MODESTUM EURASIA Journal of Mathematics, Science and Technology Education, 2021, 17(12), em2048
ISSN:1305-8223 (online)
OPEN ACCESS Research Paper https://doi.org/10.29333/ejmste/11376

Construction and Evaluation of an Instrument to Measure High School Students Biological Content Knowledge

Kathy L. Malone [1*], William J. Boone [2], Andria Stammen [3], Anita Schuchardt [4], Lin Ding [3], Zakee Sabree [5]

[1] Nazarbayev University, Nur-Sultan, KAZAKHSTAN
[2] Miami University College of Education, Cincinnati, Ohio, USA
[3] Ohio State University, College of Education and Human Ecology, Department of Teaching and Learning Columbus, Ohio, USA
[4] University of Minnesota - Twin Cities, College of Biological Sciences, Department of Biology Teaching and Learning, Minneapolis, Minnesota, USA
[5] Ohio State University, College of Arts and Sciences, Department of Evolution, Ecology, and Organismal Biology, Columbus, OH, USA

Received 25 August 2021 ▪ Accepted 6 November 2021

Abstract

Instruments for assessing secondary students' conceptual understanding of core concepts in biology are needed by educational practitioners and researchers alike. Most instruments available for secondary biology (years 9 to 12) focus only on highly specific biological concepts instead of multiple core concepts. This study describes the development of a 25-item instrument designed to fill this gap, the High School Biology Concept Inventory (HS-BCI). The HS-BCI not only assesses student knowledge of key biological concepts but also alternative conceptions. Using Rasch theory, the initial instrument was constructed from a pool of 61 instrument items using test results from 1015 students. The final 25-item instrument was validated with 1955 students. The results provide reliability and validity evidence for the HS-BCI. The findings suggest that it can be utilized to assess both conceptual knowledge and alternative conceptions.

Keywords: alternative conceptions, biology, concept inventories, instrument construction, Rasch

INTRODUCTION

There is a long history of utilizing different instrumentation to measure student traits or concepts in science education such as the Force Concept Inventory (Hestenes et al., 1992), the Science Interest Survey (Lamb et al., 2012) and the Scientific Reasoning Progress Tool (Abate et al., 2020). In this paper, the development of the High School Biology Concept Inventory (HS-BCI) for students aged 14 to 19 years is presented. The HS-BCI was developed to provide "measures" of high school students' biology understanding beyond those which are provided by existing instrumentation. The HS-BCI incorporates key concepts of high school and middle school (ages 11- 13)) biology courses as well as the developmental progression of these concepts as detailed in multiple standards (e.g., KMK, 2005; NGSS Lead States, 2013). In addition, the HS-BCI is designed to connect to the core concepts in entry-level college or university biology courses (Brewer & Smith, 2011) to determine student growth over multiple years. The assessment item choices were generated from alternative conceptions research to allow teachers and researchers to assess the conceptions held by their students and/or participants. Thus, this paper is of use to both teachers and educational researchers, and the instrument development procedures detailed here can be used by researchers in all domains.

For quasi-experimental studies testing the effectiveness of high school curriculum interventions, it is important to have measurement instruments that provide reliable and valid measures. A critical component of such robust instruments is the need to take into consideration student alternative conceptions when the items are designed in order to be able to determine students' conceptual stances and how they change over the course of instruction (Klymkowsky & Garvin-Doxas, 2020). However, while there have been several such assessments developed for college-level biology, few target high school-level biology students and their general preconceptions.

✉ klmalone60@gmail.com **(*Correspondence)** ✉ boonewjd@gmail.com ✉ andriastammen@gmail.com
✉ aschucha@umn.edu ✉ ding.65@osu.edu ✉ sabree.8@osu.edu

Contribution to the literature

- There is a need for a secondary level biology concept inventory which covers the major core concepts included in introductory biology courses in order to allow for pre and post assessments for an entire year long course.
- There is a need for valid and reliable concept inventories that allow researchers and practitioners to determine the alternative biological conceptions held by secondary students over a single year as well as across multiple educational levels.
- There is a continued need to demonstrate how one can develop concept inventories using Rasch analysis.

CONCEPTUAL BIOLOGY INSTRUMENTS – AN OVERVIEW

College/University-level Biology Instruments

While several college-level biology assessments exist, most focus on specific concepts such as natural selection, or osmosis. Of course, this is not surprising since the topics of college-level biology instrumentation may align well with the curriculum of many college-level biology classes as many target a limited number of concepts. Below is an overview of instruments developed in the last 20 years.

Many college-level instruments target the specific concepts of evolution and focus exclusively on alternative conceptions, contextual reasoning, evolutionary processes, or genetic drift (e.g., Anderson, et al., 2002; Kalinowski et al., 2016; Nadelson & Southerland, 2009; Price et al., 2014). Others are available for targeted concepts such as genetics (e.g., Todd & Romaine, 2016), cell biology (Shi et al., 2010), meiosis (Kalas et al., 2013), diffusion, and osmosis (Fisher et al., 2011) or statistical reasoning in biology (Deane et al., 2016). Todd and Romaine (2016) was unique in that it focused on learning progressions, was adapted based on a previously developed high school instrument (Todd et al., 2017), and targeted the genetic concepts focused on in Vision and Change (Brewer & Smith, 2011).

Only one college-level instrument, the 30-item multiple-choice (MC) Biology Concept Inventory attempts to assess the course content presented in introductory college biology (Klymkowsky et al., 2010). However, its primary focus is on randomness, process, and structure. Other recent instruments have been developed to focus on only specific core concepts in college general biology but not the key concepts for the entire course (Cary et al., 2019; Couch et al., 2019).

In summary, instruments for college-level populations, mostly focus on targeted concepts. Typically, these instruments were developed through the use of college student interviews, expert panel reviews, and some level of psychometric analysis. Thus, they have limited use in assessing changes in high school students' conceptual understanding.

High School Level Biology Instruments

Fewer instruments exist for high school biology and those that are available only target limited topics. Many are two-tiered instruments. Two-tier questions first examine student content knowledge in a first-tier and then assess the reasoning used in a second-tier (Treagust, 1986). Examples of these types of instruments include a 13 two-tiered item assessment for photosynthesis & respiration (Haslam & Treagust, 1987); a 13 two-tiered item assessment on flowering plants (Lin, 2004); a 12 two-tiered item assessment on breathing and gas exchange (Treagust & Mann, 1998) and a 25 two-tiered item assessment called the Internal Transport and the Human Circulatory System (Wang, 2004). Wang's (2004) assessment is unique in that it was developed for three specific grade bands (i.e., primary, middle, and high school level). This could allow for cross-grade band studies on these topics.

Other assessments for this level targeted specific genetic concepts. For example, Todd, et al., (2017) developed the 34 MC items Learning Progression-Based Assessment of Modern Genetics while Tsui and Treagust (2010) constructed a two-tiered 13-item assessment focused on both Mendelian and molecular concepts. The Learning Progression-Based Assessment of Modern Genetics (Todd et al., 2017) and its college-level adaptation (Todd & Romine, 2016) are the only instruments that could allow for studies linking high school and undergraduate biology conceptual development.

Limitations of Current Instruments

After reviewing these biology instruments, we determined that most appeared to have either a narrow conceptual focus or only targeted college-age students (e.g., Anderson et al., 2002; Todd et al., 2017). The only conceptual instrument designed to cover a broad conceptual range was constructed for college-age students (Klymkowsky et al., 2010).

These limitations can be problematic for researchers and instructors who wish to focus on year-long high school biology implementations. High school teachers have limited time and may not be able to administer an extended formative assessment instrument for every topic covered. Thus, a single instrument that covers the core concepts in an introductory course could allow for

teachers to be able to determine the alternative conceptions held by their students across multiple topics using a single instrument thus saving precious time for needed mean-making activities. In addition, due to the lack of appropriate high school instrumentation, biology education researchers attempting to evaluate different instructional pedagogies in biology at this level either use self-developed instruments (e.g., Adeoye & Abimbola, 2016; Pugh et al., 2014) or focus on qualitative assessments (e.g., Cavalho et al., 2018; Hogan, 2000). It is sometimes unclear if these self-developed instruments undergo extensive validity or reliability evaluation which can lead to their conclusions possibly being questioned. Another issue with self-developed instruments is that they are often only useful for a single evaluation (i.e., the assessment is specific to a single class at a single institution). This lack of common metrics hinders researchers, grantors, and practitioners as it is not possible to compare student abilities across multiple studies and contexts. The need for common instruments was recognized by the National Science Foundation (NSF) in the USA when they funded a compendium that listed science research instruments available for K-12 grade levels (Minner et al., 2012). While this compendium lists some biology concept tests that are knowledge-based, there are no concept assessments listed that include a focus on students' preconceptions. In addition, there are no instruments that are based upon the core concepts and developmental progressions suggested by national-level standards documents such as the Next Generation Science Standards in the USA (NGSS Lead States, 2013).

In conclusion, while there are a range of instruments to assess concept development for the field of biology education the majority of them have been developed specifically for college-age students and lack a broad conceptual focus in biology. This paper details the development and measurement properties of a biology instrument designed to help alleviate these limitations as it was designed not only for high school level students, but also to assess the major alternative conceptions held by students across the core biological concepts and to be available for researchers as a common metric.

RESEARCH GOALS

1) Design a High School – Biology Concept Inventory (HS-BCI) to provide measures of high school students' conceptual knowledge across yearlong courses that a) is easy to administer, b) focuses on broad core concepts suggested by national standards, and c) uses alternative conception research to guide item construction.
2) Evaluate the measurement properties of the HS-BCI.
3) Utilize the HS-BCI to study a sample of students enrolled in different levels of high school biology.

INSTRUMENT DEVELOPMENT

The following sections detail the development process. The sections include both the development of the item pool as well as the final item selection. An overview of the instrument development and psychometric analysis for the HS-BCI is provided in the flow chart in Figure 1.

Instrument Content and Item Structure

The content and item structure of the HS-BCI was informed by high school biology teachers. Teachers were asked what type of assessment structure and content they would find most useful. A panel of high school biology teachers, science education researchers, and biology experts (N=8) engaged in a series of meetings that focused on the most appropriate content and structure. This panel requested an instrument that would 1) be short, (i.e., administration of 30 to 40 minutes); 2) be easy to score, and 3) provide an overview of students' conceptual knowledge as defined by national standards as well as their alternative conceptions.

Item Content, Development, and Revision

Firstly, it is critical to determine the trait that is to be measured in an instrument. It is only if the items comprise one overall trait, that respondent measures can be confidently computed and have meaning (Wright & Stone, 1979). Indeed there are many numerous biology concepts covered in any one year biology course. However, the core concepts in biology can be considered to be interconnected. For that reason, 5 core high school biology concepts suggested not only by expert panelists but also described in national standards (e.g., NGSS Lead States, 2013) were chosen to guide the test item content. These 5 core biological concepts are routinely the focus of introductory biology classes. Each concept was associated with an essential question to guide item development (see Table 1). Our view was that the core concepts together define a single variable.

A total of 61 MC questions were developed, reviewed, and revised. Each question focused on concepts that covered a range of understanding progressing from concepts covered in middle school to that in undergraduate (UG) college biology classes. The item development phase included: 1) a review of student alternative biology conception research, 2) convening of two expert panels (Liu, 2010; Schmeiser & Welch, 2006), 3) conducting interviews with college and high school biology students, and 4) collecting data using a high school student questionnaire to evaluate potential instrument items.

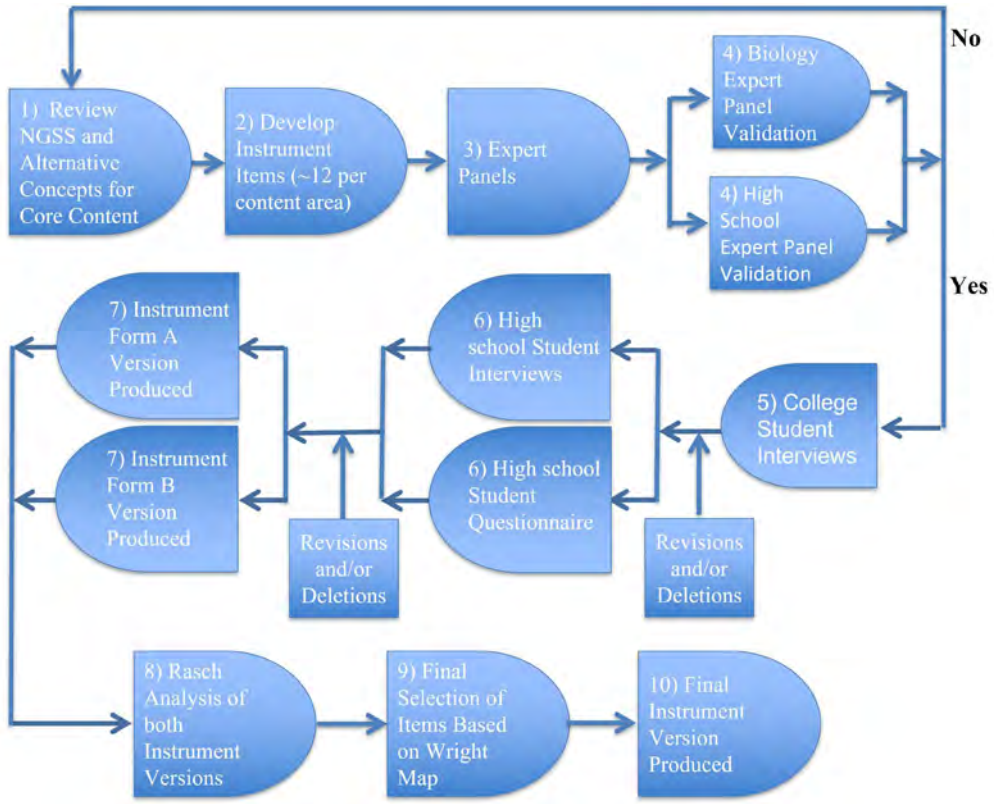

Figure 1. Instrument development flowchart (starting with the review of alternative concepts (#1) and finishing with the final instrument (#10))

Table 1. Assessment essential questions associated with the core concepts

Core Concepts (CC) in HS-BCI	Essential Questions
Core Concept 1 (CC1). Evolution and diversity	How and why do populations change over time?
Core Concept 2 (CC2). Population interactions	How and why do populations in a system interact with other populations over time?
Core Concept 3 (CC3). Growth and reproduction	How is information preserved during reproduction while still producing the variation observed in life?
Core Concept 4 (CC4). Inheritance	How are traits passed from parents to offspring?
Core Concept 5 (CC5). Energy and matter	How and why do energy and matter transfer within and across systems?

In all living organisms the functions necessary to sustain life
a. Depend on creating energy molecules because food is a source of chemical energy
b. Depend on food as a source of molecules because they provide building materials and chemical energy
c. Depend on the intake of a healthy diet because nutrients are sources of building materials
d. Depend on food absorption because water, air and mineral are needed for the environment.

Answer Choice	Alternative Conception	Reference
a.	Food is a source of energy not a source of building materials	AAAS Project 2061 (n.d.)
b.	Correct	
c.	Food is a source of building materials but not a source of chemical energy	AAAS Project 2061 (n.d.)
d.	Food is any material taken from the environment	Anderson et al., 1990

Figure 2. HS-BCI item with its associated table connecting identified alternative conceptions and answer choices

To demonstrate item development, exemplars are provided in this section. Twelve potential instrument items were developed for CC2. Although each of those items was unique, each addressed the essential question and covered the developmental progression expected from middle school through high school and the UG college level. Figure 2 highlights how each distractor of an item is linked to an identified alternative conception

Plants reproduce_____.
a. Only sexually using flowers and fertilized seeds.
b. Either sexually or asexually depending on the type of plant.
c. Only asexually through generating a genetically identical clone.
d. Only asexually if the plant has both male and female reproductive cells.

Figure 3. Initial HS-BCI question identified for revision by expert review panels

You find the following population of mice on an island with dense grasses to live in and few predators.

100 Big Mice
Stronger than small mice but find it hard to move quickly through dense grass

100 Small Mice
Can move quickly through dense grass

What would happen to the population of mice over 1000 generations if another animal population arrived on the island and equally preferred eating both big and small mice?

a. The frequency of big mice in the population will increase relative to the small mice because they are stronger and more likely to defend themselves.
b. The frequency of small mice in the population will increase relative to the big mice because they can move quickly and are more likely to escape being eaten.
c. The frequencies will remain the same because being stronger and quicker are characteristics that are balanced between the mice and will help them survive in their environment.
d. The frequencies will remain the same because all organisms are equally able to survive and reproduce under most conditions.

Figure 4. Instrument item used in high school student interviews

that was obtained from either biology research, expert panel suggestions, or student interviews.

The initial item pool was reviewed by 2 expert panels. One panel was composed of 5 faculty and staff in biology from three different universities. A second panel was composed of 8 high school biology teachers. The teachers taught at a mix of public and private schools throughout the United States. All panel members were asked to review the questions concerning factual accuracy, diagrammatic accuracy, whether distractors addressed the relevant alternative conceptions, readability, cultural bias, and age appropriateness.

Subsequently, the item in Figure 3 was identified for revision by both panels. This item is composed of three incorrect alternative answers. For example, if a student selects distractor C, then the student could have the alternative conception that sexual reproduction only occurs in animals but not in plants (Berthelsen, 1999).

The consensus of the panels was that the item exhibited a lack of precise phrasing. For example, a high school expert member wrote, "What about plants…like spider plants sending out shoots, but also forming seeds". Furthermore, a biology panel member wrote, "Depending on the type, some plants may be able to do both". These comments revealed that the use of 'or' in the intended correct answer (option B) might direct students towards thinking that a plant species can only employ a single reproductive strategy. A second biology expert member suggested the following modification to the item: "eliminate 'or' and put 'and' in answer choice B because you are asking about plants in general". Therefore, choice B was modified as suggested. Stammen et al. (2016) explains in detail the validation of the HS-BCI as a result of the panel review process. Fourteen of the 61 items were identified by the expert panels as needing revision. 7 items were removed from the potential item pool, and 7 items were retained but revised.

A series of think-aloud interviews were performed to determine if the remaining questions and their answer choices made sense to students. The first interviews were conducted with 7 university students from a large public university in the midwest of the USA. Each student was asked to explain their understanding of all remaining questions and answer choices, as well as figures and tables presented. Nine items were edited, and one item was removed leaving a total of 53 potential items.

Next, a sample of 8 high school students attending a suburban midwestern high school was interviewed. Students were asked to explain their understanding of questions, answer choices, and item figures and tables through the use of a think-aloud interview. Nineteen items were edited. One item was removed from the pool leaving 52 items in the item bank. Figure 4 is an item that was modified based on high school interviews.

On an island with dense grass, you find the following population of mice. Initially, these mice had few predators.

100 Big Mice
Stronger than small mice and find it hard to move quickly through dense grass

100 Small Mice
Can move quickly through dense grass

Another animal population arrived on the island. This animal could move fast in open areas, but it moved slowly through the dense grass that populated the island. What would happen to the population of mice over 1000 generations if this animal equally preferred eating both big and small mice?

a. The frequency of big mice in the population will increase relative to the small mice because they are stronger and more likely to defend themselves.
b. The frequency of small mice in the population will increase relative to the big mice because they can move quickly and are more likely to escape being eaten.
c. The frequencies will remain the same because being stronger and quicker are characteristics that are balanced between the mice and will help them survive in their environment.
d. The frequencies will remain the same because all organisms are equally able to survive and reproduce under most conditions.

Figure 5. Test item edits made to Figure 4 after high school student interviews. Edits are italicized

During high school think-aloud interviews, students commented that the new predator needed to be described and suggested birds, snakes, wolves, or a large cat (but not a house cat). These comments suggested that the item needed to include specific details concerning the new predator. The question was modified to include a description of how the newly introduced predator moves in the island's dense grass (see Figure 5).

A final step in the process consisted of a questionnaire given to 73 high school students attending rural and suburban schools in the midwest of the USA. Students were asked to explain in writing their understanding of each items' vocabulary terms and figures. Students were asked to underline any vocabulary that they did not understand, (Liu, 2010; Schmeiser & Welch, 2006). For example, all students indicated that they were not familiar with the term 'biomass'. An edit to this test item was made such that the word 'biomass' was changed to the phrase 'total mass of the organisms'. A total of 9 items were modified based on the student questionnaire results.

Final Pilot Item Pool

A total of nine items were removed from the item pool. Five of the nine items removed concerned the topic of "inheritance". One reason for the removal of several inheritance items was related to difficult scientific terminology such "trait, and recessive". Finding alternative terms or representations for these concepts proved difficult. To ensure an adequate number of inheritance items for potential inclusion in the final instrument, an additional item was developed and reviewed by three biology experts (one high school teacher and two university faculty), so that the final pool of inheritance items totaled 6. The final item bank included 53 items.

Psychometric Analysis of the Item Bank. To evaluate the measurement functioning of the instrument items, Rasch analysis was utilized (e.g., Boone et al., 2014a; Wright & Stone, 1979). Rasch analysis takes into consideration that raw scores are not linear, that instrument items have differing degrees of difficulty, that for an instrument to provide meaningful "measures" one trait must be measured with a set of instrument items, and evidence must be reviewed concerning the functioning of items to define a single trait. Rasch techniques have been used in a wide range of research fields (e.g., science education (Ding, 2014; Yang et al., 2018) math education (Chan et al., 2014), and medical research (Young et al., 2018)). However, Rasch has rarely been utilized to validate high school biology-focused instruments - only one out of six existing high school biology instruments appear to have used Rasch techniques (Todd et al., 2017). In addition, Rasch has been used to validate only a handful of college biology instruments (e.g., Deane et al., 2016; Price et al., 2014; Todd & Romaine, 2016). This underutilization of Rasch could be caused by a lack of knowledge of Rasch and its benefits to instrument development especially in its ability to take into consideration that all instrument items are not equal in terms of difficulty.

Table 2. Number of questions per core concept

Core Concept Name	Core Concept Number	Number of Questions
Evolution and Diversity	CC1	5
Population Interactions	CC2	5
Growth and Reproduction	CC3	5
Inheritance	CC4	4
Energy and Matter	CC5	6

This section provides an overview of the central analysis steps which were utilized to evaluate the item pool, the results of the analysis, and how the item pool was used to create two different instrument test forms. Using two test forms allowed for all 53 items to be field-tested without the risk of test fatigue. Student responses from both forms were collated, and data were analyzed using Rasch analysis to develop a final form of the instrument.

A Form A and Form B of the test were constructed with 16 common items. These common items served as "anchor items" which allowed all 53 items to be calibrated or distributed on the same unidimensional logit scale. The linking items amounted to about 47 percent, exceeding the 20 percent suggested by Liu (2010). Anchor items were selected based upon core concepts as well as predicted item difficulty. The two forms were balanced for number of questions per core concept.

The test forms were administered at the end of the school year to 1015 first-year biology students (i.e., students who were enrolled in their first high school biology class, typically 15 to 16 years of age) who attended 5 schools in three regions (i.e., midwest, southwest & southeast) of the United States. Teachers located at the school sites administered the tests, collected answer sheets, and mailed them back to the research team. Each student was given one class period, or about 40-45 minutes, to complete the assessment without the use of external support material.

The data were evaluated using Rasch techniques in a single linking analysis using the program Winsteps (Linacre, 2018). A central step to the analysis was first the identification of potentially problematic instrument items using Rasch Fit statistics (Boone et al., 2014). Rasch fit statistics such as MNSQ Outfit are commonly used to identify items that may not measure the same trait as the majority of items. In total, 12 of the 53 instrument items were identified as items that might not contribute to the optimal measurement. This decision was based, in part, upon the outfit Mean Square (MNSQ) Statistics. Common reasons for such instrument items not optimally functioning can be 1) high ability students missed low difficulty level questions, 2) low ability students correctly answered high difficulty level questions, and 3) items measuring a different trait than the majority of test items. Such items should be excluded from the final instrument. Additional techniques such as Principal Component Analysis of Residuals (PCAR, Gray et al., 2014) and step ordering (Lincare, 2001) were conducted to further investigate the functioning of the 53 pilot items.

Selection of Items for Final Version of HS-BCI. After the elimination of 12 items, item measures were computed for those remaining. Three members of the research team reviewed the distribution of items and selected 25 that spanned a range of item difficulty. The goal of this selection process was to identify items that marked different regions of the trait such that there were several items of high and low difficulty range. The reasoning for this range choice was that these items could be used as anchoring items for the future development of instruments at both the middle school and college levels. An additional consideration was to select a range of difficulty across items that spanned each of the core ideas. Table 2 shows the distribution of the final 25 instrument items.

PSYCHOMETRIC ANALYSIS OF FINAL VERSION OF THE INSTRUMENT

Following the selection of the final 25 items, the readability of the instrument was calculated using the Flesch-Kincaid Reading Level test. The instrument was assessed to be suitable for students at the 6th grade reading level. Next, data were collected from a new sample of 1,925 students attending 21 schools from 14 school districts located in the midwest of the USA. The districts included urban, suburban, and rural areas (41, 46, and 13 percent, respectively). The students taking the test were experiencing their first year of biology, both regular and honor students, as well as those taking a second advanced year of biology. The honors students were considered to be higher achieving than the regular students. Table 3 provides a summary of the student data. During this data collection teachers administered the tests. Each student was given one class period, or about 40-45 minutes, to complete the assessment without the use of external assistance of any kind. Following the data collection, a Rasch analysis was conducted to evaluate the measurement functioning of the instrument.

Rasch Analysis of the Final 25 Item HS-BCI

The Rasch analysis entailed the analysis of the instrument's dimensionality, item, and person reliability, step ordering, Wright map, construct validity, and predictive validity.

Table 3. Summary data as a function of the type of biology class

Biology Class Level	Number of students	Number of teachers
Regular (First year)	1485	23
Honors (First year)	345	5
Second Year	125	3

Table 4. Step ordering exemplar (* denotes correct answer)

Item answer alternatives	Number of Respondents	Average Person Measure (logits)
A	217	-1.12
B	464	-0.93
C*	519	-0.54
D	723	-0.75
E	8	-1.06
F	3	-1.53

Dimensionality. An important aspect of evaluating instrument functioning is to determine if the instrument measures a single trait. Two commonly utilized techniques are: 1) the fit of items and respondents (Lincare, 2002) and 2) a PCAR analysis (Lincare, 1998). A review of the items revealed a mean MNSQ Outfit of 1.03 and a mean MNSQ Infit of 1.00. All items exhibited an MNSQ between 0.7 and 1.30. These ranges of item fit fall within acceptable bounds (Boone et al., 2014).

A PCAR of residuals was also conducted. This technique provides a second method to evaluate item dimensionality. An eigenvalue of the first contrast below 2.0 provides evidence that the set of items for an instrument defines a single trait (Linacre, 2018). For this analysis, the eigenvalue of the first contrast was 1.67, supporting the conclusion that the set of items defined a single trait.

Item and Person Reliabilities. Analysis of the collected data set revealed a Rasch item reliability of 1.00 and person reliability of 0.53. Rasch reliabilities vary from a low of 0.00 to a high of 1.00, with a higher value being considered better. Many factors can impact reliability. Factors impacting person reliability include not only the range of the person's ability but also the length of the test, the number of categories per item, and the sample-item targeting but person reliability is largely uninfluenced by model fit and sample size (Malec et al., 2007; Moeini et al., 2016). The item reliability is independent of test length and model fit but does depend on the item difficulty range and sample size (Malec et al., 2007). Malec et al. (2007) suggested a target value of .80 or higher for person reliability and a target value of .90 or higher for item reliability. However, this standard for higher person reliabilities has come into question for concept inventories (Taber, 2018). This will be further discussed in the discussion section.

Step Ordering - Respondent Performance as a Function of Item Answer Alternatives. An evaluation of the performance of students as a function of their answer choice was conducted. The average measure of respondents who correctly answer a test item should be greater than the average measure of respondents who do not correctly answer. There was no occurrence of a higher average student measure for the students who selected an incorrect response than those who selected a correct response. For this analysis, if 10 or fewer respondents used a response category, then the response category average was not utilized for the analysis of average respondent performance as a function of item answer alternatives. This was completed since when one computes an average from 10 or smaller the average is uncertain and can be impacted more greatly by outliers. Table 4 presents an exemplar of this analysis for one of the 25 instrument items. Notable is that the average person measure of students selecting the correct answer, option C (-0.54) is higher than the average of students selecting any of the incorrect answers. This is the pattern one would hope to observe in a well-functioning test.

In addition, the results in Table 4 also demonstrated the need to have test alternatives tied to student alternative conceptions. In this case, the majority of the students have an alternative conception as shown by the higher number of participants choosing the incorrect option D (N=723). Thus, if known by the teacher they can use this to design their lessons to scaffold students towards the correct conception. This shows the need to have distractors in the form of strongly held alternative conceptions to tempt the respondents.

Using the Wright Map to Evaluate Item Placement, Item Targeting, and Floor to Ceiling Effects. Wright Maps are used to evaluate an instrument's functioning (e.g., Boone & Noltemeyer, 2017). Figure 6 presents the Wright Map from the analysis of the HS-BCI. The right side of the Wright Map presents the item difficulty. Items plotted toward the top of the Wright Map are more difficult items while those at the bottom are easier. The left side of the Wright Map presents the person measures of respondents. Those respondents plotted toward the top of the Wright Map are those respondents who are higher performing respondents while those at the bottom are the lower performing students. The "M" plotted on the item side of the Wright Map represents the location of the average difficulty of all items. The "M" plotted on the person side of the Wright Map represents the location of the average person ability of all respondents. The "S" marks one standard deviation

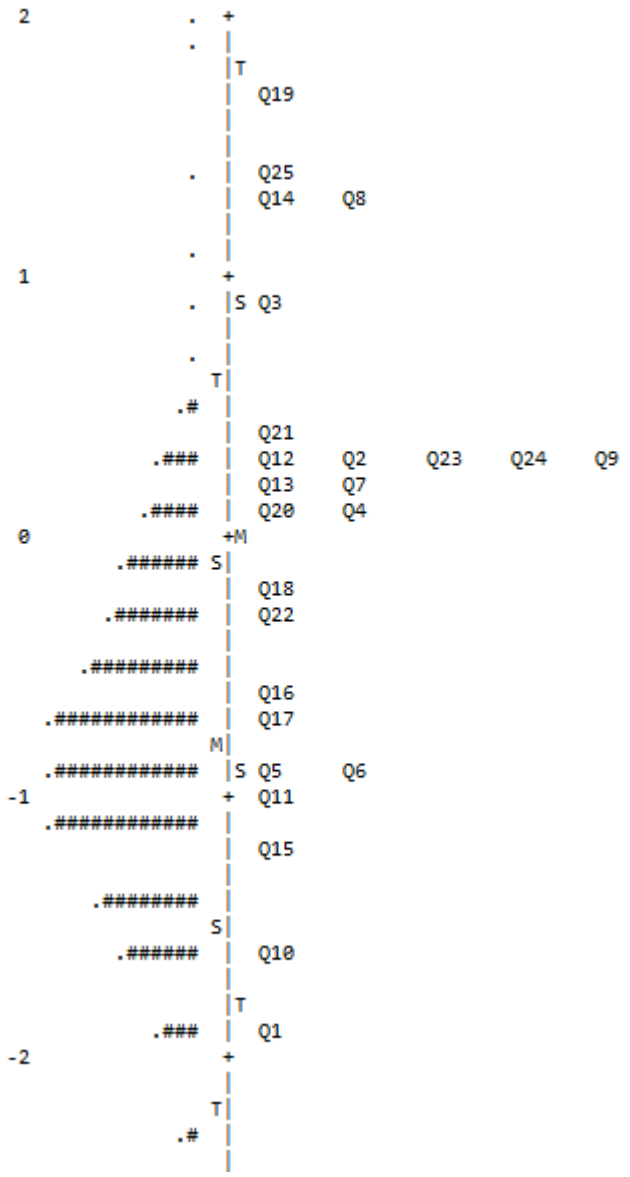

Figure 6. Wright Map for the final 25 item instrument. Items are arranged by Rasch item difficulty on the right side of the map (more difficult items toward the top of the map) and Rasch person ability on the left side of the map (higher-performing students at the top of the map). A mix of difficult items are presented to help measure the growth of students in future versions of the instrument targeting older students

and the "T" marks two standard deviations for the person measures.

A number of components of the Wright Map can be reviewed to help assess the instrument's measurement functioning. One such component is the location of instrument items along the continuum. For a well-functioning instrument, it is important to have instrument items that mark different locations of the trait. This means there should be limited regions of the continuum where there are no items. A review reveals that there is a very good distribution of marks along the continuum; however, for this sample of students, the Wright Map shows some instrument items above the person measures of many sample respondents. For example, Q19, Q25, Q14, Q8, and Q3 are some of the items that were of high difficulty for respondents.

A second technique that is used to evaluate the measurement functioning of an instrument is to evaluate whether there is a good range of item difficulty to help distinguish the performance of test-takers. In other words, are items located near all respondents? The Wright Map revealed that there are a range of items in terms of item difficulty when compared to the location of respondents.

A third technique is an appraisal of the 'test targeting" of an instrument. One method is comparing the location of the mean "M" test item (0.00 logits) and the mean "M" person ability (-0.79 logits). When the two Ms are near each other, that provides a quick assessment of good test item targeting because there are not too many items that are too difficult for respondents, and there are not too many items that are too easy for respondents. A general rule of thumb is that there

If a couple has a one-in-four risk of having a child with an inheritable disease, then _____.

a. the second child born will have a reduced risk of inheriting the disease if their firstborn child has the disease.
b. the firstborn child has the highest risk of inheriting the disease when compared to the second-born child.
c. each child born to this couple will have a one-in-four risk of inheriting the disease.
d. the chances of inheriting the disease will depend on the gender of the child.

Figure 7. The least difficult question

Which of the following organism(s) share a common ancestor with humans?
a. Chimpanzees
b. Bees
c. Oak trees
d. All of the above
e. None of the above

Figure 8. The most difficult question on the HS-BCI

should be less than a 1.00 logit difference between the average item and the average person (Finger et al., 2012). For this study, this was the case. In many ways, the offset of "M"'s seen in the Wright Map reflects what has already been mentioned, namely, there are more difficult items than needed for this sample of respondents. The selection of more difficult items was purposeful as the HS-BCI was designed to be able to reach across levels from beginning high school to second-year biology students. Plus, the more difficult questions were needed to link the HS-BCI to a future college undergraduate biology assessment (UG-BCI).

When targeting is evaluated, another issue to explore is the floor or ceiling effects observed. Velozo et al. (2006) have used a criterion of no more than 5% of respondents at the highest measure or lowest measure possible on an instrument. For the HS-BCI less than 1% of the respondents were located at the minimum or maximum measure possible.

Using the Wright Map to Evaluate Construct Validity. One technique by which the construct validity evidence was evaluated involved an appraisal of the item difficulty ordering shown on the Wright Map. If the ordering matches that predicted from theory, this is evidence supporting the assertion of the construct validity of the instrument (Green et al., 2019). This review was conducted using a team of 3 experts in the field of biology education who compared the ordering of the questions from high, middle to low difficulty. They concluded that the instrument items did follow the predicted ordering of difficulty across the entire instrument. For comparison purposes, this discussion will focus on the most and least difficult items.

When the least difficult items were compared, it was found that they encompass concepts targeted in middle school. For example, the easiest item (Q1) targeted inheritance and variation of traits (see Figure 7). Middle school standards include the use of Punnett squares and diagrams to model genetic variation in offspring (NGSS Lead States, 2013). The second least difficult question (Q10) focused on competitive interactions within ecosystems. The ordering of this question also makes sense since middle school standards require students to predict outcomes based on interactions of biotic and abiotic ecosystem components.

The most difficult question (Q19) targeted common ancestry (see Figure 8). While this topic is taught at the middle and high school levels it has been routinely determined to be a very difficult concept for all grade levels (e.g., AAAS 2061; Berti et al., 2017; Malone et al., 2019; Seoh et al., 2016; Stammen, 2018). Examples from these studies included the finding that students from primary school to college usually believe that a chimpanzee shares a common ancestor with humans while plants do not. Thus, its placement as the most challenging question in the HS-BCI is not surprising.

The other two most difficult questions dealt with biomass (Q25) and energy transfer (Q14). One of these questions asked students to determine where most of the mass of a tree came from as it grew from a seed. This type of biomass question has been determined to be difficult at all educational levels (e.g., Marmaroti & Galanopoulou, 2006; Mintzes et al., 2001). The energy transfer question tests students' ability to connect energy transfer with food chains. While this concept is taught from the lower to upper secondary levels in most countries, students have been shown to have difficulty connecting food chains with energy transfer and flow through an ecosystem (e.g., Barman et al., 1995; Opitz et al., 2017). In addition, energy concepts remain problematic at the college level (Hartley et al., 2011; Lazarowitz & Lieb, 2006). For example, Hartley et al. (2011) found that the majority of undergraduate students from introductory to upper-level biology courses experienced difficulty tracing energy and matter flow in systems. Given that these difficulties have been shown

to be problematic for many years it seems to be a must that teachers in the future are able to understand the conceptions held by their students since these concepts have proven so difficult to address.

Predictive Validity Evidence. A second type of validity analysis is the evaluation of the predictive nature of the instrument (Baghaei, 2008). Predictive validity evidence was attained by determining whether the ordering of person measures from low to high performers, as revealed by the instrument, made sense. The average person measures for each course were computed and compared. This comparison was done using a Welch ANOVA and a Games-Howell post hoc test. It would be predicted that: 1) students in regular 1st-year classes would have lower person measures than students in honors 1st-year classes, and 2) that students in 1st-year honors classes would have lower measures than students of second-year classes. For this statistical analysis, the Rasch person measures were rescaled to 0 to 1000 where 1000 represents the highest measure. The pattern observed between the average group measures was as predicted ($F(2,1952) = 106.88$, $p<.01$). The effect sizes as measured with Cohen's d showed a medium effect between each grade level (0.5 and 0.7, respectively).

DISCUSSION – RASCH ANALYSIS, IMPLICATIONS, AND LIMITATIONS

The main goal of this study was to develop a valid and reliable high school level biology concept inventory that could be used as an assessment tool by teachers and researchers, focused on broad concepts with item construction grounded in alternative conception research. The final HS-BCI is a 25-item MC instrument that can be used to measure student alternative conceptions, as well as assess instructional effectiveness. The instrument avoids biological jargon while assessing student conceptions using distractors based on common alternative conceptions. In addition, the assessment only takes one class period to administer. This assessment is the first high school biology assessment that targets the major alternative conceptions held by students within the context of five core concepts that are included in the developmental progressions outlined by multiple standard documents (e.g., Brewer & Smith, 2011; KMK, 2005; NGSS Lead States, 2013). Having such an instrument to collect student measures of core biological concepts in a single instrument is a step forward for high school biology education.

The second research goal was to evaluate the measurement properties of the HS-BCI using Rasch analysis. The use of Rasch analysis allowed a large number of items to be piloted across two test forms. Rasch item anchoring allowed the difficulty of items, regardless of test form, to be compared on the same single scale. This allowed the purposeful selection of items for the single final test form. Rasch analysis also provided linear person measures, which could then be evaluated with parametric statistical tests. Furthermore, Rasch analysis provided a wide variety of techniques to access the functioning of the instrument. The instrument has been shown to provide multiple forms of validity evidence. The two expert panels supported the validity of the instrument, the ordering of items by difficulty matched experts' predictions, and the measures of different biology classes matched expected class performance. A review of the fit of the instrument items revealed mean MNSQs all within the bounds of acceptability. From a PCAR perspective, the set of items do not suggest more than one trait. The analysis of the average measures of respondents to each answer choice for each item revealed a pattern that would be expected in a well-functioning instrument. The test item targeting fell within the range that is acceptable and no floor to ceiling effect was observed. The analysis revealed a Rasch item reliability of 1.00, which is the highest value an item can achieve. The only limitation of the HS-BCI is the low person reliability score of 0.53. There could be many factors impacting this person reliability value. A number of more difficult items were included to link the assessment to a future UG-BCI. This meant that these items were not as helpful for the computation of the majority of sample person measures as medium difficulty items would have been. The assessment also contains several very easy items that were included to link it to a middle school instrument (MS-LSCI, Stammen, 2018), such easy items did not contribute to the sample person measure computations as well as medium difficulty items would have. Having such very easy and very hard items were important for future linking to easier and harder versions of the instrument. However, having included such easy and harder items may have decreased the person reliability of the data we collected as there were by design fewer middle difficulty items. Other issues that can impact person reliability are the sample ability variance (Linacre, 2018). If a wider range of student ability had been observed a higher person reliability might have been observed. Finally, although our instrument exhibited good targeting, better population targeting might have resulted in higher person reliability (Linacre, 2018). This better targeting would have been achieved through the inclusion of more items in the difficulty range of approximately -.5 logits to -1.5 logits. Such medium difficulty items are items which are at a similar difficulty as that observed for the students with measures around the average student measure. It will be important to assess the person reliability values over time. It is important to mention, however, that high values of person reliabilities might not be desirable when developing an instrument to assess for biological conceptual understanding (Taber, 2018). Thus, the current person reliability values may be sufficient for the goals of the HS-BCI.

These instruments, MS-LSCI, HS-BCI, and UG-BCI, will allow for tracking of common biology alternative conceptions across a broad range of topics unlike targeted assessments (Todd & Romaine, 2016; Todd et al., 2017; Wang, 2004). As shown in the instrument overview there are few instruments that can measure shifts in student conceptual understanding across grade levels such as from middle school to college. This paper only located two possible candidates and they were only developed for specific concepts (i.e., genetics and internal transport) and thus were not comprehensive across an entire course or year.

Finally, the HS-BCI was shown to be able to distinguish student conceptions as predicted across grades and biological course levels. In addition, the assessment items demonstrated that student answer selections mirrored other research in student alternative conceptions. Therefore, the HS-BCI can be utilized to determine how student conceptions shift within the year for multiple biological courses as well as across grade levels.

The development of this assessment has several implications for practice and research. Currently, it is difficult to assess how different curriculum programs affect student conceptions since there is no common assessment tool for this grade level. This instrument will allow for these types of comparisons.

In addition, teachers now have an assessment that they can use only twice a year to assess how their changes in practice shift student knowledge over the course of a single year of study. This ease of assessment presentation, marking and analysis will greatly help to encourage biology teachers to routinely assess the effectiveness of their practice.

As in any study, there are of course some limitations. While a very large sample of students provided data for this study, this does not mean that if other students provided data in the future that the item ordering would be the same. In addition, as with any newly developed instrument, there can be fine-tuning of items needed over time.

CONCLUSIONS AND FUTURE RESEARCH

This study has provided researchers and practitioners with the ability to not only measure high school students' biological content knowledge but also their major prior conceptions through the use of the HS-BCI. The descriptions of the Rasch techniques used to not only develop the instrument but also to evaluate its functioning should provide researchers in all fields with the initial knowledge base to consider utilizing Rasch techniques to develop their own instruments. In addition, there are a number of recent books that can expand on the initial knowledge provided in this paper (Andrich & Marais, 2019; Boone & Staver, 2020; Boone et al., 2014).

The HS-BCI allows for several avenues of future research. Of course, one future venture would be to assess the conceptual effectiveness of year-long curricular innovations in biology. In addition, combining this data with additional data allows us to potentially develop a Rasch measure to raw score conversion table. This would allow those using the HS-BCI to compute Rasch measures of students without conducting a Rasch analysis. This would be very helpful to teachers. The use of linked biology concept inventories across grade bands allows for multiple longitudinal studies. The HS-BCI was the first step towards the construction of three linked instruments, the MS-LSCI (Stammen, 2018) and the future UG-BCI. This set of instruments would allow for comparisons of alternative conceptions across different grades as well as aid curricular efforts that bridge grade levels. The work presented here should help others utilize Rasch techniques to develop new instruments for use in both mathematics and science education.

Author contributions: All authors have sufficiently contributed to the study, and agreed with the results and conclusions.

Funding: This research was partially funded by a grant under the federally funded Math Science Partnership State Grants Program, under Grant number OH160505 and OH160511.

Acknowledgements: We would like to thank James Roth for his work on the illustrations for the HS-BCI.

Declaration of interest: No conflict of interest is declared by authors.

Data sharing: The HS-BCI is available either upon request from the corresponding author or in online supplemental materials of EJMSTE.

REFERENCES

AAAS Project 2061 (n.d.). Pilot and field test data collected between 2006 and 2010 (Unpublished raw data).

Abate, T., Michael, K., & Angell, C. (2020). Assessment of scientific reasoning: Development and validation of scientific reasoning assessment tool. *Eurasia Journal of Mathematics, Science and Technology Education, 16*(12), em1927. https://doi.org/10.29333/ejmste/9353

Adeoye, A. G., & Abimbola, O. I. (2016). Effects of senior school students' use of demo kit on their achievement in biology in Omu-Aran, Nigeria. *Electronic Journal of Science Education*, *20*(8), 88-102. https://ejrsme.icrsme.com/article/view/16226

Anderson D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*, *39*(10), 952-978. https://doi.org/10.1002/tea.10053

Anderson, C. W., Sheldon, T. H., & Dubay, J. (1990). The effects of instruction on college nonmajors'

conceptions of respiration and photosynthesis. *Journal of Research in Science Teaching, 27*(8), 761-776. https://doi.org/10.1002/tea.3660270806

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory. Measuring in the educational, social and health sciences.* Springer. https://doi.org/10.1007/978-981-13-7496-8

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions, 22*(1), 1145-1146.

Barman, C. R., Griffiths, A. K., & Okebukola, P. A. (1995). High school students' concepts regarding food chains and food webs: A multinational study. *International Journal of Science Education*, *17*(6), 775-782. https://doi-org/10.1080/0950069950170608

Berthelsen, B. (1999). Students naïve conceptions in life science. *Michigan Science Teachers Association Journal, 44*(1), 13-19.

Berti, A. E., Barbetta, V., & Toneatti, L. (2017). Third-graders' conceptions about the origin of species before and after instruction: An exploratory study. *International Journal of Science and Mathematics Education*, *15*(2), 215-232. https://doi.org/10.1007/s10763-015-9679-5

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, *4*(1), 1416898.

Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch analyses in the human sciences*. Springer International Publishing.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer Nature.

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

Cary, T. L., Wienhold, C. J., & Branchaw, J. (2019). A Biology Core Concept Instrument (BCCI) to teach and assess student conceptual understanding. *CBE – Life Sciences Education*, *18*(3), ar46. https://doi.org/10.1187/cbe.18-09-0192

Cavalho, J. C. Q. D., Beltramini, L. M., & Bossolan, N. R. S. (2018). Using a board game to teach protein synthesis to high school students. *Journal of Biological Education, 53*(2), 205-216. https://doi.org/10.1080/00219266.2018.1469532

Chan, S. W., Ismail, Z., & Sumintono, B. (2014). A Rasch model analysis on secondary students' statistical reasoning ability in descriptive statistics. *Procedia-Social and Behavioral Sciences*, *129*, 133-139. https://doi.org/10.1016/j.sbspro.2014.03.658

Couch, B. A., Wright, C. D., Freeman, S., Knight, J. K., Semsar, K., Smith, M. K., Summers, M. M., Zheng, Yi, Crowe, A. J., & Brownell, S. E. (2019). GenBio-MAPS: A programmatic assessment to measure student understanding of vision and change core concepts across general biology programs. *CBE – Life Sciences Education*, *18*(1), ar1. https://doi.org/10.1187/cbe.18-07-0117

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2016). Development of the statistical reasoning in biology concept inventory (SRBCI). *CBE – Life Sciences Education*, *15*(1), ar5. https://doi.org/10.1187/cbe.15-06-0131

Ding, L. (2014). Seeking missing pieces in science concept assessments: Reevaluating the brief electricity and magnetism assessment through Rasch analysis. *Physical Review Special Topics-Physics Education Research, 10*(1). https://doi.org/10.1103/PhysRevSTPER.10.010105

Finger, R. P., Fenwick, E., Pesudovs, K., Marella, M., Lamoureux, E. L., & Holz, F. G. (2012). Rasch analysis reveals problems with multiplicative scoring in the macular disease quality of life questionnaire. *Ophthalmology*, *119*(11), 2351-2357. https://doi.org/10.1016/j.ophtha.2012.05.031

Fisher, K. M., Williams, K. S., & Lineback, J. E. (2011). Osmosis and diffusion conceptual assessment. *CBE – Life Sciences Education*, *10*(4), 418-429. https://doi.org/10.1187/cbe.11-04-0038

Gray, J., Kim, J., Ciesla, J. R., & Yao, P. (2014). Rasch Analysis of the Lubben Social Network Scale–6 (LSNS-6). *Journal of Applied Gerontology, 35*(5), 508-528.

Green, A. L., Lambert, M. C., & Hurley, K. D. (2019). Measuring activation in parents of youth with emotional and behavioral disorders. *The Journal of Behavioral Health Services & Research*, *46*(2), 306-318. https://doi.org/10.1007/s11414-018-9627-6

Hartley, L. M., Wilke, B. J., Schramm, J. W., D'Avanzo, C., & Anderson, C. W. (2011). College students' understanding of the carbon cycle: Contrasting principle-based and informal reasoning. *BioScience*, *61*(1), 65-75. https://doi.org/10.1525/bio.2011.61.1.12

Haslam, F., & Treagust, D. F. (1987). Diagnosing secondary students' misconceptions of photosynthesis and respiration in plants using a two-tier multiple choice instrument. *Journal of Biological Education*, *21*(3), 203-211. https://doi.org/10.1080/00219266.1987.9654897

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *Physics Teacher, 30*, 141-158. https://doi.org/10.1119/1.2343497

Hogan, K. (2000). Assessing students' systems reasoning in ecology, *Journal of Biological Education*, *35*(1), 22-28. https://doi.org/10.1080/00219266.2000.9655731

Kalas, P., O'Neill, A., Pollock, C., & Birol, G. (2013). Development of a meiosis concept inventory. *CBE—Life Sciences Education*, *12*(4), 655-664. https://doi.org/10.1187/cbe.12-10-0174

Kalinowski, S. T., Leonard, M. J., & Taper, M. L. (2016). Development and validation of the Conceptual Assessment of Natural Selection (CANS). *CBE-Life Sciences Education*, *15*(4), ar64. https://doi.org/10.1187/cbe.15-06-0134

Klymkowsky, M. W., & Garvin-Doxas, K. (2020). Concept inventories: Design, application, uses, limitations, and next steps. In *Active Learning in College Science* (pp. 775-790). Springer. https://doi.org/10.1007/978-3-030-33600-4_48

Klymkowsky, M. W., Underwood, S. M., & Garvin-Doxas, R. K. (2010). Biological Concepts Instrument (BCI): A diagnostic tool for revealing student thinking. *arXiv preprint arXiv:1012.4501*

KMK [Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der BRD] (Ed.). (2005). Bildungsstandards im fach biologie für den Mittleren Schulabschluss [Biology education standards for the Mittlere Schulabschluss]. Wolters Kluwer.

Lamb, R. L., Annetta, L., Meldrum, J., & Vallett, D. (2012). Measuring science interest: Rasch validation of the science interest survey. *International Journal of Science and Mathematics Education*, *10*(3), 643-668. https://doi.org/10.1007/s10763-011-9314-z

Lazarowitz, R., & Lieb, C. (2006). Formative assessment pre-test to identify college students' prior knowledge, misconceptions and learning difficulties in biology. *International Journal of Science and Mathematics Education*, *4*(4), 741-762. https://doi.org/10.1007/s10763-005-9024-5

Lin, S. W. (2004). Development and application of a two-tier diagnostic test for high school students' understanding of flowering plant growth and development. *International Journal of Science and Mathematics Education*, *2*(2), 175-199. https://doi.org/10.1007/s10763-004-6484-y

Linacre J.M. (2001). Category, step and threshold: Definitions & disordering. *Rasch Measurement Transactions*, *15*(1), 794.

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal components analysis. *Rasch Measurement Transactions*, *12*(2), 636.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean?. *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2018). *Winsteps ministep: Rasch-model computer programs*. https://www.winsteps.com/winman/copyright.htm

Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch modeling approach*. Information Age Pub.

Malec, J. F., Torsher, L. C., Dunn W. F., Wiegmann, D. A., Arnold, J. J., Brown, D. A., & Phatak V. (2007). The Mayo high performance teamwork scale: Reliability and validity for evaluating key crew resource management skills. *Journal of the Society for Simulation in Healthcare*, 2(1), 4-10. https://doi.org/10.1097/SIH.0b013e31802b68ee

Malone, K. L., Schuchardt A. M., & Sabree, Z. (2019). Models and modeling in evolution. In U. Harms, & M. J. Reiss (Eds), *Evolution education re-considered: understanding what works* (pp. 207-226). Springer International Publishing.

Marmaroti, P., & Galanopoulou, D. (2006). Pupils' understanding of photosynthesis: A questionnaire for the simultaneous assessment of all aspects. *International Journal of Science Education*, *28*(4), 383-403. https://doi.org/10.1080/09500690500277805

Minner, D., Ericson, E., Wu, S., & Martinez, A (2012, November). *Compendium of research instruments for STEM education part 2: Measuring students' content knowledge, reasoning skills, and psychological attributes.* http://www.cadrek12.org/resources/compendium-research-instruments-stem-education-part-ii-measuring-students-content-knowledg

Mintzes, J. J., Wandersee, J. H., & Novak, J. D. (2001). Assessing understanding in biology. *Journal of Biological Education*, *35*(3), 118-124. https://doi.org/10.1080/00219266.2001.9655759

Moeini, S., Rasmussen, J. V., Klausen, T. W., & Brorson, S. (2016). Rasch analysis of the Western Ontario Osteoarthritis of the Shoulder Index–the Danish version. *Patient Related Outcome Measures*, *7*, 173. https://doi.org/10.2147/PROM.S87048

Nadelson, L. S., & Southerland, S. A. (2009). Development and preliminary evaluation of the measure of understanding of macroevolution: Introducing the MUM. *The Journal of Experimental Education*, *78*(2), 151-190. https://doi.org/10.1080/00220970903292983

NGSS Lead States. 2013. *Next Generation Science Standards: For States, By State*s. The National Academies Press.

Opitz, S. T., Blankenstein, A., & Harms, U. (2017). Student conceptions about energy in biological contexts. *Journal of Biological Education*, *51*(4), 427-440. https://doi.org/10.1080/00219266.2016.1257504

Price, R. M., Andrews, T. C., McElhinny, T. L., Mead, L. S., Abraham, J. K., Thanukos, A., & Perez, K. E. (2014). The genetic drift inventory: A tool for measuring what advanced undergraduates have

mastered about genetic drift. *CBE – Life Sciences Education*, *13*(1), 65-75. https://doi.org/10.1187/cbe.13-08-0159

Pugh, K. J., Koskey, K. L., & Linnenbrink-Garcia, L. (2014). High school biology students' transfer of the concept of natural selection: A mixed-methods approach. *Journal of Biological Education*, *48*(1), 23-33. https://doi.org/10.1080/00219266.2013.801873

Schmeiser, C., B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational Measurement*. Praeger Publishers.

Seoh, K. H. R., Subramaniam, R., & Hoh, Y. K. (2016). How humans evolved according to grade 12 students in Singapore. *Journal of Research in Science Teaching*, *53*(2), 291-323. https://doi.org/10.1002/tea.21256

Shi, J., Wood, W. B., Martin, J. M., Guild, N. A., Vicens, Q., & Knight, J. K. (2010). A diagnostic assessment for introductory molecular and cell biology. *CBE – Life Sciences Education*, *9*(4), 453-461. https://doi.org/10.1187/cbe.10-04-0055

Stammen, A. (2018). *The development and validation of the Middle School-Life Science Concept Inventory (MS-LSCI) using Rasch Analysis* (Doctoral dissertation, Ohio State University).

Stammen, A., Lan, D., Schuchardt, A., Malone, K., Ding, L., Sabree, Z., & Boone, W. (2016). Development of the Secondary-Biology Concept Inventory (S-BCI): A study of content and construct validation. In *ICMST Conference Committee (Ed.), Education Research Highlights in Mathematics, Science and Technology 2016*, Egiten Publishing.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Todd, A., & Romine, W. L. (2016). Validation of the learning progression-based assessment of modern genetics in a college context. *International Journal of Science Education*, *38*(10), 1673-1698. https://doi.org/10.1080/09500693.2016.1212425

Todd, A., Romine, W. L., & Cook Whitt, K. (2017). Development and validation of the learning progression–based assessment of modern genetics in a high school context. *Science Education*, *101*(1), 32-65. https://doi.org/10.1002/sce.21252

Treagust, D. (1986). Evaluating students' misconceptions by means of diagnostic multiple choice items. *Research in Science Education*, *16*(1), 199-207. https://doi.org/10.1007/BF02356835

Treagust, D. F., & Mann, M. (1998). A pencil and paper instrument to diagnose students' conceptions of breathing, gas exchange and respiration. *Australian Science Teachers Journal*, *44*(2), 55-59.

Tsui, C. Y., & Treagust, D. (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education*, *32*(8), 1073-1098. https://doi.org/10.1080/09500690902951429

Velozo, C. A., Choi, B., Zylstra, S. E., & Santopoalo, R. (2006). Measurement qualities of a self-report and therapist-scored functional capacity instrument based on the Dictionary of Occupational Titles. *Journal of Occupational Rehabilitation, 16* (1), 109-122. https://doi.org/10.1007/s10926-005-9014-z

Wang, J. R. (2004). Development and validation of a two-tier instrument to examine understanding of internal transport in plants and the human circulatory system. *International Journal of Science and Mathematics Education,* *2*(2), 131-157. https://doi.org/10.1007/s10763-004-9323-2

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Mesa Press.

Yang, Y., He, P., & Liu, X. (2018). Validation of an instrument for measuring students' understanding of interdisciplinary science in grades 4-8 over multiple semesters: A Rasch measurement study. *International Journal of Science and Mathematics Education, 16*(4), 639-654. https://doi.org/10.1007/s10763-017-9805-7

Young, C. A., Quincey, A. M. C., Wong, S. M., & Tennant, A. (2018). Quality of life for post-polio syndrome: A patient derived, Rasch standard scale. *Disability and Rehabilitation, 40*(5), 597-602. https://doi.org/10.1080/09638288.2016.1260650