and 0.78, which is in an acceptable range for reliability (37).

*Internal structure validity: items function across student abilities.* An ICC assessed each item's capacity to discriminate student answers of varying student ability (Fig. 3). Four items (Q4, Q6, Q8, Q10) had very steep curves, indicating that they have high probability of a correct answer for students at a tightly defined ability score ($-0.5$ to 0.5). Ten questions captured students from a middle range of WLE values ($-2$ to 2), but did not discriminate at a distinct ability. Three items (Q1, particles diffuse from high to low concentration; Q12, $K^+$

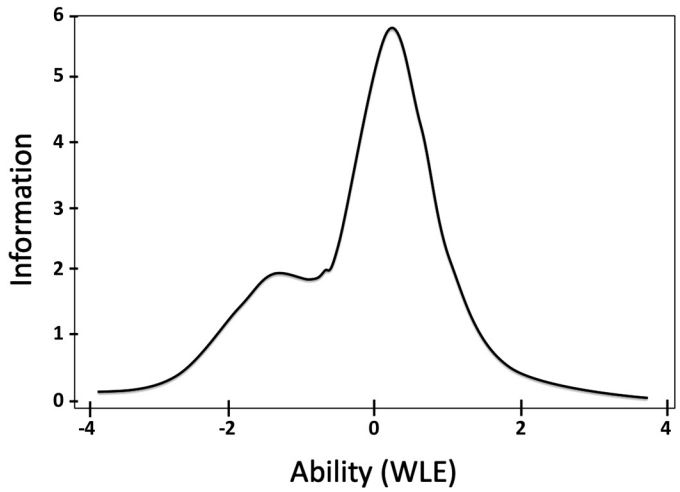


Fig. 2. Test item function determined that our assessment discriminated scores of those student with average ability [−2 to 1.5 weighted likelihood estimate (WLE)], but not those with the highest ability (>2 WLE).

ion movement across the membrane; Q16; modulation of signal intensity) had a very high probability of correct answer and did not capture much variation in student ability.

A Wright map provides a more detailed visualization of the assessment as it displays both the ability score where students have a 50% likelihood of answering correctly (item difficulty), as well as a histogram of students at that ability score (Fig. 4). The Wright map showed that paired questions (e.g., Q5/6, Q7/8), where a "what" question was paired with a "how" follow-up item, functioned as expected. The "how" questions (Q6 and Q8) discriminated students of higher ability better than the "what" question in the pair (Q5 and Q7). Item discrimination centered around students of average ability, where the bulk of respondents occurred (−2 to 1.5). There is a noticeable gap in item discrimination of very advanced ability (>2), as there are fewer examples of such students in the course. Additionally, Q1, which asked students to reason about concentration gradients, had a near 100% correct answer rate with students of all ability scores. Our questions were well matched to the range of most of the student abilities in the course, and Q1 remains important because, when paired with Q2, it uncovers an additional underlying weakness in student reasoning.

*Internal structure validity: DIF.* We used DIF analysis to determine whether individual items did not perform equitably for students with respect to their binary gender, ethnicity, and economic status. All items performed equally for all groups of interest (AIC = 7654.17).

*Capturing differences in ability among populations.* To support the use of EGAD for quantifying differences in performance among student populations, we investigated if the EGAD could differentiate mean student abilities (WLE) between the two time points. We used linear mixed-effect regression, and, when controlling for student GPA and demographic factors, we were able to detect differences in mean student WLE (model $r^2 = 0.47$, $P < 0.001$, $n = 444$) (Fig. 5). While these differences could be due to a number of factors, including student learning, student affect, or student experience (e.g., the students were tired or had multiple exams in other courses the day of the assessment), these results, nevertheless, support the use of the EGAD to evaluate differences in performance among student populations.

Student ability estimate was strongly correlated with cumulative GPA ($r^2 = 0.47$, $P < 0.001$, $n = 444$, Table 3). We also found that men and non-URM students had higher ability estimate on both assessment time points ($P = 0.009$ and $P < 0.001$, respectively). This evidence supports the use of the EGAD to determine differences in student responses across demographic groups in introductory physiology students. The number of lectures a student had the opportunity to attend between the assessment time points was also positively correlated with the students' ability estimate ($P = 0.02$). While we did not measure student learning through outside analyses, this is evidence of student learning and helps address a test-retest phenomena, where students taking the same version of the test improve performance simply by familiarity with the questions. If proximity to the test material had an effect on performance, we would expect to see an inverse relationship between days of instruction and EGAD performance.

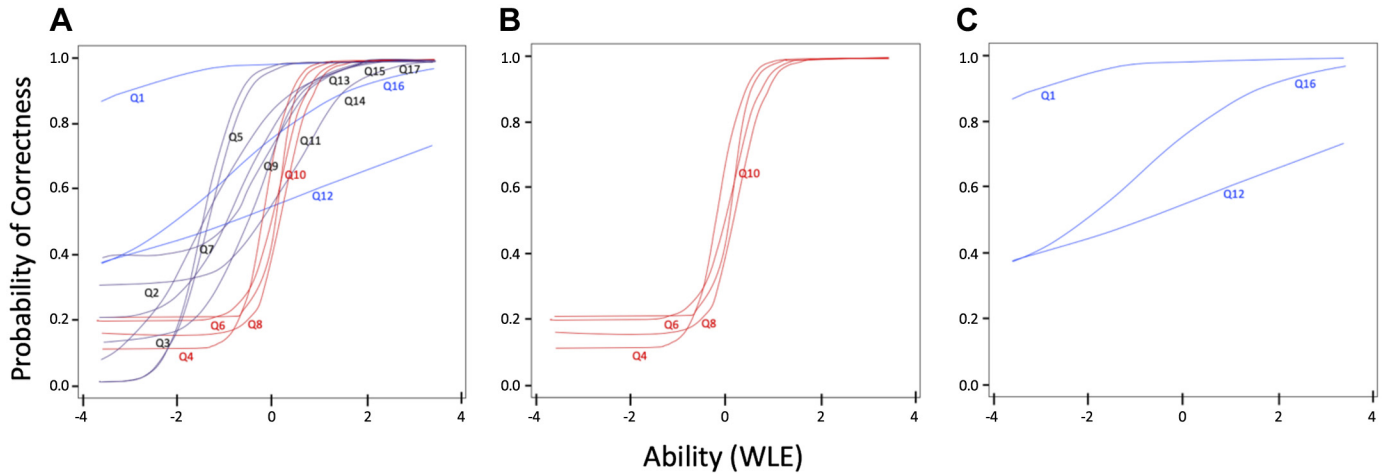


Fig. 3. Item characteristic curve assessed an individual item's ability to discriminate student answers of varying student ability. *A*: 10 items (in purple) captured students from a middle range of weighted likelihood estimate (WLE) values (−2 to 2). *B*: 4 items [*question 4* (Q4), Q6, Q8, and Q10; red) have high probability of a correct answer for students at a tightly defined WLE (−0.5 to 0.5). *C*: 3 items (Q1, Q12, and Q16; blue) had a very high probability of a correct answer and did not capture the variation in student ability.

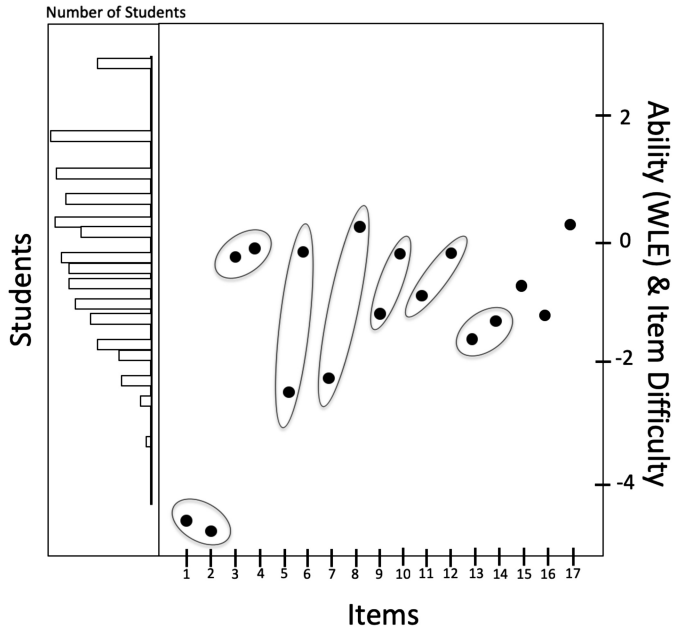


Fig. 4. The Wright map shows paired questions (circled) functioned as expected, with "how" items (*left* member of the pair) discriminating students of higher ability than their "what" counterpart (*right* member). The left *y*-axis is a histogram of the number of students at each ability score on the right *y*-axis, which shows both student ability score for the histogram and item difficulty for the dot-map, both in logits on the same scale. Item discrimination centered around students of average ability, where the bulk of respondents occurred (−2 to 1). There is a noticeable gap in item discrimination for very advanced ability (>1). WLE, weighted likelihood estimate.

## DISCUSSION

The goal of this project was to develop and provide validity evidence based on the AERA/APA/NCME Standards of a multiple-choice assessment instrument that could rigorously measure student understanding of introductory neurophysiology. The EGAD can be completed online outside of class time in ~20 min by students, thereby not encroaching on face-to-face class time. In this study, the EGAD functioned equitably for a diverse student population and was able to capture changes in student learning.

We believe that the EGAD is able to capture the concepts that are difficult for students to master in learning neurobiology. Our assessment captured not only changes in learning "what" happens in various neurophysiological situations, but, due to its two-tiered design, EGAD also captured important changes in student reasoning, the "how." This is a trend that other investigators have noted and have seen as a major value of the two-tier format (11, 14). We noticed that, based on IRT analysis, the "what" questions were less difficult than the second tier "how" questions, where reasoning was assessed. While being able to correctly predict what happens will improve exam scores, to be successful in upper level courses student will have to master answering the "how" questions. Therefore, as instructors, it is imperative that we are able to assess how a student reasons about core course concepts. Using the EGAD, instructors can monitor how student reasoning is changing as a result of instruction.

We have evidence of the validity of the EGAD across a wide range of students enrolled in introductory physiology. Only one question (Q1) did not effectively discriminate students in our population, as most all students answered correctly. This question addressed concentration gradients, a concept most all students have mastered before entering introductory physiology. We did not remove this item, as we believe it is important for establishing a baseline for the population, as well as alerting the instructor to the fact that some students may not be meeting this baseline. Students who struggle with this item may possess fundamental misunderstandings or are missing information and could benefit from targeted instructional support or instructor intervention.

The Wright map generated from the IRT model shows that the assessment is most reliable for students performing near the middle ability score within our study population, which is likely typical of large state university introductory physiology students. For this reason, we believe the EGAD is best suited to introductory physiology students. The EGAD will allow instructors to confirm that students are not only improving in factual knowledge (i.e., able to answer "what" questions), but also integrating the reasoning for physiological mechanisms. Since students' mechanistic reasoning may transfer to increased performance on questions that require students to analyze, synthesize, or evaluate physiological situations, instructors can utilize the results of the EGAD to understand and improve the mechanistic reasoning of their students.

*Equity and diversity in validation.* It is imperative that the EGAD be free of bias regarding gender, ethnicity, and first-generation and economic status of the student. While we observed differential performance on the assessment as a whole, we found that no individual item presented an unfair bias to any particular group of interest. This finding is a critical result of our validation analysis, because it means no individual item is disproportionally more difficult for any particular demographic group, and, therefore, the EGAD results are an

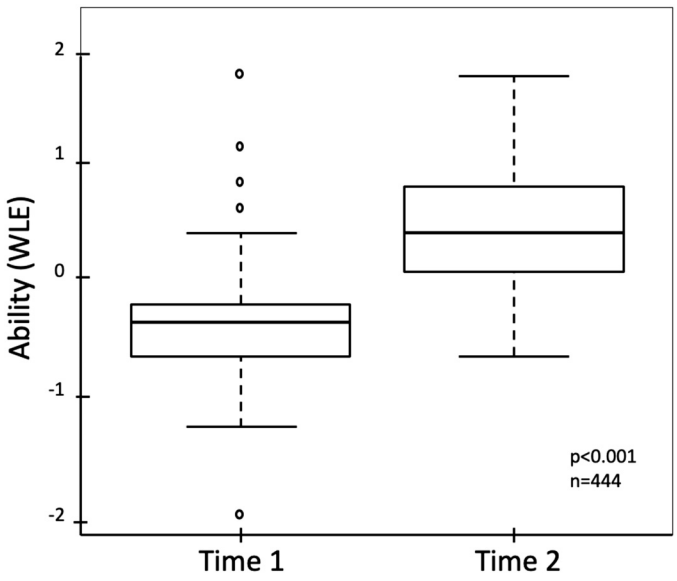


Fig. 5. Differences in mean student ability [weighted likelihood estimate (WLE)] between two time points, demonstrating the Electrochemical Gradients Assessment Device (EGAD) can capture ability differences between populations (model $r^2 = 0.47$, $P < 0.001$, $n = 444$). We report an effect size of 1.522 for WLE, a difference of ~4–5 questions correct. The linear mixed-effect regression box plot defines the center line as the mean, the box as the interquartile range, the whiskers as the range of the data, and the circles are defined as outliers by the model as 1.5× the interquartile range.

Table 3. *Estimates, SE, and P value for linear, mixed-effect regression*

| Variable | Estimate | SE | *P* Value |
|---|---|---|---|
| Time point (T1/T2) | 1.522 | 0.069 | **<0.001** |
| Gender, male | 0.258 | 0.098 | **0.009** |
| GPA | 0.611 | 0.127 | **<0.001** |
| EOP/non-EOP | −0.026 | 0.15 | 0.86 |
| URM-Caucasian | 0.388 | 0.117 | **0.001** |
| URM-International | −0.209 | 0.241 | 0.39 |
| URM-URM | 0.092 | 0.205 | 0.65 |
| URM-other | 0.039 | 0.139 | 0.78 |
| Lecture number | 0.113 | 0.049 | **0.02** |

EOP, Educational Opportunity Program; GPA, grade point average; T1 and T2, time point at the beginning and end of the week's instruction, respectively; URM, underrepresented minority in science. *P* values in bold are below $P = 0.05$.

unbiased assessment. However, on the assessment as a whole, we found that men performed higher than women and Caucasian and Asian-American students performed higher than those of underrepresented minorities. These results mirror broader trends in academic performance that have been recorded in other studies of introductory biology exams (10, 47).

It is likely that social factors, such as test anxiety, imposter syndrome, or possible systematic discrepancies in preparation for college may explain our finding that women scored lower than men on the assessment (28). Anxiety, in particular, has been shown to negatively impact performance in women (25, 26), and, while this assessment was not graded for correctness, only for participation, latent test anxiety could exist in any assessment. Furthermore, it has been demonstrated that students of underrepresented minorities and women can suffer from stereotype threat (where students feel at risk of conforming to stereotypes about their social group) or imposter syndrome (feeling as if one has "gamed" the system to belong in the course), and these can also negatively affect exam performance (41, 43, 44). Other research shows that test format can impact student performance, with women tending to perform lower on multiple-choice questions than on constructed response open-ended questions in Science, Technology, Engineering, and Mathematics (45). As the EGAD is a multiple-choice assessment, this may have contributed to the gender differences in performance.

Collectively, it is possible that these factors, in part, explain lower performance on the EGAD of some groups of students, rather than implicit bias of the assessment. It is also possible that students of these groups do in fact have a higher ability on this assessment as a whole. We recommend instructors be cognizant of these discrepancies in performance when developing course instruction and using the EGAD to assess student learning.

*Limitations.* While we demonstrate that the EGAD functions effectively for the majority of students in introductory neurophysiology, there are limitations of the assessment. Some items in the assessment (e.g., Q1, particles diffuse from high to low concentration) were particularly easy and did not discriminate student ability in our sample. These questions might still be useful in other courses with more relaxed prerequisites or where there are more students with weaker scientific backgrounds. We also note that even the most difficult items did not discriminate the highest performing students in the population. We also expect that courses that emphasize neuro- and electrophysiology (e.g., neuroscience courses) to have more students in the very advanced ability range (in comparison to our students) and might require more difficult questions. Consequently, we will continue to create more challenging neurophysiology questions to be added to future versions of EGAD. For traditional introductory physiology courses that spend four to six lectures on electrophysiology, the EGAD is well aligned, and for courses where students spend less time on electrophysiology, our easier questions (Q1, Q5, Q7, Q12) might be more discriminatory of those student populations.

The EGAD is designed to be a short, formative assessment that instructors can use to determine how well their students understand basic electrophysiology. It is important to note that, in an effort to minimize the number of questions, and therefore the amount of time for students to complete, items are not evenly distributed among learning objectives. We retained the more challenging versions of questions that addressed similar learning objectives. Therefore, instructors should focus more of their attention on which questions students are struggling with, in addition to the changes in performance on the whole assessment. It is also important to note that, while the EGAD can capture differences in populations of students, these differences could be due to a number of factors, including student learning, student affect, or student experience (e.g., other course demands).

*Implications for teaching and learning.* The EGAD focuses on the fundamental neurophysiology concepts associated with membrane and action potentials. EGAD is a very practical and logistically simple tool to use, as it is a multiple-choice assessment taken online outside of course time and completed by students in ~20 min.

As 650,000 students take introductory biology and 450,000 take anatomy and physiology courses (28) in preparation for careers in the health field, helping this large number of students gain a deeper understanding of the principles of neurophysiology could have far reaching consequences. The EGAD can provide important feedback to instructors on their students' reasoning, identifying difficult content areas and efficacy of teaching practices. We suggest faculty first use the EGAD to inform how they organizing their curricula around neurophysiology learning objectives (48). Our results show that concepts such as diffusion and material movement down concentration gradients are less challenging for students and only need limited instructional attention. We also suggest that faculty focus on mechanistic reasoning that helps students explain how and why processes occur. This provides students with much-needed practice to master understanding of the complexities associated with electrochemical gradients. In the future, the EGAD could also be modified to include more challenging questions to capture students of higher ability score ($>1$) in introductory courses.

Instructors can use the EGAD as a precourse assessment to determine where their students have difficulty in neurophysiology. As we have aligned each question with our stated learning objectives, it would be important for instructors to look at which specific questions their students struggled with rather than just using the total score. Precourse results can assist instructors in focusing instruction and designing specific practice opportunities for the students. Results from the EGAD can be used to address particular sticking points in student

reasoning. During in-class activities, EGAD questions or isomorphic questions (i.e., questions that test the same concepts but with different superficial features) can be used as formative assessment. Using EGAD at the end of the neurophysiology unit will help the faculty member assess the effectiveness of their teaching activities and inform the redesign of the next iteration of the neurophysiology unit. We do, however, caution against using the EGAD for summative assessment, as higher performing students will reach the maximum score on the assessment as most questions target the middle of the population. The Vision and Change report emphasized the need for data-driven support of in-class instruction (1, 16). We propose that faculty use the EGAD instrument as a means to monitor both student understanding of concepts covered in the assessment, and the effectiveness of innovative teaching methods created to help students master the challenges of electrophysiology. An easily administered assessment like the EGAD can provide biology faculty with data they can use to develop instruction that matches the level of conceptual understanding of their current students. Instructors who are aware of their students' misconceptions can prepare interactive pedagogies (7, 12, 13) to improve student learning of complex biology concepts.

*Conclusions.* The Electrochemical Gradients Assessment Device (EGAD) is a 17-item multiple-choice assessment that can be completed in ~20 min. The EGAD has good internal structure and is reliable for undergraduates diverse in gender, race/ethnicity, and economic status. The questions in the EGAD are of intermediate difficulty and can discriminate among the majority of students. The EGAD is best at differentiating populations of high- middle- and low-performing students. EGAD can capture performance differences among student populations, both within a single administration of the assessment and between time points. This assessment is ideal for introductory physiology courses with a neurophysiology component, but is limited in scope for more advanced courses, as items did not discriminate the highest performing students.

## ACKNOWLEDGMENTS

We thank Dr. Jenny McFarland (Edmonds Community College), who administered earlier versions of the EGAD, as well as the students who completed the EGAD assessment at the University of Washington. We also thank Dr. Denise Pope (Center for the Integration of Research, Teaching and Learning) and four neurophysiology instructors who provided expert feedback on the EGAD. We thank Patricia Martinkova for help with the statistics. Emily Scott provided critical feedback on the manuscript. Other SimBiotic Software employees aided in setting up the Action Potentials Extended laboratory for the study.

## GRANTS

This work was supported by a grant from the National Science Foundation (no. 1661263) and Department of Biology laboratory startup funds for J. H. Doherty at the University of Washington.

## DISCLOSURES

K. J. Kim and E. Meir are employees of SimBiotic Software, the publisher of the Action Potentials Extended laboratory discussed in this paper and receive remuneration based on sales of the laboratory. No other conflicts of interest, financial or otherwise, are declared by the authors.

## AUTHOR CONTRIBUTIONS

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

## REFERENCES

1. **American Association for the Advancement of Science.** *Vision and Change in Undergraduate Biology Education: A Call to Action*. Washington, DC: AAAS, 2011.
2. **American Educational Research Assoc., American Psychological Association, National Council on Measurement in Education.** *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, 2014.
3. **Aho K, Derryberry D, Peterson T.** Model selection for ecologists: the worldviews of AIC and BIC. *Ecology* 95: 631–636, 2014. doi:10.1890/13-1452.1.
4. **de Ayala RJ.** *The Theory and Practice of Item Response Theory*. New York: Guilford, 2008.
5. **Bass KM, Drits-Esser D, Stark LA.** A primer for developing measures of science content knowledge for small-scale research and instructional use. *CBE Life Sci Educ* 15: rm2, 2016. doi:10.1187/cbe.15-07-0142.
6. **Boone WJ.** Rasch analysis for instrument development: why, when, and how? *CBE Life Sci Educ* 15: rm4, 2016. doi:10.1187/cbe.16-04-0148.
7. **Brewer EP.** Successful techniques for using human patient simulation in nursing education. *J Nurs Scholarsh* 43: 311–317, 2011. doi:10.1111/j.1547-5069.2011.01405.x.
8. **Cardozo DL.** A model for understanding membrane potential using springs. *Adv Physiol Educ* 29: 204–207, 2005. doi:10.1152/advan.00067.2004.
9. **Crowther GJ.** Which way do the ions go? A graph-drawing exercise for understanding electrochemical gradients. *Adv Physiol Educ* 41: 556–559, 2017. doi:10.1152/advan.00111.2017.
10. **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.
11. **Fisher KM, Williams KS, Lineback JE.** Osmosis and diffusion conceptual assessment. *CBE Life Sci Educ* 10: 418–429, 2011. doi:10.1187/cbe.11-04-0038.
12. **Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, Wenderoth MP.** Active learning increases student performance in science, engineering, and mathematics. *Proc Natl Acad Sci USA* 111: 8410–8415, 2014. doi:10.1073/pnas.1319030111.
13. **Freeman S, Haak D, Wenderoth MP.** Increased course structure improves performance in introductory biology. *CBE Life Sci Educ* 10: 175–186, 2011. doi:10.1187/cbe.10-08-0105.
14. **Fulmer GW, Chu H-E, Treagust DF, Neumann K.** Is it harder to know or to reason? Analyzing two-tier science assessment items using the Rasch measurement model. *Asia-Pac Sci Educ* 1: 1, 2015. doi:10.1186/s41029-015-0005-x.
15. **Gómez-Benito J, Sireci S, Padilla J-L, Hidalgo MD, Benítez I.** Differential Item Functioning: Beyond validity evidence based on internal structure. *Psicothema* 30: 104–109, 2018. doi:10.7334/psicothema2017.183.
16. **Goodman BE, Barker MK, Cooke JE.** Best practices in active and student-centered learning in physiology classes. *Adv Physiol Educ* 42: 417–423, 2018. doi:10.1152/advan.00064.2018.
17. **Guy R.** Overcoming misconceptions in neurophysiology learning: an approach using color-coded animations. *Adv Physiol Educ* 36: 226–228, 2012. doi:10.1152/advan.00047.2012.
18. **Haudek KC, Prevost LB, Moscarella RA, Merrill J, Urban-Lurain M.** What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE Life Sci Educ* 11: 283–293, 2012. doi:10.1187/cbe.11-08-0084.
19. **Hestenes D, Wells M, Swackhamer G.** Force concept inventory. *Phys Teach* 30: 141, 1992. doi:10.1119/1.2343497.
20. **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.
21. **Huggins-Manley AC.** Psychometric consequences of subpopulation item parameter drift. *Educ Psychol Meas* 77: 143–164, 2017. doi:10.1177/0013164416643369.
22. **Jin H, Anderson CW.** Developing assessments for a learning progression on carbon-transforming processes in socio-ecological systems. In: *Learning Progressions in Science*. Rotterdam, the Netherlands: Sense, 2012, p. 151–181. doi:10.1007/978-94-6091-824-7_8.

23. **Kalinowski ST, Leonard MJ, Taper ML.** Development and validation of the conceptual assessment of natural selection (CANS). *CBE Life Sci Educ* 15: ar64, 2016. doi:10.1187/cbe.15-06-0134.
24. **Kim K, Meir E.** *Action Potentials Extended*. Missoula, MT: SimBio, 2015.
25. **Maloney EA, Beilock SL.** Math anxiety: who has it, why it develops, and how to guard against it. *Trends Cogn Sci* 16: 404–406, 2012. doi:10.1016/j.tics.2012.06.008.
26. **Maloney EA, Waechter S, Risko EF, Fugelsang JA.** Reducing the sex difference in math anxiety: the role of spatial processing ability. *Learn Individ Differ* 22: 380–384, 2012. doi:10.1016/j.lindif.2012.01.001.
27. **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.
28. **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.
29. **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.
30. **Michael J, Modell H, McFarland J, Cliff W.** The "core principles" of physiology: what should students understand? *Adv Physiol Educ* 33: 10–16, 2009. doi:10.1152/advan.90139.2008.
31. **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. *Am J Physiol* 277: S127–S135, 1999. doi:10.1152/advances.1999.277.6.S127.
32. **Modell H, Michael J, Wenderoth MP.** Helping the learner to learn: the role of uncovering misconceptions. *Am Biol Teach* 67: 20–26, 2005. doi:10.1662/0002-7685(2005)067[0020:HTLTLT]2.0.CO;2.
33. **Modell HI.** How to help students understand physiology? Emphasize general models. *Adv Physiol Educ* 23: 101–107, 2000. doi:10.1152/advances.2000.23.1.S101.
34. **National Research Council**. *Science Teaching Reconsidered: A Handbook*. Washington, DC: National Academies, 1997. doi:10.17226/5287.
35. **National Research Council.** *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. Washington, DC: National Academies, 2000. doi:10.17226/9853.
36. **Neumann I, Neumann K, Nehm R.** Evaluating instrument quality in science education: rasch-based analyses of a nature of science test. *Int J Sci Educ* 33: 1373–1405, 2011. doi:10.1080/09500693.2010.511297.
37. **Nunnally J, Bernstein I.** *Psychometric Theory*. New York: McGraw-Hill, 1994.
38. **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.
39. **Price RM, Pope DS, Abraham JK, Maruca, Meir E.** Observing populations and testing predictions about genetic drift in a computer simulation improves college students' conceptual understanding. *Evol Educ Outreach* 9: 8, 2016. doi:10.1186/s12052-016-0059-6.
40. **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.
41. **Shapiro JR, Neuberg SL.** From stereotype threat to stereotype threats: implications of a multi-threat framework for causes, moderators, mediators, consequences, and interventions. *Pers Soc Psychol Rev* 11: 107–130, 2007. doi:10.1177/1088868306294790.
42. **Silverthorn DU.** Uncovering misconceptions about the resting membrane potential. *Adv Physiol Educ* 26: 69–71, 2002. doi:10.1152/advan.00012.2002.
43. **Spencer SJ, Steele CM, Quinn DM.** Stereotype threat and women's math performance. *J Exp Soc Psychol* 35: 4–28, 1999. doi:10.1006/jesp.1998.1373.
44. **Steele C, Spencer S, Aronson J.** Contending with group image: the psychology of stereotype and social identity threat. *Adv Exp Soc Psychol* 34: 379–440, 2002. doi:10.1016/S0065-2601(02)80009-0.
45. **Wilson K, Low D, Verdon M, Verdon A.** Differences in gender performance on competitive physics selection tests. *Phys Rev Phys Educ Res* 12: 020111, 2016. doi:10.1103/PhysRevPhysEducRes.12.020111.
46. **Wilson M.** Making measurement important for education: the crucial role of classroom assessment. *Educ Meas* 37: 5–20, 2018. doi:10.1111/emip.12188.
47. **Wright CD, Eddy SL, Wenderoth MP, Abshire E, Blankenbiller M, Brownell SE.** Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE Life Sci Educ* 15: ar23, 2016. doi:10.1187/cbe.15-12-0246.
48. **Xu X, Lewis JE, Loertscher J, Minderhout V, Tienson HL.** Small changes: using assessment to direct instructional practices in large-enrollment biochemistry courses. *CBE Life Sci Educ* 16: ar7, 2017. doi:10.1187/cbe.16-06-0191.