education sciences

MDPI

Article

Describing the Development of the Assessment of Biological Reasoning (ABR)

Jennifer Schellinger [1,*], Patrick J. Enderle [2], Kari Roberts [3], Sam Skrob-Martin [1], Danielle Rhemer [1] and Sherry A. Southerland [1]

[1] School of Teacher Education, Florida State University, 1114 W Call St, Tallahassee, FL 32306, USA; sks14b@my.fsu.edu (S.S.-M.); dvandezande@fsu.edu (D.R.); ssoutherland@admin.fsu.edu (S.A.S.)
[2] Department of Middle and Secondary Education, Georgia State University, Atlanta, GA 30302, USA; penderle@gsu.edu
[3] Center for Integrating Research and Learning, National High Magnetic Field Laboratory, Tallahassee, FL 32310, USA; kari.roberts@magnet.fsu.edu
* Correspondence: jls09h@fsu.edu

Abstract: Assessments of scientific reasoning that capture the intertwining aspects of conceptual, procedural and epistemic knowledge are often associated with intensive qualitative analyses of student responses to open-ended questions, work products, interviews, discourse and classroom observations. While such analyses provide evaluations of students' reasoning skills, they are not scalable. The purpose of this study is to develop a three-tiered multiple-choice assessment to measure students' reasoning about biological phenomena and to understand the affordances and limitations of such an assessment. To validate the assessment and to understand what the assessment measures, qualitative and quantitative data were collected and analyzed, including read-aloud, focus group interviews and analysis of large sample data sets. These data served to validate our three-tiered assessment called the Assessment of Biological Reasoning (ABR) consisting of 10 question sets focused on core biological concepts. Further examination of our data suggests that students' reasoning is intertwined in such a way that procedural and epistemic knowledge is reliant on and given meaning by conceptual knowledge, an idea that pushes against the conceptualization that the latter forms of knowledge construction are more broadly applicable across disciplines.

Keywords: scientific reasoning; biological reasoning; assessment; three-tiered assessment; Assessment of Biological Reasoning

Citation: Schellinger, J.; Enderle, P.J.; Roberts, K.; Skrob-Martin, S.; Rhemer, D.; Southerland, S.A. Describing the Development of the Assessment of Biological Reasoning (ABR). *Educ. Sci.* **2021**, *11*, 669. https://doi.org/10.3390/educsci11110669

Academic Editors: Moritz Krell, Andreas Vorholzer and Andreas Nehring

Received: 14 July 2021
Accepted: 6 October 2021
Published: 21 October 2021

1. Introduction

Enhanced learning in science moves beyond memorization and recitation of fundamental concepts to encompass a much larger collection of sense-making activities that resemble the cognitive, procedural, epistemic and social work of scientists [1–3]. Although investigative activities occur in science classrooms in myriad ways, they often limit or even neglect to deeply engage students in the explanatory and evaluative spheres of the scientific enterprise that are essential to the development of scientific understandings [3–6]. Greater emphasis on engaging students in practices reflecting the investigative, explanatory and evaluative spheres of science require supporting students in understanding not only the conceptual elements involved but also the procedural and epistemic function of such practices [3,7]. Such learning not only helps students participate in the development of evidence-based arguments, explanations and models, but also helps them learn to evaluate the quality of different elements of these products and how the processes involved in developing them connect with each other [8,9]. Thus, science learning grounded in these practices also necessitates engaging students in various forms of scientific reasoning where they connect these different activities and products of science in complex yet coherent

ways [10]. We define scientific reasoning as the process that encompasses "the skills involved in inquiry, experimentation, evidence evaluation and inference that are done in the service of conceptual change or scientific understanding" [11] (p. 172), a process that brings together conceptual (i.e., content), procedural and epistemic aspects of knowledge [3,12].

Research on learners' engagement in scientific reasoning activities demonstrates the complexity of such processes, particularly as they engage in the evaluative and explanatory aspects of science [3,12]. A multitude of factors can shape how students engage in scientific practices that serve as manifestations of scientific reasoning. Scholars have pointed to the need to create time and space in classrooms where students are afforded opportunities to develop the epistemic agency required to engage in reasoning activities to construct knowledge [13]. Research into students' participation in episodes of critique highlight structural and dialogical elements of argumentation activities reliant on scientific reasoning [14]. Studies of instruction centered around students developing and refining scientific models demonstrate that the concepts that serve as the cognitive objects involved in their reasoning must have a robust quality before students can connect them to broader conceptual elements of models [15]. Enhancing students' reasoning using scientific models requires engaging their creativity, while also supporting their ability to understand the multiple goals that models can help achieve [16]. However, it is important to note that students' proficiency with procedural aspects of scientific reasoning, including experimentation and data analysis, are supportive of their learning of conceptual objects and epistemological characteristics [17–19]. There is some evidence to suggest that the cognitive and motivational characteristics of students are also predictive of their ability to reason across broader disciplinary contexts [5].

Much of the research into students' scientific reasoning when engaged in scientific practices involves intensive qualitative analytical approaches that rely on products resulting from relevant activities [15,20,21]. The composition and quality of students' arguments [21], models [15,16] and constructed responses to open-ended questions [20,22] can be coded by multiple raters to inductively develop thematic findings or deductively assess the alignment of students' products to theoretically derived frameworks. Such analyses can be extended by or complemented through separate explorations of students' reasoning as they are engaged in various types of individual interviews, which are then qualitatively coded [15,23]. Other researchers explore students' reasoning in action, relying on various analytical approaches employing discourse analysis [14] or observation protocols [6,17] that still necessitate qualitative coding or scoring approaches amongst multiple raters.

Another influential aspect of these analytical approaches concerns the conceptual and disciplinary contexts within which they occur. Many of the studies identified remain tied to particular conceptual areas within specific scientific disciplines, including thermal conductivity in chemistry [21], evolutionary theory or genetics in biology [22,23] and carbon cycling and climate change in Earth science [15]. Limited studies exist where researchers have employed more scalable, quantitative instruments that explore connections between students' scientific reasoning and broader disciplinary contexts [5], multiple reasoning competencies and skills that can be employed internationally [24,25] and measure competencies among various age ranges [26]. Assessments that do exist have been criticized because they are not psychometrically sound [26]. Additionally, most large-scale measures across various disciplines remain focused on students' conceptual understanding, limiting the inferential capacity of such work to gain understanding about students' scientific reasoning [27–30]. Thus, measuring students' scientific reasoning across a discipline and across dimensions of scientific reasoning through more scalable quantitative approaches remains an ongoing challenge for science education research, something that limits the research that can be conducted.

In light of these challenges, this study focuses on the iterative development and validation of a multiple-choice instrument using qualitative and psychometrically sound quantitative approaches aimed at assessing dimensions of students' scientific reasoning across ten focal topic areas within biology, entitled Assessment of Biological Reasoning

(ABR). As part of a broader study exploring the influence of teachers sustaining productive classroom talk on student sensemaking [31], the effort described here involved adapting a previously used measure of students' ability to construct scientific explanations through two-tiered, open-ended questioning [17]. Using the instrument and previously analyzed student response data, we developed a three-tier multiple choice assessment exploring each biological topic through a conceptually oriented first tier, a procedural explanatory second tier and a newly developed epistemic third tier exploring students' reasoning supporting their scientific explanations. The study presented below was guided by the following research questions:

What does a three-tier multiple choice assessment measure about students' scientific reasoning across a variety of scenarios relying on fundamental biology concepts?

What are the affordances and limitations of using this approach to measure students' scientific reasoning in biology?

2. Literature Review

2.1. Scientific Reasoning

Scientific reasoning, a central feature of scientific sensemaking, has suffered from the absence of a coherent definition. Early conceptualizations of scientific reasoning present reasoning as a process by which one can develop understandings of science by controlling variables and making causal inferences based on the outcomes of those tests [32,33]. This model, which closely aligns with one methodological approach of science, that of controlled experimentation, represents an overly narrow view of science [34] and does not capture the complex set of reasoning strategies encompassed in the coordination of theories (prior knowledge and beliefs) and evidence needed to generate new knowledge [35,36].

The examination of how these strategies interact and inform one another requires that one engages in the investigative, evaluative and explanatory spheres of science described by Osborne [37] as the spheres that position students to address questions such as "What is nature like?", "Why does it happen?", "How do we know?" and "How can we be certain?" (p. 181). As students engage in exploring these questions, they make observations to understand natural phenomena and to figure out why something happens by constructing and testing models and explanatory hypotheses through empirical investigations and/or data collection that serves as a basis for argumentation and critique, a process by which students consider explanations, the strength of those explanations and how those explanations are supported by evidence [38,39]. When students come to interact in all aspects of these spheres, they are positioned to better engage in a more holistic representation of reasoning which includes conceptual (i.e., content), procedural and epistemic aspects of knowledge [3,12].

Discussions as to whether such reasoning is broadly applicable across domains or is domain-specific exist. Shavelson [40] argued that scientific reasoning can be used when considering everyday decisions. Chinn and Duncan [41] argue that such applicability can be applied to evaluate the trustworthiness of claims about larger scientific issues (e.g., global climate change) presented by the scientific community. These arguments connect with ideas that many of the reasoning aspects, such as a claim that must be supported by evidence, occur across disciplines (e.g., history and literature).

Other argue that scientific reasoning is domain-specific. Samarapungavan [42] suggests that epistemic reasoning is tied to the role of evidence (i.e., what counts as evidence in a knowledge claim, to what extent does it count and why does it count) in bridging conceptual knowledge with practice within a specific disciplinary context. Kind and Osborne [12] describe conceptual (i.e., content), procedural and epistemic aspects of scientific reasoning to require domain-specific concepts or the ontological entities of a disciple to answer questions about "What exists?", the procedures and constructs that help establish knowledge claims and answer causal questions about "Why it happens?" and the epistemic constructs, values and applications that support the justification of these knowledge claims to answer questions about "How do we know?" (p. 11).

Whether domain specific or broadly applicable, scientific reasoning that connects conceptual, procedural and epistemic aspects of knowledge push against the traditional focus in K-12 education of correctly reciting information about content [43]. Instead, by emphasizing these forms of knowledge, students are asked to demonstrate an understanding of content in ways that integrate how they know what they know. For example, when the object of reasoning is to understand whether species are living or nonliving things (i.e., conceptual), students must understand criteria for separating these species (i.e., procedural) and they must understand the role that categorization serves in identifying distinguishing characteristics of living from nonliving things and the particular constructs needed to explain the phenomena (i.e., epistemic [12]).

2.2. Reasoning in Practice

Curriculum and instruction in recent years have focused on positioning students to engage with the forms of knowledge construction involved in reasoning through such activities as model-based and argumentation-driven inquiry. Zagori et al. [15] and others [44–48] suggest that models serve as tools for reasoning because they are developed based on prior knowledge, they are used to make prediction and to generate scientific explanation about how and why a phenomenon works, they are informed based on data collected through investigations and observations and they serve as artifacts of new understandings when the initial model is evaluated and revised. Zagori and her colleagues [15] conducted a quasi-experimental comparative study to understand how modeling-enhanced curricular interventions supported students' model-based explanations (e.g., conceptual understanding and reasoning). They found that students had statistically significant gains in their model-based explanations about water and geosphere interactions as measured through a pre- and post-unit modeling task when supported with a rigorous curricular intervention that provided opportunities for students to engage in scientific modeling practices (intervention 2) compared to an intervention that provided only pre- and post-unit supplementary lessons and tasks involving modeling (intervention 1). These findings were based on a quantitative score for each student across five epistemic features of modeling, including components (i.e., model elements), sequences (i.e., component relationships), mapping (i.e., relationship of model to the physical world), explanatory process (i.e., the connections articulated between cause and effect of system processes) and scientific principle (i.e., connections to underlying scientific theory). When examining these results further, the researchers noted that the features of components and explanatory processes explained the difference in the aggregated feature scores. While the scores of these particular features helped explain students' gains in model-based explanations, they provided less insight into how students themselves conceptualized these and other features in their models. To further understand the results, the researchers examined students' scores on the components and explanatory process features in conjunction with student interview data. One key finding from this examination was that students' models served as reasoning tools to explain how and why water flows underground when students' models included hidden elements under the subsurface of the earth. Swartz and colleagues [47], similarly, found that models can serve as reasoning tools in which students improve their understandings and develop new knowledge that encompasses the explanatory mechanisms and relationships between components of a phenomenon, findings that required the analysis of construct maps and focus group interviews to understand how students construct and use models.

2.3. Assessments of Reasoning

We present these studies not only to acknowledge that efforts are being made in science education to provide opportunities to engage students in scientific reasoning but also to acknowledge the effort and work required to assess students' reasoning capabilities. Such assessments require qualitative examination of student work products (e.g., models, drawings, written work and answers to open-response questions), student interviews, students' discourse and engagement in reasoning tasks and activities [4,14,15,17,23]. Similar

effort and work is required in assessing students' reasoning capabilities in argumentation, the results of which highlight that students often struggle to understand why the construction and generation of claims based on evidence are necessary for science learning [49–51], to analyze and discern quality evidence to substantiate their claims [52,53] and to provide justification for the relationship between claims and evidence to support their argument [50–55].

These findings, while useful in helping us understand students' reasoning capabilities, many of which are tied to specific concepts within a scientific discipline (e.g., groundwater and water systems), are not necessarily sustainable or scalable. In response to issues of scale that go beyond just measuring conceptual understanding, a prominent feature of many large-scale assessments [27–30], instruments to measure students' scientific reasoning have been developed [26]. In a review of 38 test instruments measuring scientific reasoning, Opitz and colleagues [26] found that most tests were related to reasoning skills associated with hypothesis generation, evidence generation, evidence evaluation and drawing conclusions within specific scientific domains, biology being the most common (N = 13). They found that newer assessments, those developed from 2002 to 2013 (N = 27), measure scientific reasoning competencies as a coordinated set of domain-specific skills as compared to the older assessments (N = 11 developed from 1973 to 1989). Additionally, they found that, of the newer assessments, only 17 reported reliability measures and fewer reported validity measures, a finding that led the authors to call the "overall state of psychometric quality checks" unsatisfactory (p. 92). Only 14 of the 38 tests were multiple choice and most were of a closed format following a tiered structure.

Tiered assessments present interconnected questions such as two-tiered assessments that measure content knowledge in tier 1 and related, higher order thinking and explanatory reasoning in tier two [56–59]. For instance, Strimaitis and colleagues [60] developed a two-tiered multiple-choice instrument to measure students' abilities to critically assess scientific claims in the popular media. The 12-item assessment presented students with two modified articles (i.e., dangers of high heels and energy drinks) and asked them to evaluate aspects of the claims presented in each article (tier one) and the logic (tier two) they used to determine their response to tier one. Such tests not only provide opportunities to quantitatively measure students' underlying reasons for their answer choices but they also provide opportunities to assess the alternative conceptions that many students hold related to the particular topic being assessed [61].

While a two-tiered assessment can provide a diagnostic measure of student content knowledge and their explanatory reasoning related to that knowledge, it can suffer from over- or under-estimations of student conceptions [62] or alternative conceptions [63–65], meaning it can fail to differentiate mistakes from such things as lack of knowledge or correct answers due to guessing [66]. To account for these estimation errors, instruments with three and four tiers have been developed. Three-tiered assessments add a third item that provides a measure of the student's confidence in their answer to the first two content and reasoning items [63]. Four tier assessments add additional items to measure the test takers confidence in their prior answers. In a four-tiered assessment, tier one measures content knowledge, tier two measures the student's level of confidence in their answer to tier one, tier three measures reasoning for tier one and tier four measures the student's confidence related to their reasoning in tier three [64]. The inclusion of additional tiers to assess confidence serves as a measure of the student's belief in their own accuracy and provides a level of validity to their answers [67]; however, these tiers do not provide additional measures of a student's higher order reasoning skills nor do they attend to the interrelated conceptual, procedural and epistemic aspects of scientific reasoning that can be difficult to assess quantitatively and are not often assessed in this way.

Informed by the previous work that has been conducted in terms of assessments of students reasoning and motivated by a need for a psychometrically sound measure of students' content knowledge and reasoning skills in biology, this research study focuses

on the development, fine-grained analysis and validation of a multiple-choice instrument aimed at assessing students' scientific reasoning across ten focal topic areas within biology.

3. Methods

This research project is part of a broader professional development study focused on supporting biology teachers' practice to engage students in scientific reasoning through productive scientific discourse [31]. The goal of this assessment is to measure students' explanation of biological phenomena. This assessment was developed based on an existing constructed response assessment used to measure students' conceptual knowledge in biology necessary to evaluate scientific claims [17]. Major concepts in the discipline were selected as foci for the questions, allowing the instrument to serve as an assessment of student learning in both secondary and post-secondary biology courses. The topics that the assessment addresses include cell theory, meiosis, mitosis, photosynthesis and cellular respiration, nutrient cycling, species concepts, evolution and natural selection.

This assessment was designed to understand three dimensions of students' biological reasoning of the 10 focal phenomena listed above operationalized within the four styles of reasoning put forth by Kind and Osborne [12]. These styles include experimental evaluation, hypothetical modeling, categorization and classification, and historical-based evolutionary reasoning, and represent key practices in scientific knowledge generation. Experimental evaluation relates to empirical investigations to establish patterns, differentiate objects and test predictions. Three focal topics fall within this style, including respiration, natural selection and photosynthesis. Hypothetical modeling relates to the construction of models. The focal topics of Medelian genetics, mitosis and evolution fall within this style. Categorization and classification relate to ordering based on variety and taxonomy. Biological species concept and cell theory align with this style of reasoning. Lastly, historical-based evolutionary reasoning relates to the construction of historical derivations of explanations and development, which include meiosis and nutrient cycling.

The three dimensions of biological reasoning were operationalized within each of these styles of reasoning, including conceptual knowledge (i.e., object of reasoning), procedural knowledge (i.e., use of conceptual knowledge required for reasoning within a specific context) and epistemic knowledge (i.e., ability to justify conclusions based the application of that knowledge). To allow for this structure, each question was framed with an introductory scenario targeting the focal phenomenon with relevant imagery, including graphics, tables, or charts. The first item of the 3-tier question was directed at understanding students' knowledge of specific biological concepts relevant to the focal phenomenon. The second question was aimed at students' use of knowledge, or their application of biological concepts to develop explanations for the focal phenomenon. Finally, the third question asked students to apply reasoning for their explanation by asking them to indicate how relevant biological concepts lead to the explanation of the focal phenomenon. Each tiered question had four answer choices that included a correct choice and distractors, which were developed from expert responses and/or known student responses from previous assessments.

Assessments of this nature should be validated for research purposes with the participant populations that they are intended to be used with. Although multiple views exist on the specific procedures that should be followed for developing educational testing instruments [68–70], a shared consensus suggests that varied pieces of evidence should be collected to demonstrate the properties of an instrument and the validity of the instruments' measurements. Figure 1 provides a graphic identifying the multiple lines of evidence we developed to demonstrate the validity of the ABR. For construct validity, we relied on the input of experts to develop the instrument items and assess how well items measured the targeted, theoretically grounded biological constructs. Experts were comprised of five of the six authors and one high school biology teacher. Of the five authors, three hold two post-secondary degrees in biology and two hold post-secondary degrees in biology and in education. One of the experts holding a post-secondary degree in biology and in education

was also a teacher, represented as teacher #2 in Section 3.1.4. Additionally, the high school biology teacher (Teacher #1), who administered the assessment in her class (discussed in Section 3.1.4), has both a teaching credential and a doctorate in biology. For criterion-related validity, we recruited participants from different populations that theoretically differ in their learning about the focal biological concepts. Finally, we conducted several procedures that improved and demonstrated the reliability of the developed items, including their interpretability by participants, analysis of distractor responses and the internal consistency of the items. We also examined the factor structure of the respondent data to explore how the scores from the instrument should be interpreted.

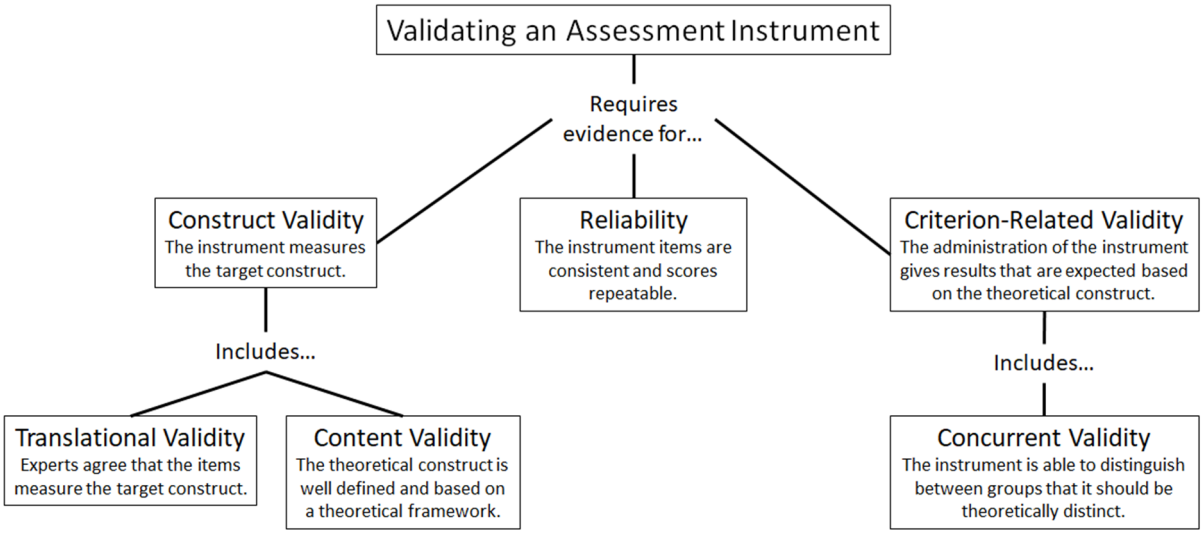

Figure 1. Validation framework used for the Assessment of Biological Reasoning (ABR).

3.1. Data Collection

3.1.1. Exploration of Wording and Coherency Issues

Two rounds of initial testing, including self-recorded read-alouds (see Section 3.1.2) and focus group (see Section 3.1.3) interviews, were conducted to identify possible wording and coherency issues in the 10, three-tiered questions. Each round consisted of a qualitative focus on students' understanding of what each question was asking and the ideas they used to answer the question.

3.1.2. Read-Aloud Interviews

The initial round of testing occurred through self-recorded read-alouds taking approximately from 30 min to 1 h. Seven participants took part in individual read-alouds; one individual had completed high school Biology Honors, three participants were high school biology students and the other three were enrolled in a post-secondary General Biology Laboratory course for non-biology majors. During the read-aloud, each participant read each test item aloud, discussed how they answered the item (e.g., how they arrived at the right choice and why they eliminated certain answer choices), they identified any parts they found difficult or confusing and they made suggestions for item improvement.

3.1.3. Focus Group Interviews

Two virtual focus group interviews were conducted through Zoom. Two participants, one high school Biology and one post-secondary General Biology Laboratory student, took part in the first focus group, which took 2 h and 20 min. During this focus group, the participants answered, annotated and discussed 6 of the 10 questions, including Mendelian genetics, natural selection, nutrient cycling, cell theory, photosynthesis and mitosis. Because of time limitations, the participants in this group answered and made notes on the remaining four questions not addressed in the meeting within one day of the interview. The

second focus group interview had six participants—one high school Biology Honors, two high school Biology and three post-secondary General Biology Laboratory students—and took 2 h and 15 min to conduct. The sequence of the questions was changed for this interview to ensure that feedback for the items that the first group did not have time for in their focus group were examined. In this case, participants answered, annotated and discussed questions related to meiosis, evolution, species concepts and respiration before answering the six questions that the first group started with (i.e., Mendelian genetics, natural selection, nutrient cycling, cell theory, photosynthesis and mitosis).

Both focus group interviews were led by the first author. She followed the same protocol for each interview. In this protocol, participants were introduced to the general structure of the assessment (i.e., three-tiered), they were provided a link to an individual Google document with the assessment questions and then they were asked to work through one three-tiered question individually before coming back together to discuss the question as a group. Students were asked to annotate questions indicating the correct answer, the pieces of the questions that helped them arrive at their answer and to mark any parts that were confusing. Once all students completed the question, the interviewer asked all participants to describe how they solved the problem, the essential pieces of the question that helped them answer it and whether they found any parts of the question or the language of the question difficult, challenging, or confusing. This pattern continued until students cycled through all or most questions. At the end of the interview, participants were asked to discuss if they noticed any changes in how they thought about or read the item for each question and if there were any directions or markers that they wished they had been provided when answering the questions.

3.1.4. Large Sample Data Collection

After completion of the qualitative analysis of the assessment, the instrument was administered to a larger population of students in two rounds to identify if there were any problematic items that potentially needed adjustments. Each round required students to complete the assessment and these data were analyzed for internal consistency.

The first round of analysis focused on examining student assessment data from two teachers (Table 1), one who taught high school Advanced Placement (N = 45 students) and International Baccalaureate Biology classes (N = 15 students) and one who taught a post-secondary General Biology Laboratory course (N = 27 students). The purpose of this analysis was to determine if there were any problematic items and if any items needed to be adjusted. This round also allowed for an examination of item distractors to ensure that they aligned with the internal consistency analyses.

Table 1. Participant information for round 1 data collection.

Teacher	School Type	Course Title	Number of Students
1	high school	Advanced Placement Biology	45
1	high school	International Baccalaureate Biology	15
2	post-secondary	General Biology Laboratory	27

The second round of testing focused on completing final factor analysis, as well as a reexamination of distractor and consistency data. In the distractor analyses, any item with more students selecting an incorrect item choice than the correct item choice was flagged for follow-up review. For internal consistency, Cronbach's alpha was calculated for each scale and alpha values for the scale if each item was removed. We looked for any items where the deletion of the item would increase the scale reliability. More details on the analysis and results can be found in Sections 3.2 and 4. For this purpose, data from three teachers' classrooms collected at the end of the semester were included in the data set. Data were collected from two high school biology teachers (Table 2), one of which participated in the first round of quantitative data collection who taught Advanced Placement (N = 72 students) and International Baccalaureate (N = 20 students) Biology courses and one teacher,

denoted as teacher #3 in Table 2, who taught Advanced Placement Biology (N = 7 students). Additionally, data were collected from teacher #2 who participated in the first round of quantitative data collection. Seven post-secondary students enrolled in her General Biology Laboratory took the assessment in this round.

Table 2. Participant information for round 2 data collection.

Teacher	School Type	Course Title	Number of Students
1	high school	Advanced Placement Biology	72
1	high school	International Baccalaureate Biology	20
2	post-secondary	General Biology Laboratory	7
3	high school	Advanced Placement Biology	7

3.2. Data Analyses

3.2.1. Qualitative Analyses of Individual and Group Interviews

Transcripts were produced for each individual read-aloud and focus group interview for the relevant analyses. For the individual read-alouds, the transcripts were analyzed by the team to identify areas that needed to be clarified, changed, or improved in the assessment. The research team reviewed the participants' responses to the assessment items and the reflective questions concerning clarity of the text and conceptual coherence. As each item and question set was reviewed, the research team identified specific similar challenges mentioned by at least 3–4 students. Similarly, transcripts of the focus group responses for each question set were reviewed. With these transcripts, any issue that maintained the focus of the group's discussion for a significant amount of time was given priority. For the analysis of the individual read-aloud interviews, the research team maintained a stronger emphasis on clarity of the text and how well participants were able to interpret the instrument. For the focus group analysis, greater attention was given to the participants' grasp of the concepts and explanations being provided by the instrument. For both analyses, the researchers collectively identified patterns in the participants' responses and negotiated the manner in which they were addressed as a group. These changes are discussed further in Section 4. Changes occurred after each round of analysis and the revised assessment was used in the next round of data collection.

3.2.2. Quantitative Analyses of Students' Responses to the Instrument

For both rounds of quantitative data analysis presented in this paper, the analyses were conducted using a classical test theory (CTT) approach. The purpose of the first two round of testing for this instrument was to provide preliminary validity evidence before the team conducted large-scale data collection. CTT analyses are more appropriate for smaller sample sizes and provide baseline evidence for the instrument's validity so that the team could begin large-scale data collection for future item response theory (IRT) models with greater confidence in the instrument. The first round of quantitative data was analyzed to assess how well the questions and responses were interpreted by students. For this round of analysis, the research team primarily focused on the distractor analysis and the percentage of students selecting the preferred response. The items that resulted in participants responding with distractor choices for over 50% of the sample were reviewed for clarity. These metrics were determined using SPSS 27. After completion of this analysis, three items were adapted in order to improve performance, where text was altered to clarify distinctions between popular distractor responses and preferred responses. Further, this analysis explored how students in the different courses performed on the assessment to understand the instrument's ability to distinguish between theoretically distinct groups.

The analysis of the second round of quantitative data involved several procedures aimed at assessing the reliability of the instrument as a measure of students' scientific reasoning in biology. The second round of quantitative data analyses focused on the several psychometric properties of the items in the assessment, including item difficulty and discrimination, distractor analysis, internal consistency analysis and exploratory factor

analysis. For distractor analysis, the frequencies of the responses to all four options of each item were calculated using SPSS 27. Any items with distractors which had a higher percentage of students selecting a distractor over the correct answer were flagged for further review. In addition to distractor analysis, we also calculated item difficulty (percentage of students obtaining the item correct or *p*-value) and item discrimination (a point-biserial correlation between the dichotomous variable for obtaining the item correct and the student's summed score on the rest of the items). The results of the item difficulty and discrimination analyses are presented in Table 3. To evaluate the internal consistency of the instrument, we calculated Cronbach's alpha to measure internal consistency using SPSS 27 for the overall instrument and for each of the tiers in the assessment. Finally, to test for the dimensionality of the instrument, we conducted an exploratory factor analysis (EFA). Dichotomously coded variables were used, with 0 indicating that a student obtained the item incorrect and 1 representing that the student obtained the item correct. The EFA was conducted in Mplus 8.4 [71], using the weighted least squares mean and variance adjusted (WLSMV) estimator.

Table 3. Item Difficulty and Discrimination.

Item Number	Difficulty (*p*-Value)	Discrimination (Point-Biserial Correlation)
Tier 1		
Q1.1	0.533	0.603
Q2.1	0.598	0.574
Q3.1	0.411	0.599
Q4.1	0.673	0.287
Q5.1	0.411	0.457
Q6.1	0.411	0.428
Q7.1	0.645	0.596
Q8.1	0.626	0.607
Q9.1	0.617	0.515
Q10.1	0.514	0.511
Tier 2		
Q1.2	0.561	0.383
Q2.2	0.579	0.429
Q3.2	0.495	0.292
Q4.2	0.262	0.260
Q5.2	0.514	0.368
Q6.2	0.355	0.233
Q7.2	0.439	0.405
Q8.2	0.673	0.339
Q9.2	0.383	0.282
Q10.2	0.533	0.328
Tier 3		
Q1.3	0.542	0.448
Q2.3	0.607	0.463
Q3.3	0.336	0.406
Q4.3	0.234	0.193
Q5.3	0.477	0.570
Q6.3	0.430	0.432
Q7.3	0.439	0.349
Q8.3	0.589	0.511
Q9.3	0.533	0.469
Q10.3	0.430	0.197

4. Results

4.1. Evidence for Construct Validity—Initial Item Development and Review

As stated previously, the ARB instrument arose from the adaptation of a previously developed and validated instrument aimed at measuring students' ability to construct scientific explanations using core biology ideas [9]. That instrument consisted of two tiers of open-ended, constructed response questions aligned with several theoretical frameworks describing fundamental biological knowledge [10]. This assessment was reviewed by several biologists and biology educators and found to have translational validity, in that all the experts agreed that the instrument measured important concepts and explanations in biology, thus also supporting the construct validity of the ARB. Experts developed ideal answers for the constructed response version that were used to develop the scoring rubrics for the open-ended version of the first- and second-tier questions. For the current ARB instrument, the expert-generated rubrics served as the guide for developing the correct multiple-choice responses for all of the first- and second-tier questions in the ARB. Further, the authentic student responses from data collected in previous studies were reviewed by the research team to develop the distractor responses for the first and second tiers. To establish construct validity for the third-tier questions and responses, a new panel of experts, all who had a minimum of two post-secondary degrees in biology and advanced study in education, reviewed the third-tier questions and agreed they assessed biological reasoning. The third-tier responses also aligned with theoretical descriptions of how core science ideas are used to develop scientific explanations through reasoning [8,72]. Taken together, these efforts support the construct validity for the ARB instrument.

4.2. Evidence for Validity—Outcomes from Qualitative Interview Stages

The analysis of the two rounds of interview data led to several changes in the original iteration of the instrument. One major revision resulting from the initial round of think-aloud individual interviews entailed creating a relatively standardized structure for each tier in the question set for each topic area. The original question stems for the first and second tiers mirrored the question stems from the original constructed response instrument and the third-tier stem followed a general structure of "Which of the following **best describes** your **reasoning** for the choice you made in the previous question (#2)? (2nd tier question)". However, participants experienced difficulty in distinguishing the intent of the third-tier reasoning question from the second-tier question asking them to develop an explanation of the presented scenario using the focal concept from the first tier. Confusion between developing an explanation or the role of evidence in argumentation with the underlying reasoning has been noted in other studies, thus the students' struggle was not surprising [8,73]. To address this issue, all second-tier questions, which originally varied greatly in structure, were aligned more closely to a general form of "**Use** your knowledge of *X* (Focal concept in 1st tier) to select the statement that **best explains** *Y* (Focal scenario for each topic)."

This revised standardized structure was used during the focus group interviews and this set of students described the structure to be clear and logically presented. For instance, they discussed how the first-tier questions required that they pull from their prior knowledge about the concept, the second-tier ones required that they apply that knowledge to a scenario that they considered to have real world applications, which were sometimes novel to them, and the third-tier ones required that they describe their reasoning for that choice. In addition to this group understanding this structure and feeling comfortable in answering the question based on this structure, they also identified that the consistency of this structure helped them understand the nature of the assessment and the connection between the tiers as they progressed through the questions.

Further issues emerging from the analysis of several rounds of interview data broadly related to the semantic structures of the items and potential responses. Several of these issues surfaced as participants considered several of the distractor answer choices. Both individually and in the focus groups, some distractor answer choices seemed too attractive

when compared to desired answer choices. As the research team reviewed these items, the appeal of these distractors followed one of two trends. The first trend involved the distractor response using more generalized language while mainly differentiating through one or two critical terms from the desired response, which typically used slightly more technical wording. The slight variation in ease of comprehension led to the selection of the distractor over the desired response. To address this trend, the responses were edited to limit the level of technicality of each response and to expand the critical elements of the distractor to be more apparent. The second trend in participants' preference for certain distractor responses related to variation in the volume and length of text in the possible responses in the questions. If a particular response was longer and greater in word volume, participants typically deliberated more about their appropriateness and selected those distractors, even if the desired response had less length and volume. To address this trend, the length and volume of all responses for each question set were revised so that they were relatively equal to each other.

One last structural issue that arose for particular question sets involved the nature of the graphics used to accompany the focal scenario for each question set. Specifically, the graphics used in the questions about cell theory, mitosis, photosynthesis and cellular respiration went through several revisions to enhance the clarity of the image and provide a more nuanced representation of the scenario. The photosynthesis and cellular respiration question sets rely on the same experimental scenario using indicators to note the production and use of carbon dioxide in test tubes with plants and animals. The original image used involved black and white graphics only at the beginning of the questions. However, after some revisions, participants engaged in more thoughtful reasoning when color was added to the graphics and the answer choices were aligned to repeated elements from the overarching graphic. As these two questions rely on the evaluation of experimental data, rather than already analyzed forms of data, these revisions appeared to be particularly helpful in supporting participants' engagement with those questions.

4.3. Initial Evidence for Reliability—Outcomes of Quantitative Data Collection and Analyses

For the first round of quantitative data collection, the research team analyzed the results to determine how well the revisions to the textual structure and complexity of the responses supported participants selecting the desired response compared to the distractors. From this analysis, two issues arose that required attention to certain questions and responses. The first issue involved trends in responses to several first-tier questions, which asks respondents to select an answer that best described or defined the focal science concept for the question set. The analysis showed that, for four of these first-tier questions, participants selected one or two distractor responses at levels that were 10–25% greater than the desired response level. Upon review of these first-tier questions, all four followed a similar structure of asking a "negative" question, such as "Select the answer that does NOT represent the products of meiosis." Based on this pattern in the larger data set, the research team chose to revise those first-tier questions to a more affirmative format, such as "Select the answer that best represents the products of meiosis." The second issue concerned further challenges involving high similarity between the desired response and a particular distractor for three questions, which were revised further to distinguish between the two selections.

The second round of quantitative data collection provided more participant responses than the first round of data collection, while also allowing all course groups to complete their course of study in biology. The analyses for this data set aimed to explore several psychometric properties of the instrument to provide preliminary evidence for reliability and validity of the instrument. The first analytical step involved further distractor analysis for each item. The results from this analysis demonstrated that only two items had response rates which were significantly higher for a particular distractor (>10%) than the desired response. These particular items included the second- and third-tier questions for the question set involving cellular respiration. For both questions, the more popular distractor

response involved a critical error that misrepresented the role of oxygen in the process of cellular respiration, where O_2 was treated as a reactant rather than a product of the process. Understanding this specific role of oxygen is a key element of a sophisticated understanding of cellular respiration and more advanced reasoning through the experimental scenario presented in the question. Thus, the research team chose to retain these items in their forms as the distractor can help discern learners with more advanced biological reasoning. Only two other distractor responses garnered a slightly higher response rate than their corollary desired response item (<10%), but the review of those items did not demonstrate a compelling need for revision. All other distractors did not reach a response level higher than the desired correct response for the other questions. See Table 3 for a summary of item difficulty and discrimination.

The next psychometric analysis involved assessing the internal consistency of the instrument as a whole and of the three different tiers of question types by calculating a Cronbach's alpha for each subset of the data (see Table 4). For all items together, Cronbach's alpha was 0.905. When looking at the individual tiers within the assessment, the first- and third-tier subsets met the commonly adopted threshold of 0.7 [74]. The second tier had an alpha value slightly below 0.7. Follow-up analyses of item statistics for the second tier showed the deletion of any one item would not have increased the overall internal consistency for this tier, indicating that no item was problematic enough that deleting it from the instrument increased the overall reliability. The reduced internal consistency for the second-tier questions was not unexpected, as these questions are the most unique individually due to the different scenarios presented for each biological topic. Thus, the nature of the appropriate explanations for each scenario involved different reasoning processes, including experimental evaluation, application of analogical models and comparison of classification structures [12].

Table 4. Internal consistency for instrument and question tiers (Cronbach's alpha).

Overall Instrument	1st Tier	2nd Tier	3rd Tier
0.905	0.830	0.672	0.744

To test for initial dimensionality of the instrument, we conducted an EFA in Mplus version 8.4 using the WLSMV estimator. For this, dichotomously coded variables were used, with 0 indicating that a student obtained the item incorrect and 1 representing that the student obtained the item correct. The resulting scree plot is presented in Figure 2. To interpret the scree plot, we first identified the elbow point in the plot, indicating the number of factors at which point factors stop explaining significant portions of the variation and only considered factors to the left of that point significant. Our plot has an elbow point at 2 factors, indicating that only a one-factor model should be considered, based on these data. The plot provides preliminary evidence for a one-factor structure of the item response data. With an elbow point at factor 2, the plot indicates that only the first factor explains a significant amount of variance. For this analysis, two- and three-factor structures were considered. A three-factor structure would be plausible considering the conceptual, procedural and epistemic characters of the different question tiers. A two-factor structure would be plausible in light of the intertwined nature of the procedural and epistemic tiers with respect to the responses. The result of the one-factor structure is intriguing in light of scientific reasoning, as it lends support to the notion that all three elements of reasoning are necessary and possibly inseparable for an instrument in this format. However, we consider these factor analysis results to be preliminary due to the small sample size available. In the future, we plan to distribute the model to a large sample of students and we will conduct a more thorough examination of dimensionality through both exploratory and confirmatory factor analyses as preliminary stages to our planned IRT models.

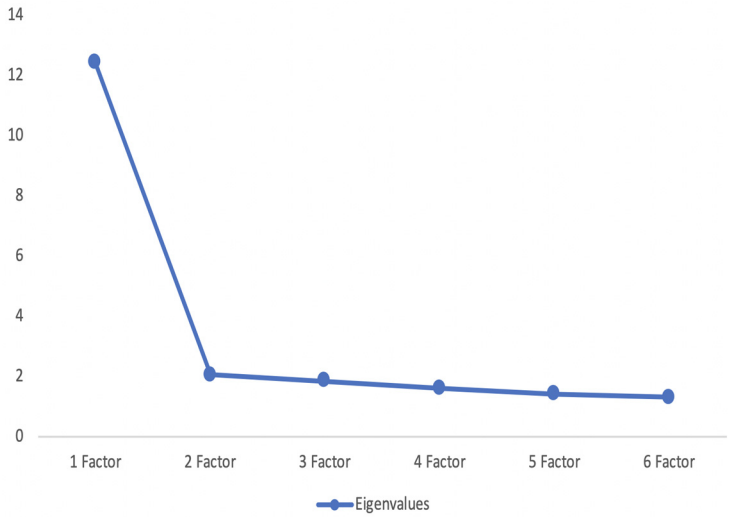

Figure 2. Exploratory factor analysis plot.

4.4. Evidence for Concurrent Validity

Another outcome from the analyses of the larger sets of data is the development of preliminary evidence for the instrument's ability to distinguish between groups of learners who are theoretically distinct. The instrument was administered at the beginning and end of a spring semester course sequence for all three groups. However, the group of participants in the International Baccalaureate Biology course completed a previous semester of biology instruction in the fall. Due to the school schedule structure, a semester-long course in this school equaled what is typically considered a year of typical instruction in most schools. Students in the Advanced Placement Biology and post-secondary General Biology Laboratory course were just beginning their continued study of biology, thus having to rely more on remembered prior knowledge to complete the instrument. That said, the post-secondary students, accepted for study at a research-level university, would be reasonably expected to have at least a slightly more developed conceptual capability in the sciences than Advanced Placement Biology students, who mostly had received introductory-level instruction in life science a few years prior. Thus, it is reasonable to expect the International Baccalaureate Biology students to score better than the other course groups, as they experienced the most recent direct instruction in biology. Further, due to their advanced experience with schooling, it was expected that the post-secondary students would score better than the Advanced Placement Biology students. Disaggregating the data by the different course groups confirmed these expectations, as seen in Table 5, offering evidence to support the concurrent validity of the instrument's ability to distinguish between theoretically different groups.

Table 5. Average percent correct responses across all items by course group, round 1 data.

	Advanced Placement Biology	International Baccalaureate Biology	General Biology Laboratory
Average mean correct	0.35	0.69	0.42
Standard deviation of mean correct	0.10	0.18	0.18

To test for significant differences in the average scores across these groups in the first round of data collection, we ran a Kruskal–Wallis one-way ANOVA in SPSS version 27. The non-parametric Kruskal–Wallis was selected because the data in our sample were not normally distributed, which would have resulted in a violation of assumptions in a traditional one-way ANOVA. The test indicated that, overall, there were significant differences between the groups ($H(2) = 23.130$, $p < 0.001$). Post hoc tests revealed signifi-

cant differences between Advanced Placement Biology and International Baccalaureate biology ($p < 0.001$) and between International Baccalaureate Biology and General Biology Laboratory ($p = 0.001$), but no significant difference between General Biology Laboratory and Advanced Placement Biology (Table 6).

Table 6. Average percent correct responses across all items by course group, round 2 data.

	Advanced Placement Biology	International Baccalaureate Biology	General Biology Laboratory
Average mean correct	0.43	0.70	0.50
Standard deviation of mean correct	0.10	0.15	0.24

As in round one, we examined the average scores across the different course types to establish concurrent validity for the ARB using the second data set. To test for significant differences in the average scores across these groups, we ran a Mann–Whitney U test in SPSS version 27. The non-parametric Mann–Whitney test was selected for round two data because the sample size for General Biology Laboratory was not large enough to test for statistical significance and the data was not normally distributed, consistently with round one data. The Mann–Whitney test indicated a significant difference between the scores of Advanced Placement and International Baccalaureate Biology ($U = 314.5$, $p < 0.001$).

5. Discussion

Using the collection of evidence described above, we assert that the preliminary evidence supports the Assessment of Biological Reasoning as a valid assessment instrument for measuring high school students' reasoning capabilities across several major biological topic areas. The resulting ABR assessment consists of 30 questions divided into 10 question sets connected to 10 biological topic areas, with each set including three tiered questions with four answer choices each (see Supplementary Materials for the full instrument). The three-tiered nature of the question sets align with the three recognized dimensions of scientific reasoning [3,12], including a conceptually oriented question comprising the primary object of reasoning, a procedural oriented question that engages the student in developing scientific explanations for the scenarios grounding the question and an epistemically oriented question exploring how a respondent uses the focal science concept to construct their preferred explanatory response. Using a validation framework stemming from the work by Trochim [70] and used in previous validation work by the authors [9], we collected an assemblage of evidence that demonstrates the construct validity, criterion validity and reliability of the ABR instrument.

Through the development of the ABR, the research team gained some insight into the nature of students' reasoning in biology. When developing the instrument, we were not sure if the multiple tiers of questions within a set would be reliant or independent of each other, as each set had a specific focus on a specific ontological/conceptual component but each tier of questions focused on a different component of reasoning. This question regarding the interactive nature of the components stems from descriptions that primarily place domain specificity within the ontological/conceptual component, while the procedural and epistemic components of scientific reasoning have more domain general characteristics [12]. Based on the EFA analysis conducted with the largest sample of responses, the one-factor structure confirmed for the ABR provides preliminary evidence that domain-specific/general distinctions among the three components are not borne out. Rather, although procedural and epistemic dimensions of reasoning may broadly be applied across disciplines, as all science disciplines involve experimental design, modeling and classification, our results suggest that those reasoning components are given meaning by their ontological element. That is, investigating students' ability with certain scientific reasoning activities must pay attention to the ontological/conceptual components of the

activity. This conclusion resonates with other studies that demonstrate that conceptual awareness can improve the overall quality of the verbal argumentation that students engage in, but it is important to indicate that students' epistemic practices can improve separately from conceptual awareness [17]. Additionally, it is important to note that the ABR is mute regarding this point, as the design of the ABR negates this possibly, even if it is sound, given the design of this standardized measure.

The analyses of students' thinking and reasoning during the qualitative data collection also support the intertwined nature of the three tiers of questions within each set. Considering the outcomes described above, an interesting pattern emerged when we examined the questions for which students' expressed difficulties—particularly interpretive difficulties as opposed to simple unfamiliarity with the concept. In the instances, when students encountered interpretive difficulties with a particular question set, we came to understand the students' self-generated descriptions of the focal concepts became a standard by which the students' judged the phrasing of the other response items. It seems that students assessed the language in the responses for the second and third tier through their personal understanding of the focal concepts. This pattern offers an explanation for why the negatively phrased first-tier questions in a previous iteration of the ABR did not produce high correct response rates. This relationship can also help understand how the role of graphics changed and enhanced students' ability to reason through the scenarios, as they provided a conceptual anchor for those questions that could have assisted students in navigating the second- and third-tier questions. The importance of conceptual clarity for respondents' reasoning resonates with findings of earlier studies that speak to the importance of the quality of the cognitive objects involved in students' reasoning [15].

6. Limitations and Implications

The research team recognizes that the ABR instrument and the current validation efforts do have some limitations that should be acknowledged. First, the assessment, while focusing on key biology topics covered in high school and post-secondary education is limited in nature because of this focus. As our results suggest, the ontological/conceptual component of the assessment are interconnected such that the application and reasoning components cannot be disentangled. As such, the ABR instrument is limited in use to biology classes.

Second, the nature of the assessment, while allowing the quantitative assessment of scientific reasoning to be conducted in a controlled format that can be uniformly implemented and easily scored in a short amount of time for a large sample of students, has its limitations [75,76]. One such limitation is that the multiple-choice format is constrained and does not assess reasoning that may occur in what Chinn and Duncan [41] call "the wild". By this they mean that multiple-choice and, even, assessments with open-ended questions do not capture students' reasoning that is observable during performance tasks, inquiry activities, or through direct open-ended, person-centered questioning (questions related to students' ideas) that can be employed by teachers in situ [41,75]. Additionally, while multiple choice tests may have advantages over open-response questions, which often also assess a student's writing ability, they are open to issues of guessing and test taking strategies such as using clues provided by particular words or statements in a question [75].

Third, as a multiple-choice style assessment, there are valid critiques that the wording of response items requires students to comprehend and use language that may not be familiar or representative of their thinking [77]. However, we endeavored to make the language of the response items more accessible by generating many of them from previously recorded student responses and iteratively refining the instrument based on qualitative data from interviews. Similarly, the language used in the question sets is relatively complex and may present challenges for some students. The inclusion of the graphics for each question works to support the interpretability of the questions, but those may not be sufficient and further scaffolding to support students' interpretation of meaning may necessitate further investigation of the ABR. Although the sample sizes for this study were

not overly large, further research being conducted will provide a much larger data set that will help advance the validation of the ABR and the findings related to measuring students' biological reasoning.

Much of the groundbreaking work into students' reasoning in science has been necessarily content-embedded and heavily descriptive, often relying on participant observations and analysis of students' work products and discourse [14,15,20,21]. Given the intensive nature of such investigations, such work is simply not scalable, something that limits the advancement of this line of research. In response to this and to the need for psychometrically sound assessments [26], the ABR represents a contribution to research into secondary students' reasoning in biology, as it is domain- and grade level-specific for measuring students' reasoning in secondary level biology. Although some in-depth assessments of students' reasoning with certain biological topics already exist [22,23], extant assessments across the discipline of biology are primarily limited to measuring conceptual understanding [28]. Thus, the introduction of the ABR represents an advanced tool for the field to use to measure more complex learning and reasoning in secondary biology classrooms, something needed if the field is to move toward larger scale studies involving students' biological reasoning.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/educsci11110669/s1, Assessment of Biological Reasoning (ABR) Instrument.

Author Contributions: Conceptualization, J.S., P.J.E., S.S.-M., D.R. and S.A.S.; methodology, J.S., P.J.E., K.R. and S.A.S.; validation, J.S., P.J.E., K.R., S.S.-M., D.R. and S.A.S.; formal analysis, K.R.; investigation, J.S., P.J.E., S.S.-M., D.R. and S.A.S.; writing—original draft preparation, J.S., P.J.E., K.R., S.S.-M., D.R. and S.A.S.; writing—review and editing, J.S., P.J.E., K.R. and S.A.S.; visualization, J.S.; supervision, J.S., P.J.E. and S.A.S.; project administration, J.S. and S.A.S.; funding acquisition, P.J.E. and S.A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This material is based upon work supported by the National Science Foundation under DRL #1720587 and DMR #1644779. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Florida State University (STUDY00001609 approved August 17, 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy issues.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Crawford, B. From inquiry to scientific practices in the science classroom. In *Handbook of Research on Science Education, Volume II*; Routledge: New York, NY, USA, 2014.
2. Duschl, R. Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Rev. Res. Educ.* **2008**, *32*, 268–291. [CrossRef]
3. Osborne, J. The 21st century challenge for science education: Assessing scientific reasoning. *Think. Ski. Creat.* **2013**, *10*, 265–279. [CrossRef]
4. Walker, J.P.; Sampson, V.; Southerland, S.; Enderle, P.J. Using the laboratory to engage all students in science practices. *Chem. Educ. Res. Pract.* **2016**, *17*, 1098–1113. [CrossRef]
5. Nehring, A.; Nowak, K.H.; Upmeier zu Belzen, A.; Tiemann, R. Predicting students' skills in the context of scientific inquiry with cognitive, motivational, and sociodemographic variables. *Int. J. Sci. Educ.* **2015**, *37*, 1343–1363. [CrossRef]
6. Walker, J.; Sampson, V. Learning to argue and arguing to learn: Argument-Driven Inquiry as a way to help undergraduate chemistry students learn how to construct arguments and engage in argumentation during a laboratory course. *J. Res. Sci. Teach.* **2013**, *50*, 561–596. [CrossRef]
7. Ford, M. Educational implications of choosing "practice" to describe science in the Next Generation Science Standards. *Sci. Educ.* **2015**, *99*, 1041–1048. [CrossRef]

8. Sampson, V.; Enderle, P.; Grooms, J. Argumentation in science education. *Sci. Teach.* **2013**, *80*, 30. [CrossRef]
9. Grooms, J.; Enderle, P.; Sampson, V. Coordinating scientific argumentation and the Next Generation Science Standards through argument driven inquiry. *Sci. Educ.* **2015**, *24*, 45–50.
10. National Research Council. *A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*; National Academies Press: Washington, DC, USA, 2012.
11. Zimmerman, C. The development of scientific thinking skills in elementary and middle school. *Dev. Rev.* **2007**, *27*, 172–223. [CrossRef]
12. Kind, P.; Osborne, J. Styles of scientific reasoning: A cultural rationale for science education? *Sci. Educ.* **2017**, *101*, 8–31. [CrossRef]
13. Stroupe, D.; Moon, J.; Michaels, S. Introduction to special issue: Epistemic tools in Science Education. *Sci. Educ.* **2019**, *103*, 948–951. [CrossRef]
14. Gonzalez-Howards, M.; McNeill, K. Acting with epistemic agency: Characterizing student critique during argumentation discussions. *Sci. Educ.* **2020**, *104*, 953–982. [CrossRef]
15. Zangori, L.; Vo, T.; Forbes, C.; Schwarz, C.V. Supporting 3rd-grad students' model-based explanations about groundwater: A quasi-experimental study of a curricular intervention. *Int. J. Sci. Educ.* **2017**, *39*, 1421–1442. [CrossRef]
16. Svoboda, J.; Passmore, C. The strategies of modeling in Biology education. *Sci. Educ.* **2013**, *22*, 119–142. [CrossRef]
17. Grooms, J.; Sampson, V.; Enderle, P. How concept familiarity and experience with scientific argumentation are related to the way groups participate in an episode of argumentation. *J. Res. Sci. Teach.* **2018**, *55*, 1264–1286. [CrossRef]
18. Koerber, S.; Osterhaus, C. Individual differences in early scientific thinking: Assessment, cognitive influences, and their relevance for science learning. *J. Cogn. Dev.* **2019**, *20*, 510–533. [CrossRef]
19. Reith, M.; Nehring, A. Scientific reasoning and views on the nature of scientific inquiry: Testing a new framework to understand and model epistemic cognition in science. *Int. J. Sci. Educ.* **2020**, *42*, 2716–2741. [CrossRef]
20. Salmon, S.; Levy, S. Interactions between reasoning about complex systems and conceptual understanding in learning chemistry. *J. Res. Sci. Teach.* **2019**, *57*, 58–86. [CrossRef]
21. Sampson, V.; Clark, D. The impact of collaboration on the outcomes of scientific argumentation. *Sci. Educ.* **2009**, *93*, 448–484. [CrossRef]
22. Haskel-Ittah, M.; Duncan, R.G.; Yarden, A. Students' understandings of the dynamic nature of genetics: Characterizing undergraduate' explanations for interactions between genetics and environment. *ICBE Live Sci. Educ.* **2020**, *19*, ar37. [CrossRef]
23. To, C.; Tenenbaum, H.; High, H. Secondary school students' reasoning about evolution. *J. Res. Sci. Teach.* **2016**, *54*, 247–273. [CrossRef]
24. Krell, M.; Mathesius, S.; van Driel, J.; Vergara, C.; Krüger, D. Assessing scientific reasoning competencies of pre-service science teachers: Translating a German multiple-choice instrument into English and Spanish. *Int. J. Sci. Educ.* **2020**, *42*, 2819–2841. [CrossRef]
25. Krell, M.; Redman, C.; Mathesius, S.; Krüger, D.; van Driel, J. Assessing pre-service science teachers' scientific reasoning competencies. *Res. Sci. Educ.* **2020**, *50*, 2305–2329. [CrossRef]
26. Opitz, A.; Heene, M.; Fischer, F. Measuring scientific reasoning—A review of test instruments. *Educ. Res. Eval.* **2017**, *23*, 78–101. [CrossRef]
27. Barbera, J. A psychometric analysis of the chemical concepts inventory. *J. Chem. Educ.* **2013**, *90*, 546–553. [CrossRef]
28. Garvin-Doxas, K.; Klymkowsky, M.W. Understanding randomness and its impact on student learning: Lessons learned from building the Biology Concept Inventory (BCI). *CBE Life Sci. Educ.* **2008**, *7*, 227–233. [CrossRef]
29. Hestenes, D.; Wells, M.; Swackhamer, G. Force concept inventory. *Phys. Teach.* **1992**, *30*, 141–158. [CrossRef]
30. Pollock, S.J. Comparing student learning with multiple research-based conceptual surveys: CSEM and BEMA. In *AIP Conference Proceedings*; American Institute of Physics: College Park, MD, USA, 2008; Volume 1064, pp. 171–174.
31. Southerland, S.A.; Granger, E.; Jaber, L.; Tekkumru-Kisa, M.; Kisa, Z. Learning through Collaborative Design (LCD): Professional Development to Foster Productive Epistemic Discourse in Science. National Science Foundation, DRL #1720587. 2017. Available online: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1720587. (accessed on 11 June 2021).
32. Inhelder, B.; Piaget, J. *The Growth of Logical Thinking: From Childhood to Adolescence*; Parsons, A.; Milgram, S., Translators; Basic Books: New York, NY, USA, 1958. [CrossRef]
33. Zimmerman, B.J. Attaining Self-Regulation: A Social Cognitive Perspective. In *Handbook of Self-Regulation*; Boekaerts, M., Pintrich, P.R., Zeidner, M., Eds.; Academic Press: San Diego, CA, USA, 2000; pp. 13–39.
34. Mayr, E. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*; Harvard University Press: Cambridge, MA, USA, 1982.
35. Kuhn, D. *Education for Thinking*; Harvard University Press: Cambridge, MA, USA, 2005.
36. Kuhn, D.; Dean, D. A bridge between cognitive psychology and educational practice. *Theory Pract.* **2004**, *43*, 268–273. [CrossRef]
37. Osborne, J. Teaching scientific practices: Meeting the challenge of change. *J. Sci. Teach. Educ.* **2014**, *25*, 177–196. [CrossRef]
38. Sandoval, W.A. Conceptual and epistemic aspects of students' scientific explanations. *J. Learn. Sci.* **2003**, *12*, 5–51. [CrossRef]
39. Sandoval, W.; Reiser, B. Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Sci. Educ.* **2004**, *88*, 345–372. [CrossRef]

40. Shavelson, R.J. Discussion of papers and reflections on "exploring the limits of domain-generality". In *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*; Fischer, F., Chinn, C.A., Engelmann, K., Osborne, J., Eds.; Routledge: New York, NY, USA, 2018; pp. 112–118.
41. Chinn, C.A.; Duncan, R.G. What is the value of general knowledge of scientific reasoning? In *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*; Fischer, F., Chinn, C.A., Engelmann, K., Osborne, J., Eds.; Routledge: New York, NY, USA, 2018; pp. 77–101.
42. Samarapungavan, A. Construing scientific evidence: The role of disciplinary knowledge in reasoning with and about evidence in scientific practice. In *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*; Fischer, F., Chinn, C.A., Engelmann, K., Osborne, J., Eds.; Routledge: New York, NY, USA, 2018; pp. 56–76.
43. Banilower, E.R.; Smith, P.S.; Malzahn, K.A.; Plumley, C.L.; Gordon, E.M.; Hayes, M.L. *Report of the 2018 NSSME+*; Horizon Research, Inc.: Chapel Hill, NC, USA, 2018.
44. Jackson, S.L.; Stratford, S.J.; Krajcik, J.; Soloway, E. Making dynamic modeling accessible to precollege science students. *Interact. Learn. Environ.* **1994**, *4*, 233–257. [CrossRef]
45. Penner, D.E. Complexity, emergence, and synthetic models in science education. In *Designing for Science*; Psychology Press: Mahwah, NJ, USA, 2001; pp. 177–208.
46. Sins, P.H.; Savelsbergh, E.R.; van Joolingen, W.R. The Difficult Process of Scientific Modelling: An analysis of novices' reasoning during computer-based modelling. *Int. J. Sci. Educ.* **2005**, *27*, 1695–1721. [CrossRef]
47. Schwarz, C.V.; Reiser, B.J.; Davis, E.A.; Kenyon, L.; Achér, A.; Fortus, D.; Shwartz, Y.; Hug, J.; Krajcik, J. Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *J. Res. Sci. Teach. Off. J. Natl. Assoc. Res. Sci. Teach.* **2009**, *46*, 632–654. [CrossRef]
48. Berland, L.; Reiser, B. Making sense of argumentation and explanation? *Sci. Educ.* **2009**, *93*, 26–55. [CrossRef]
49. Erduran, S.; Simon, S.; Osborne, J. TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Sci. Educ.* **2004**, *88*, 915–933. [CrossRef]
50. Clark, D.B.; Sampson, V. Personally-seeded discussions to scaffold online argumentation. *Int. J. Sci. Educ.* **2007**, *29*, 253–277. [CrossRef]
51. Jimenez-Aleixandre, M.P.; Bugallo Rodriguez, A.; Duschl, R.A. "Doing the lesson" or "doing science": Argument in high school genetics. *Sci. Educ.* **2000**, *84*, 287–312. [CrossRef]
52. Sandoval, W.A.; Millwood, K.A. The quality of students' use of evidence in written scientific explanations. *Cogn. Instr.* **2005**, *23*, 23–55. [CrossRef]
53. Osborne, J.; Erduran, S.; Simon, S. Enhancing the quality of argumentation in school science. *J. Res. Sci. Teach.* **2004**, *41*, 994–1020. [CrossRef]
54. Ryu, S.; Sandoval, W. Improvements to elementary children's epistemic understanding from ssutatin argumentation. *Sci. Educ.* **2012**, *96*, 488–526. [CrossRef]
55. Zohar, A.; Nemet, F. Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *J. Res. Sci. Teach.* **2002**, *39*, 35–62. [CrossRef]
56. Adadan, E.; Savasci, F. An analysis of 16–17-year-old students' understanding of solution chemistry concepts using a two-tier diagnostic instrument. *Int. J. Sci. Educ.* **2012**, *34*, 513–544. [CrossRef]
57. Chen, C.C.; Lin, H.S.; Lin, M.L. Developing a two-tier diagnostic instrument to assess high school students' understanding-the formation of images by a plane mirror. *Proc. Natl. Sci. Counc. Repub. China Part D Math. Sci. Technol. Educ.* **2002**, *12*, 106–121.
58. Griffard, P.B.; Wandersee, J.H. The two-tier instrument on photosynthesis: What does it diagnose? *Int. J. Sci. Educ.* **2001**, *23*, 1039–1052. [CrossRef]
59. Treagust, D.F. Development and use of diagnostic tests to evaluate students' misconceptions in science. *Int. J. Sci. Educ.* **1988**, *10*, 159–169. [CrossRef]
60. Strimaitis, A.M.; Schellinger, J.; Jones, A.; Grooms, J.; Sampson, V. Development of an instrument to assess student knowledge necessary to critically evaluate scientific claims in the popular media. *J. Coll. Sci. Teach.* **2014**, *43*, 55–68. [CrossRef]
61. Tan KC, D.; Taber, K.S.; Goh, N.K.; Chia, L.S. The ionisation energy diagnostic instrument: A two-tier multiple-choice instrument to determine high school students' understanding of ionisation energy. *Chem. Educ. Res. Pract.* **2005**, *6*, 180–197. [CrossRef]
62. Chang, H.P.; Chen, J.Y.; Guo, C.J.; Chen, C.C.; Chang, C.Y.; Lin, S.H.; Su, W.J.; Lain, K.D.; Hsu, S.Y.; Lin, J.L.; et al. Investigating primary and secondary students' learning of physics concepts in Taiwan. *Int. J. Sci. Educ.* **2007**, *29*, 465–482. [CrossRef]
63. Caleon, I.; Subramaniam, R. Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *Int. J. Sci. Educ.* **2010**, *32*, 939–961. [CrossRef]
64. Caleon, I.S.; Subramaniam, R. Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions. *Res. Sci. Educ.* **2010**, *40*, 313–337. [CrossRef]
65. Peşman, H.; Eryılmaz, A. Development of a three-tier test to assess misconceptions about simple electric circuits. *J. Educ. Res.* **2010**, *103*, 208–222. [CrossRef]
66. Hasan, S.; Bagayoko, D.; Kelley, E.L. Misconceptions and the certainty of response index (CRI). *Phys. Educ.* **1999**, *34*, 294–299. [CrossRef]
67. Renner, C.H.; Renner, M.J. But I thought I knew that: Using confidence estimation as a debiasing technique to improve classroom performance. *Appl. Cogn. Psychol. Off. J. Soc. Appl. Res. Mem. Cogn.* **2001**, *15*, 23–32. [CrossRef]

68. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing*; American Educational Research Association: Washington, DC, USA, 2014.
69. Kline, P. *The Handbook of Psychological Testing*, 2nd ed.; Routledge: New York, NY, USA, 2000.
70. Trochim, W.M. *The Research Methods Knowledge Base*, 2nd ed.; Atomic Dog: Cincinnati, OH, USA, 1999.
71. Muthén, L.K.; Muthén, B.O. *Mplus: Statistical Analysis with Latent Variables: User's Guide*; Version 8; Muthén & Muthén: Los Angeles, CA, USA, 2017.
72. McNeill, K.L.; Krajcik, J. Inquiry and scientific explanations: Helping students use evidence and reasoning. *Sci. Inq. Second. Setting* **2008**, *121*, 34.
73. McNeill, K.; Knight, A. Teachers' pedagogical content knowledge of scientific argumentation: The impact of professional development on K-12 Teachers. *Sci. Educ.* **2013**, *97*, 936–972. [CrossRef]
74. Taber, K.S. The use of Cronbach's alpha when developing and reporting research instruments in science education. *Res. Sci. Educ.* **2018**, *48*, 1273–1296. [CrossRef]
75. Harlen, W. *Assessment & Inquiry-Based Science Education: Issues in Policy and Practice*; Global Network of Science Academies: Trieste, Italy, 2013.
76. Simkin, M.G.; Kuechler, W.L. Multiple-choice tests and student understanding: What is the connection? *Decis. Sci. J. Innov. Educ.* **2005**, *3*, 73–98. [CrossRef]
77. Lee, O.; Quinn, H.; Valdes, G. Science and language for English language learners in relation to Next Generation Science Standards and with Implications for Common Core State Standards for English Language Arts and Mathematics. *Educ. Res.* **2013**, *42*, 223–233. [CrossRef]