education sciences

MDPI

Article

Modeling and Measuring Pre-Service Teachers' Assessment Literacy Regarding Experimentation Competences in Biology

Cora Joachim [1,*], Marcus Hammann [2], Claus H. Carstensen [3] and Susanne Bögeholz [1,*]

1 Biology Education, University of Göttingen, 37073 Göttingen, Germany
2 Center for Biology Education, University of Münster, 48143 Münster, Germany; hammann.m@uni-muenster.de
3 Psychological Methods of Educational Research, University of Bamberg, 96047 Bamberg, Germany; claus.carstensen@uni-bamberg.de
* Correspondence: cora.joachim@biologie.uni-goettingen.de (C.J.); sboegeh@gwdg.de (S.B.)

Received: 27 April 2020; Accepted: 11 May 2020; Published: 14 May 2020

Abstract: Assessment literacy is a crucial aspect of teachers' professional knowledge and relevant to fostering students' learning. Concerning experimentation, teachers have to be able to assess student achievement when students form hypotheses, design experiments, and analyze data. Therefore, teachers need to be familiar with criteria for experimentation as well as student conceptions of experimentation. The present study modeled and measured 495 German pre-service teachers' *knowledge of what to assess regarding experimentation competences in biology*. We applied an open-answer format for the measurement instrument. For modeling we used item response theory (IRT). We argue that *knowledge of what to assess* regarding experimentation competences is a one-dimensional construct and we provide evidence for the validity of the measurement. Furthermore, we describe qualitative findings of pre-service teachers' *knowledge of what to assess*, in particular difficulties concerning the assessment of student conceptions as well as the use of scientific terms in the assessments. We discuss the findings in terms of implications for science teacher education and further research perspectives.

Keywords: assessment literacy; teacher education; experimentation; competence; biology

1. Introduction

When student experimentation competences are fostered, teachers' assessment skills come into focus. *Assessment literacy* of teachers has been found to have a significant effect on students' learning [1]. Educational assessment is closely connected to instruction and takes place regularly [2] (p. 1). It is a prerequisite to planning lessons and adapting instruction to the students' needs. Moreover, "assessment information provides feedback to the student", which can enhance achievement [2] (p. 7) [3].

A central learning objective in biology are experimentation competences [4,5]. Experimentation competences are acquired successively in high school. One challenge for students is to understand how new findings in biology are gained. Specifically, German students often have misconceptions regarding experimentation [6] (p. 199). Therefore, the formation of hypotheses, the design of experiments, and the analysis of data must be practiced, and mistakes should be discussed [7]. Teachers have to be able to assess students' experimentation competences and related conceptions adequately to adapt their instruction, thus enhancing students' understanding [2] (p. 10). Assessing experimentation competences requires a range of knowledge. Disciplinary knowledge (cf. content knowledge, CK), and pedagogical content knowledge (PCK) are essential for the assessments [8] (p. 156). Teacher education in universities has to establish the essential knowledge and skills regarding assessment literacy. "knowledge is converted to skills" with increasing competence [9] (p. 70) [10]. Skills do not include

"mastering a technique [...] [without] the use of systematic knowledge" in this contribution [11] (p. 374). Assessment literacy can be expanded in the proceeding teaching practice and the teaching career [12]. Assessment literacy is "defined as a basic understanding of educational assessment and related skills to apply such knowledge to various measures of student achievement" [8] (p. 149) cf. [13].

To date, only a few research studies have been conducted on teachers' assessment literacy for inquiry concerning biology. One of the studies has analyzed pre-service teachers' *diagnostic competence* for experimentation competences via a questionnaire with a closed answer format, for which the Cronbach's alpha for the measure of *diagnostic competence* ($\alpha = 0.50$) was low [14] (pp. 67ff., p. 189). Alternatively, the present study specifically focuses on subject-specific research regarding teacher education in biology in Germany. The paper-pencil questionnaire study applies open-ended tasks presenting classroom scenarios of experimental biology lesson activities, which meet real-life demands in assessment more closely than a closed-answer format. It aims to develop a more reliable instrument to capture pre-service teachers' knowledge of assessment criteria for experimentation and their ability to apply these criteria. Our research goals are modeling and measuring pre-service biology teachers' knowledge in the area of assessment literacy. Thereby, we aim to gain qualitative insights into pre-service teachers' strengths and weaknesses in assessing experimentation competences.

1.1. Assessment Literacy as Part of Professional Knowledge

Knowledge of assessment is part of teachers' professional knowledge [15]. Assessment literacy encompasses four areas of knowledge: (1) knowledge of assessment purposes, (2) knowledge of what to assess, (3) knowledge of assessment strategies, and (4) knowledge of assessment interpretation and action-taking [16]. With regard to the first area of knowledge, teachers should be familiar with the aims of assessment such as "Providing data for instructors on which to base instructional decisions" [16] (p. 213). The second area of knowledge acknowledges that knowledge of what to assess is linked to "curricular goals and to values of what is important to learn and how learning occurs" [16] (p. 214), suggesting that to assess students adequately, teachers must be knowledgeable and proficient in curriculum topics and skills. In addition, knowledge of students' misconceptions is an essential component of this area of knowledge [16] (p. 216f.). The third area of knowledge, knowledge of assessment strategies, encompasses different ways that can be applied to assessment. Teachers should be familiar with strategies for formal and informal assessment. Abell and Siegel emphasize that knowledge of assessment strategies also includes "knowledge of topic-specific assessment tasks" and "knowledge of response strategies" [16] (p. 214). Finally, the knowledge of assessment interpretation and action-taking is hallmarked, for example, by being able to use assessment results to adapt instruction [16] (p. 215).

Of these four areas of knowledge, *knowledge of what to assess* is especially content specific and fundamental, highlighting the relevance in taking a closer look at pre-service teachers' *knowledge of what to assess* regarding experimentation competences in biology in the following.

1.2. Assessment of Students' Experimentation Competences

Teachers' *knowledge of what to assess* comprises knowledge of concepts and processes regarding experimentation that students need to acquire and an understanding of student conceptions and difficulties. We describe teachers' knowledge of concepts and processes as well as student conceptions and difficulties in the following section.

Students need to be able to apply scientific knowledge, such as knowledge of science and knowledge about science [17]. One learning objective is procedural knowledge to understand how scientific knowledge is generated. A central method in science to gain new findings is experimentation [18] (p. 15) [19] (p. 323).

Following the general model of *Scientific Discovery as Dual Search* (SDDS) [20], experimentation competences comprise the three phases: searching hypotheses, testing the hypotheses, and evaluating the evidence [21] (p. 8). Next, we summarize requirements regarding the core facets of the three phases.

Experiments serve to examine causal relationships. Hypotheses state assumed relationships between independent and dependent variables [22] (p. 45f.). Hypotheses should be "fully specified and testable" [21] (p. 8). Hypotheses can be theoretically founded based on previous knowledge [23] (p. 9f.). Research has shown that it can be difficult for younger students to think of different explanations for a phenomenon and generate alternative hypotheses. Often, the formation of hypotheses is incomprehensive [24] (p. 245) [7] (p. 298).

For testing hypotheses, it is essential to design structured experiments and vary the independent variables (the potential causes) systematically. All other variables must be kept constant to achieve unambiguous results [18] (p. 18) [19] (p. 323) [7] (p. 292f). Furthermore, it is important to observe and accurately measure the dependent variable (potential effect) [25] (p. 7) [26] (p. 43). Many students, however, have been shown to have misconceptions. For example, students may think that the goal of an experiment is to create an effect (engineering mode) instead of examining causal relationships [21] (p. 12) [27] (p. 860ff.). The experiment has to be precisely described so that it can be repeated [25] (p. 7). It requires consideration of appropriate methods and conducting experiments in a standardized way [25] (p. 7).

Finally, the data must be analyzed precisely. When data are assessed, errors have to be analyzed and taken into account [28] (p. 155). Furthermore, students should "differentiate experimental error [. . .] from experimental effect" [21] (p. 7). Results are compared with the hypothesis, which is accepted, rejected, or further examined [21] (p. 9) [19] (p. 324). It is essential to "guard against one's own confirmation bias in data interpretation" [21] (p. 7). Confirmation bias can influence the reasoning in that specific data that do not support the hypothesis are ignored. It can be difficult for students to reject a hypothesis due to their beliefs [21] (p. 9) [24] (p. 84f.).

Learning outcomes relevant to experimentation are prescribed in the German National Educational Standards [4]. According to the standards, students are expected to be able to plan, conduct, and analyze experiments at the end of grade 10 [4] (p. 14). Teachers have to know the learning goals and understand the student conceptions in order to conduct assessments that serve learning [2] (p. 2, 10). Therefore, this situation requires CK and PCK, i.e., knowledge of experimen-tation and student conceptions in biology.

Constructs that are related to the *knowledge of what to assess* regarding experimentation competences are *examining competence* and *diagnostic competence*. *Examining competence*, similar to experimentation competences, comprises the facets of questions, hypotheses, design and performance, and analysis and interpretation, which has been the focus of analysis for pre-service science teachers [29] (p. 40ff.). Assessments of experimentation competences that serve to learn have only been focused on by a few research studies so far. One example is a study conducted with biology pre-service teachers: Dübbelde [14] investigated the *diagnostic competence* for experimentation competences. The tasks performed in a closed answer format captured pre-service teachers' ability to apply given criteria, such as linking the conclusion to the hypothesis, but not the *knowledge of what to assess*.

Besides *knowledge of what to assess*, efficacy beliefs can influence the performance in our test. High self-efficacy beliefs can enhance the useful application of knowledge [30,31] (p. 211). Personal teaching efficacy of student interns correlated, e.g., with their lesson presenting behavior and questioning behavior [32] (p. 413).

Studies of *scientific reasoning*, *scientific inquiry*, and *experimentation competences* described conflicting findings regarding the dimensionality of the constructs [22,33,34]. Weak and intermediate latent correlations of 0.33–0.73 between the subscales related to question, hypothesis, planning, and interpretation (condensed label of subscales used by the authors) indicate that different skills are necessary for the different phases of *scientific reasoning* [33] (p. 58). Wellnitz's study of *scientific inquiry*, on the contrary, found higher latent correlations between the scales question, hypothesis, experimental design, and data analysis (0.80–0.95) [22] (p. 132) so that the authors of this study argue that comprehensive skills are necessary for all phases of experimentation.

Teachers need to possess experimentation competences to be able to evaluate student achievement. In particular, explicit knowledge of criteria and misconceptions enables teachers to assess student achievement against curricular expectations. When experimenting in class, teachers can focus on one of the three phases of experimentation. To convey an understanding of scientific inquiry, however, it is helpful for students to engage themselves in the whole process [25] (p. 5). Hence, teachers should have an understanding of all three phases of experimentation.

1.3. Research Questions and Hypotheses

The goal of the study is to model and measure pre-service biology teachers' knowledge and skills regarding the assessment of high school students' experimentation competences. Depending on theoretical background, two different models of *knowledge of what to assess* regarding experimentation competences in biology can be derived: a one-dimensional (1D) model comprising the three phases of experimentation, and a three-dimensional (3D) model taking into account the different requirements for forming hypotheses, planning experiments and analyzing data. By modeling and measuring it can be learned more about the dimensionality and quality of pre-service biology teachers' assessment literacy on *what to assess*. Therefore, a reliable and valid measurement instrument is necessary.

This type of query led to three research questions:

The first question concerns the construct dimensionality and test quality.

1. In what way can *knowledge of what to assess* regarding experimentation competences in biology be modeled and measured?

For investigating the validity of our conclusions, the following constructs related to *knowledge of what to assess* regarding experimentation competences are relevant: Given the knowledge and skills that are necessary to assess experimentation competences, *examining competence* and *diagnostic competence* for experimentation competences should be closely related to the construct measured in our study. Moreover, an analysis of correlations between *knowledge of what to assess* and learning outcomes as well as an analysis of differences between known groups is interesting regarding validation, leading to the second research question.

2. To what extent is the *knowledge of what to assess* related to similar constructs and learning outcomes? To what extent can differences be found in the *knowledge of what to assess* between students at the undergraduate and graduate levels?

Regarding research question two, we expect correlations between *knowledge of what to assess* and *diagnostic competence* as well as *examining competence.* Both, the instrument for *diagnostic competence* regarding experimentation competences and the instrument for *examining competence* share a focus on experimentation with our construct. A lower correlation than between *knowledge of what to assess* and *diagnostic competence* and *examining competence* is expected between *knowledge of what to assess* and *self-efficacy beliefs* regarding teaching biology since *self-efficacy beliefs* are based on a broader range of knowledge than the *knowledge of what to assess* regarding experimentation competences.

We expect correlations between *knowledge of what to assess* regarding experimentation competences and grades as an indicator for learning outcomes in high school biology, in biology at university, and biology teacher education courses at university. Since biology teacher education courses can deal with assessment and student conceptions, the grade in biology teacher education is expected to correlate highest with our construct. Moreover, we expect correlations between *knowledge of what to assess* regarding experimentation competences and the number of respective learning opportunities.

Students at the graduate level are hypothesized to outperform students at the undergraduate level because the former are expected to have acquired more *knowledge of what to assess* during their teacher education studies than students at the undergraduate level. Assessment, knowledge and skills in experimentation, and knowledge of student conceptions in biology are prescribed contents for biology teacher education [12]. Therefore, we hypothesize the following:

Students at the graduate level reach higher person abilities in the *knowledge of what to assess* regarding experimentation competences than students at the undergraduate level.

Once a reliable and valid measurement instrument has been developed, the third research question aims at providing information about pre-service teachers' *knowledge of what to assess*.

3. What are the strengths and weaknesses of pre-service biology teachers regarding *knowledge of what to assess* regarding experimentation competences in biology?

2. Methods

2.1. Participants and Data Collection

The study was conducted from October 2014 to February 2015, including pre-service biology teachers from 18 German universities in seven federal states. We analyzed questionnaire answers of $n = 495$ pre-service biology teachers (78.1% female, mean age = 23.15 years, $SD = 3.20$ years; the gender distribution represents the higher percentage of female pre-service teachers in Germany). Five people of $N = 500$ were excluded from analyses due to missing data or improper handling of the questionnaire. The participants of the study covered a range of different semesters in Bachelor, Master, or State Examination studies (34.3% Bachelor, 41% Master, 24.7% State Examination). In the following the term *students at the undergraduate level* comprises students in their Bachelor studies as well as students striving for the State Examination degree ≤ semester 6. The term *students at the graduate level* comprises students in their Master studies as well as students striving for the State Examination degree ≥ semester 7. The study participants were seeking to become primary, secondary, and vocational school teachers or special education teachers. In Germany, both the Master and First State Examination degree qualify for teaching practice in the second phase of teacher education.

Data were collected using a paper-pencil questionnaire which recorded (a) demographic and academic information, (b) *knowledge of what to assess* regarding experimentation competences in biology, and (c) *diagnostic competence* or *self-efficacy beliefs* for teaching biology. An instrument to measure *examining competence* [35] was part of a parallel conducted study focusing on teaching competences for experimentation in biology [36]. The study on teaching competences for experimentation in biology and our study has an overlapping sample of pre-service teachers who answered both questionnaires. Two research associates and one student assistant surveyed data collection using a standardized procedure at 18 universities in seven federal states.

2.2. Measurement Instrument

For the measure of *knowledge of what to assess* regarding experimentation competences in biology, we built on the instrument of Bögeholz et al. [36], keeping well-functioning items and shortening the instrument to not exceed 90 minutes in testing time. These measures facilitated testing in sessions of seminars and made the testing time acceptable for pre-service teachers outside of seminars. Furthermore, it supported (test) performance by preventing a decrease in motivation and an increase in fatigue [37]. Thus, seven out of the initial 27 scenarios portraying different phases and competences of experimentation were chosen and adapted accordingly from Bögeholz et al. [36].

Each of the seven scenarios described an experimentation assignment for a biology lesson with hypothetical high school students and the response of a single student or a group of students (Figure 1). Pre-service teachers were asked to assess the response of the hypothetical student(s). For some scenarios, they had to explain the student conception that influenced his/her procedure and in some cases to correct the solution in addition. The applied contexts covered the required basic curricular content. Relevant information for the experiments was given so that no additional content knowledge about the contexts was required.

The measurement instrument for *knowledge of what to assess* regarding experimentation competences in biology consisted of seven biology lesson scenarios (see Figure 1 for an example) covering the phases

of hypothesis formation, design of an experiment, and analysis of data. Each phase was focused on at least in two scenarios in different contexts that were chosen in consideration of German core curricula for biology [38].

Mrs. Nell discusses the characteristics of enzymes with her students in class 10. She asks her students to examine at which temperature α-amylase breaks down starch the fastest.

Mrs. Nell gives instructions on how to design the experiment:
Three test tubes are each filled with the same amount of starch solution. Then the starch solutions in the test tubes are placed in different water baths to reach the desired temperatures: 10°C, 40°C, and 70°C. To each starch solution, α-amylase of the same temperature (10°C, 40°C, 70°C) is added. α-amylase breaks down starch into maltose and glucose. Then every minute, a drop of the starch solution of every test tube is taken and added to brown iodine solution. When starch is added to the iodine solution, the color of the mixture changes from brown to blue. When maltose and glucose are added to the iodine solution, the brown color does not change.

Steps conducted by the student Bea:

Bea's hypothesis

α-amylase breaks down starch the faster, the higher the temperature of the starch solution is.

Bea's design and performance

Test tube 1: Starch solution, α-amylase, 10°C
Test tube 2: Starch solution, α-amylase, 40°C
Test tube 3: Starch solution, α-amylase, 70°C

A drop of each solution is added after 1, 2, 3, 4, and 5 minutes to 2 drops of iodine solution, respectively.

Results of the experiment

	1 min	2 min	3 min	4 min	5 min
10°C	*blue*	*blue*	*blue*	*brown*	*brown*
40°C	*blue*	*brown*	*brown*	*brown*	*brown*
70°C	*blue*	*blue*	*blue*	*blue*	*blue*

blue: Starch is still present
brown: Starch has been broken down into maltose and glucose

Bea's conclusion

The activity of α-amylase increases with rising temperature.

Tasks for pre-service teachers:

1. Assess Bea's data analysis. Give reasons. (Item 15)
2. Explain how Bea could have come to her conclusion. (Item 16)

Figure 1. Biology lesson scenario with assessment tasks for pre-service teachers (slightly adapted layout).

The context seed germination (scheduled for grades five and six) was represented in three scenarios. The contexts photosynthesis (scheduled for grade seven and eight) and enzymology (scheduled for grade nine and ten) were each represented in two scenarios [38]. The composition of the questionnaire is shown in the matrix of Table 1. The corresponding item list is displayed in Table A1 in Appendix A.

Table 1. Matrix of contexts x phases of experimentation with scenarios and corresponding items (see Table A1 in Appendix A).

	Seed Germination	**Photosynthesis**	**Enzymology**
Hypothesis formation	Scenario 1 with items 1, 2, 3	Scenario 4 with items 4, 5, 17, 18	
Design of an experiment	Scenario 2 with items 6, 7, 19	Scenario 5 with items 13, 14, 20	Scenario 6 with items 8, 9, 10
Analysis of data	Scenario 3 with items 11, 12		Scenario 7 with items 15, 16

The task format required study participants to assess students' experimentation competences according to central criteria. The following criteria were used for the phase of hypothesis formation: comprehensive hypothesis formation and, concerning single hypotheses, being testable and founded. Regarding the phase of designing an experiment, the following criteria were used: systematic variation of variables and precise design. Furthermore, the following criteria were used for assessing the planning of the performance: accurate measurement procedures and standardization. Concerning the phase of data analysis, the following criteria were used: correct data analysis, precise data analysis, error analysis, and conclusion with a link to the hypothesis.

Moreover, the two student conceptions engineering mode of experimentation and confirmation bias had to be assessed. The implementation of the criteria of all three phases and student conceptions was realized in different categories of items: assessing student conceptions, assessing correct student solutions, and assessing incorrect student solutions. Because the tasks measuring pre-service biology teachers' assessment literacy were based on scenarios, we expected them to have curricular validity and to be motivating. The task format was close to real-world performance tasks and focused on the criteria for experimentation that are relevant for learning outcomes in biology at high school.

2.3. Coding of Knowledge of What to Assess Regarding Experimentation Competences

For each biology lesson scenario, two to four items were coded (see Tables 2 and 3 and Table A1 in Appendix A). The coding was a further development of the coding applied in the pilot study [36]. It was equally distributed to four persons and carried out according to a manual which was deductively and inductively developed [39]. The scoring of the answers to the tasks considered correctness, completeness, and accuracy (Table A1). Ten trichotomous items had a maximum score of 2 (scores 0, 1, 2) (Tables 2 and 3); ten dichotomous items had a maximum score of 1 (0, 1). For the dichotomous items, the maximum score was relativized to 2, assigning all items the same weight. A randomly chosen representative tenth of the test booklets, i.e., 52 test booklets, was analyzed by all four persons to investigate the inter-coder reliability. A sufficient power of kappa was reached with this sub-sample [40]. However, an analysis of Krippendorff's alpha was preferred for ordinal data. Krippendorff's alpha was analyzed for the most differentiated version of the scoring rubrics before item steps were combined. Four of the 20 items reached a Krippendorff's alpha below 0.70. For the other 16 items, Krippendorff's alpha was between 0.70 and 0.87. A low Krippendorff's alpha could be explained by the open-ended tasks and the original superfine scoring. After combining item steps, it can be assumed that Krippendorff's alpha improved [41].

Table 2. Scoring of Item 15 (experimentation phase: analysis of data, criterion: incorrect data analysis) – task of item 15: "Assess Bea's data analysis. Give reasons." (cf. Table A1 in Appendix A).

	Scoring	Exemplary Answers
Score 2	The criterion is named and explained.	The data analysis is wrong. No transformation could also be detected at 70 °C. (1.11) Bea's data analysis is not detailed enough. The efficiency of α-amylase increases up to 40 °C, but above that, no splitting takes place at all. Therefore, Bea's conclusion is wrong. (1.13)
Score 1	The criterion is named.	The data analysis is incomplete since not all data have been taken into account. (1.9) Her conclusion is wrong. It is possible that the relationship is not clear to her: that higher enzyme activity can explain the splitting of starch and therewith the change to a brown color. (1.17)
Score 0	The criterion is neither named nor explained.	The table would have been better the other way around. 10 °C 40 °C 70 °C 10 °C 40 °C 70 °C 1 min x x x 4 min o 2 min x 5 min o 3 min x (1.1) The data analysis in a table is good. The intervals increase constantly and everywhere equally. (1.7)

Table 3. Scoring of Item 16 (experimentation phase: analysis of data, criterion: confirmation bias) – task of item 16: "Explain how Bea could have come to her conclusion." (cf. Table A1 in Appendix A).

	Scoring	Exemplary Answers
Score 2	The criterion is explained completely. The explanation includes both of the following aspects: (1) student ignores the observation (of the 70 °C test tube) OR the student does not consider the result (of the 70 °C test tube) due to certain reasons. (2) student has a specific belief concerning the outcome of the experiment OR the student tends to confirm the hypothesis.	Bea looks for clues that confirm her hypothesis. She ignores other results of her experiment since they don't fit her belief. (confirmation bias effect?) (1.16) She might conclude, due to previous knowledge, that reactions take place faster at higher temperatures. With the experiment, she verifies her own expectations and ignores contradicting results. (1.70)
Score 1	The criterion is explained in parts. The explanation includes one of the two following aspects: (1) student ignores the observation (of the 70 °C test tube) OR the student does not consider the result (of the 70 °C test tube) due to certain reasons. (2) student has a particular belief concerning the outcome of the experiment OR the student tends to confirm the hypothesis.	Bea ignored the results of the 70 °C test tube. (1.13) Bea might have only compared the 10 °C and 40 °C and excluded 70 °C as a mistake. (1.111)
Score 0	The criterion is not explained.	Maybe she read her table falsely. To the right there are more and more brown fields that indicate that starch has been broken down. (1.1) Bea might have mixed up the variables time and temperature in her statement. (1.9)

2.4. Validation Instruments

In addition to demographic and academic information, *diagnostic competence*, *examining competence* and *self-efficacy beliefs* for teaching biology, were measured for validation purposes, each for a sub-sample.

Diagnostic competence for experimentation competences in biology was assessed with an instrument developed by Dübbelde [14]. This instrument was shortened from 17 to 12 items for the use in our study. The original 17 and remaining 12 items dealt with central conditions for experimentation, such as the foundation of the hypothesis, distinction between observations and conclusions, or link of conclusion to the hypothesis. The instrument consisted of hypothetical educational materials and products, i.e., high school students' worksheets and students' notes taken during an experiment, and an assessment sheet for pre-service teachers with 12 items focusing on the phases hypothesis formation, design of an

experiment, performance of the experiment in the sense of documentation and analysis of data. In the items, the pre-service teachers had to indicate whether certain conditions of experiments, such as performance of error analysis, had been fulfilled by the hypothetical students (nine items: "yes", "no", and "don't know") or identify the correct answer out of four choices (one item), out of three choices (one item) or out of three options, among that "don't know" (one item). The closed answer format (two choices plus "don't know") led to a high probability of guessing. The instrument with the original 17 items reached a Cronbach's alpha of 0.50 [14] (p. 189). The shortened instrument with 12 items that we applied for validation purposes had a Cronbach's alpha of 0.36 (n = 136). The instrument on *diagnostic competence* shared the focus on the assessment of students' experimentation competences with our instrument on *knowledge of what to assess*. However, the instrument of Dübbelde [14] asked for the estimation of the given criteria allowing guessing. The instrument did not focus on the personal *knowledge* of pre-service teachers on *what to assess*. Besides giving the criteria for experimentation, the *diagnostic competence* instrument differed from ours in that the instrument tested neither the knowledge of student conceptions nor the correction of specific incorrect hypothetical student solutions.

Examining competence in biology was assessed using a short scale (12 multiple-choice items) developed by Krüger et al. [35]. The instrument included the experimental phases of question formation, hypothesis formation, design of experiments, and analysis of data. Pre-service teachers had to select either a suitable question for an examination, a hypothesis that can be derived from observation, a hypothesis that is the basis of the examination, a design for the experiment that is suitable to test a specific hypothesis, or the correct data analysis of the experiment. For all choices to be taken, one out of four answers was correct. Criteria for experimentation, such as holding the independent variables constant in an experiment, have to be applied to select the correct answer. Moreover, the instrument captures contents of the knowledge base for the assessment of experimentation competences. It differs from our test on *knowledge of what to assess* regarding experimentation competences in that the criteria for experimentation do not have to be named, explained, or described. A "feeling" for how to design an experiment, for instance, is sufficient to solve the tasks. And again, guessing can also lead to the correct answer, up to 25% of the time. The instrument reached a Cronbach's alpha of 0.39 (n = 239) in our study.

The third instrument applied for validation purposes measured pre-service teachers' *self-efficacy beliefs* for teaching biology. On a Likert scale, pre-service teachers had to indicate their expected abilities concerning, for instance, planning and conducting lessons in consideration of research results on biology education, such as research results regarding student conceptions (four items) and planning lessons in consideration of core concepts ("Basiskonzepte") of biology, such as structure and function, and competences for biology (two items) [42]. Both, research results on biology education, as well as core concepts and competences for biology, comprise information relevant for experimentation in the classroom: The ability to plan and conduct lessons in consideration of research results on biology education includes the knowledge of and ability to use research findings on students' biological conceptions. The competences for biology comprise experimentation competences.

2.5. IRT Modeling and Further Analyses

Data analysis was conducted using the partial credit model [43]. Item Response Theory (IRT) analyses were conducted with ConQuest [44]. For item related analyses, the average person's ability was set to zero (=case-centered analysis, constraints = cases). Due to this procedure, also the item difficulty of the last test item could be estimated correctly. For person related analyses, the average item difficulty was set to zero (=item-centered analysis, constraints = items) [44].

The data quality was checked for the one- and 3D model via fit statistics ($0.8 \leq$ wMNSQ ≤ 1.2; $-2 \leq$ t-value ≤ 2) resulting from case-centered IRT analyses [45] (p. 164 ff.) [46] (p. 270ff.). Item-centered analyses were conducted to estimate person-measures and compare the fit of the two models. The deviance, as well as Bayesian information criterion (BIC) and Akaike's information criterion (AIC) and latent correlations between the dimensions, were computed. An analysis of differential item

functioning (DIF) was conducted with ConQuest to identify items that were biased for the educational level or gender.

For validation, *knowledge of what to assess* regarding experimentation competences and the related constructs *diagnostic competence* with 12 items [14] and *examining competence* with 12 items [35] were analyzed by multidimensional modeling, and latent correlations were examined. The multi-dimensional case-centered analysis (one dimension for each of the three constructs above) provided fit statistics for the items of the three scales. Moreover, manifest correlations between *knowledge of what to assess* regarding experimentation competences and different *self-efficacy beliefs* concerning *planning and conducting lessons in consideration of research results on biology education* and *planning lessons in consideration of core concepts and competences for biology* [42] were analyzed. Correlations between *knowledge of what to assess* regarding experimentation competences, grades, and the number of learning opportunities, were computed for a further check of validity. In addition, a Mann-Whitney-U-test was applied with person measures to examine whether students at the graduate level outperform students at the undergraduate level in their *knowledge of what to assess* regarding experimentation competences in biology. Using item difficulties, we examined which criteria of experimentation are easy or difficult to assess for pre-service biology teachers (strengths and weaknesses of pre-service teachers). We compared item difficulties of different item groups using one-way ANOVA and a post hoc Tukey HSD test for specific group comparisons. In sum, we used the steps of Figure 2 to analyze the data.

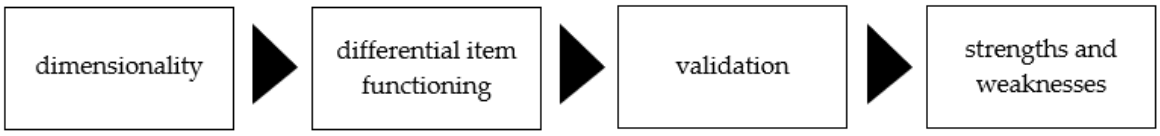

Figure 2. Foci of data analyses of pre-service teachers' *knowledge of what to assess.*

3. Results

3.1. Modeling and Measuring Knowledge of What to Assess Regarding Experimentation Competences

3.1.1. Dimensionality

The case-centered 1D modeling of *knowledge of what to assess* regarding experimentation competences as well as 3D modeling with the dimensions hypothesis formation, design of an experiment, and analysis of data reached comparable reliabilities and item parameters. The EAP/PV for the 1D model was 0.60 and lay between 0.50 and 0.54 for the three dimensions of the 3D model (hypothesis formation: 0.50, design of experiments: 0.54, analysis of data: 0.51). The item fit was good for both models: The wMNSQ values ranged from 0.92 to 1.06 in the 1D modeling, and the corresponding t-values ranged from −0.9 to 1.7. The 3D modeling yielded wMNSQ values of 0.92–1.08 and t-values of −0.8–2.0.

Comparing the fit of the 1D and 3D model, item-centered analyses revealed the following (Table 4): the BIC that considers the model complexity indicated a better fit for the 1D model. Regarding the deviance and AIC, the 3D model (deviance = 13,236, *AIC* = 13,335.08) fit (slightly) better to the data than the 1D model (deviance = 13,279, *AIC* = 13,340.76). The latent correlations between the three dimensions ranged from 0.57 to 0.80 (Table 5). These rather low latent correlations indicated that the three dimensions captured different knowledge dimensions. Considering the construct that should be measured, however, the 1D modeling was more appropriate than 3D modeling. It covered *knowledge of what to assess* regarding experimentation competences more comprehensively and with an adequate number of items and was therefore applied for the following analyses.

Table 4. Comparison of the 1D and 3D model (item-centered analysis, n = 495).

Models	Deviance	Parameter	BIC	AIC
1D	13,279	31	13,471.10	13,340.76
3D	13,263	36	13,486.45	13,335.08

Table 5. Item-centered analysis of latent correlations between hypothesis formation, design, and analysis of data of *knowledge of what to assess* regarding experimentation competences (n = 495).

	Hypothesis Formation	Design
Hypothesis formation	—	
Design	0.68	—
Analysis of data	0.57	0.80

3.1.2. Test and Item Parameters

The case-centered 1D IRT modeling revealed acceptable reliabilities (EAP/PV reliability = 0.60, item separation reliability = 0.994) (Table 6). The variance of 0.14 indicated that the differentiation between persons was low. As stated in the section "dimensionality", the item fit was good ($0.8 \leq$ wMNSQ ≤ 1.2; $-2 \leq$ t-value ≤ 2). The item difficulties of the 1D model ranged from −1.34–2.22 logits. All item steps had been reached by at least 5% of the pre-service teachers, except for Item 6, step 2. The item step was maintained due to the relevance of the content: knowledge of the student conception engineering mode of experimentation. The discrimination of the items reached acceptable values above 0.25 [47] (p. 147) except for Item 6, focusing on the student conception engineering mode of experimentation (0.15) and Item 18 dealing with the correction of an unfounded hypothesis (0.21). Both items were kept due to the relevance of their content.

Table 6. Parameters of the case-centered 1D IRT modeling of *knowledge of what to assess* regarding experimentation competences.

	1D Model
Total Number of items (dichotomous/trichotomous)	20 (10/10)
EAP/PV reliability, item separation reliability	0.60, 0.99
Variance	0.14
Item difficulty: min to max	−1.34–2.22
Person ability: min to max	−2.85–1.33
wMNSQ: min to max	0.92–1.06
T value: min to max	−0.9–1.7
Discrimination: min to max	0.15–0.45

3.1.3. Differential Item Functioning

Considering the 1D modeling of *knowledge of what to assess* (scale with 20 items), students at the undergraduate level (n = 253) scored 0.32 logits lower than students at the graduate level (n = 224). Differential item functioning (DIF) existed for Item 7 unsystematic variation of variables (logit difference = 0.74). Item 7 was the easiest item for students of both groups. It was considerably easier for students at the graduate level (solved by 98% of students at the graduate level and 87% of students at the undergraduate level). The logit difference for all other items was below 0.4. Thus, it was not regarded as a considerable DIF [48] (p. 12). No considerable DIF occurred for gender (maximum logit difference = 0.28).

The Wright Map (Figure 3) shows that nine items/item steps out of 30 items/item steps were complicated. Consequently, they did not differentiate very well between person abilities. For the range of −1.00–0.70 logits, the distribution of the item difficulties matched the person's abilities well, except for a minor gap of items between Item 8 and 9.

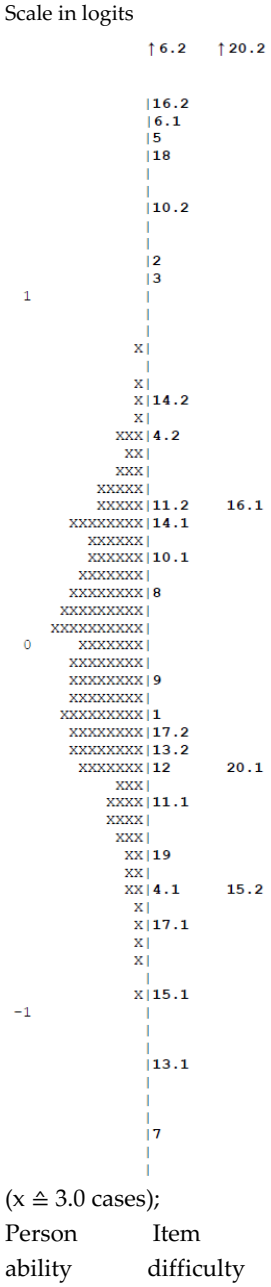

Figure 3. Wright Map of the case-centered 1D IRT modeling of *knowledge of what to assess* regarding experimentation competences (↑ = the item difficulty is greater than presented).

3.2. Validation of Knowledge of What to Assess with Related Constructs, Educational Outcomes, and Comparison of Known Groups

3.2.1. Relationship to Related Constructs

The 3D (case-centered) IRT modeling of *knowledge of what to assess* regarding experimentation competences, *diagnostic competence*, and *examining competence* showed that the fit of all items of the three instruments was good ($0.8 \leq$ wMNSQ ≤ 1.2; $-2 \leq$ t-value ≤ 2).

Analyses revealed that pre-service teachers reached the highest person measures for *diagnostic competence* (mean person ability = 1.10) (Table 7). Lower person measures were reached for *examining competence* (mean person ability = 0.16) and the lowest for *knowledge of what to assess* regarding experimentation competences (mean person ability = −0.14). Thus, it was the most difficult construct. Table 8 shows the latent correlations between the constructs.

Table 7. Parameters of the item-centered 3D IRT modeling (n = 128).

	Mean	Variance	EAP/PV
Knowledge of what to assess	−0.14	0.09	0.58
Diagnostic competence	1.10	0.27	0.45
Examining competence	0.16	0.29	0.59

Table 8. Latent correlations between the three constructs (item-centered analysis) (n = 128).

	Knowledge of What to Assess
Diagnostic competence	0.37
Examining competence	0.78

The highest latent correlation existed *between knowledge of what to assess* regarding experimentation competences and *examining competence* (0.78). The latent correlation between *knowledge of what to assess* regarding experimentation competences and *diagnostic competence* was relatively low (0.37).

The analysis of correlations (Spearman) between *knowledge of what to assess* regarding experimentation competences and *self-efficacy beliefs* for teaching biology revealed the following results: *Knowledge of what to assess* regarding experimentation competences of students at the graduate level correlated with their *self-efficacy beliefs* regarding the ability to *plan and conduct lessons in consideration of research results on biology education* s (r = 0.20, $p < 0.05$, n = 146) as well as *self-efficacy beliefs* regarding the ability to *plan lessons in consideration of core concepts and competences for biology* (r = 0.22, $p < 0.01$, n = 147). In contrast, no relationship between the variables was found for students at the undergraduate level ($p > 0.05$; n = 178, n = 181).

3.2.2. Relationship to Grades and Learning Opportunities

Table 9 shows the correlations of person abilities in the *knowledge of what to assess* regarding experimentation competences with educational variables. Better grades in high school biology as well as in courses of biology and biology teacher education at university correlated positively with person measures: The more achieved points at high school (r = 0.19, $p < 0.01$) and the lower (that is, the better) the university grade in biology (r = −0.16, $p < 0.01$) and biology teacher education (r = −0.28, $p < 0.01$), the higher were the person abilities. There was a strong correlation between university grades in biology and biology teacher education (r = 0.63, $p < 0.01$).

The amount of learning opportunities correlated with *knowledge of what to assess* regarding experimentation competences: The more courses in biology teacher education pre-service teachers had completed, the higher the person abilities in the *knowledge of what to assess* regarding experimentation competences.

Table 9. Correlations between *knowledge of what to assess* and educational variables.

	High School	University		
Variable	Last grade in biology in high school	Average grade in university courses in biology	Average grade in university courses in biology teacher education	Number of com-pleted courses in biology teacher education
Person ability	$0.19^{2}s$ (n = 446)	$-0.16^{2}s$ (n = 377)	$-0.28^{2}s$ (n = 265)	$0.21^{2}p$ (n = 406)

Legend. s = Spearman, p = Pearson, ^{2} *= $p < 0.01$;* person ability: test result (20 items of *knowledge of what to assess*); last grade in biology in high school: 1 = very poor, up to 15 = very good; average grade in courses in biology as well as biology teacher education: 1.0–1.3 = very good, 1.7–2.3 = good, 2.7–3.3 = satisfactory, 3.7–4.0 = sufficient; the number of completed courses in biology teacher education: 1 = 1, up to 10 = 10.

3.2.3. Comparison of Known Groups

Regarding the 1D model of *knowledge of what to assess* regarding experimentation competences, Q-Q-plots indicated that data were normally distributed for the students at the graduate level (n = 224) but not for students at the undergraduate level (n = 254, negatively skewed). The Levene-test indicated

that homogeneity of variances could not be assumed for the two groups ($p = 0.025$). The Mann-Whitney U-test was applied: There was a statistically significant difference in person abilities between students at the undergraduate level ($M_{Rank} = 207.25$) and at the graduate level ($M_{Rank} = 276.07$), $U = 20{,}255.50$, $Z = -5.447$, $p < 0.001$, with a moderate effect size ($r = 0.25$). Consequently, the person ability increased in the course of academic studies (Figure 4), which is in line with our hypothesis.

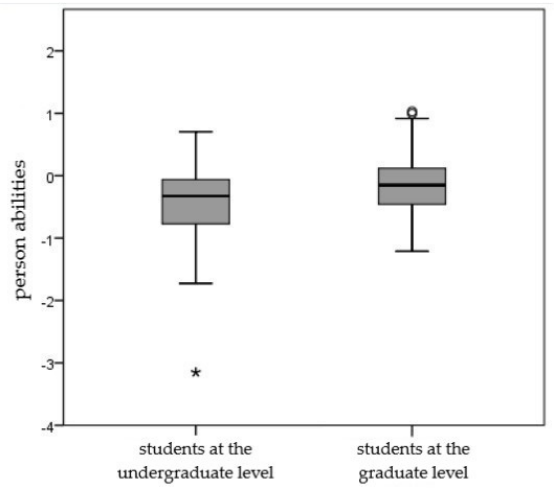

Figure 4. Person abilities of students at the undergraduate and graduate level (item-centered analysis, 1D modeling; undergraduate level: Bachelor + State Examination degree ≤ semester 6, graduate level: Master + State Examination degree ≥ semester 7).

3.3. Strengths and Weaknesses Concerning Knowledge of What to Assess Regarding Experimentation Competences

Analyzing the distribution of items on the Wright Map (Figure 5), we were able to identify specific contents that influence item difficulty. We were able to group these contents into four categories. Category i focusses on the assessment of student conceptions. Category ii deals with the assessment of correct student solutions. The further two categories comprise the assessment of incorrect student solutions: Category iii focusses on the assessment of the planning of the performance with regard to standardization and accuracy and category iv on the assessment of further incorrect student solutions. In the following, we describe the contents of the four categories in order of decreasing difficulty.

Ad (i) Student conceptions were displayed in the trichotomous Item 6 (2.22 logits, Figure 5) and Item 16 (0.98 logits). Item 6 focused on the student conception engineering mode of experimentation. Item 16 dealt with the student conception confirmation bias. The assessment of student conceptions was very difficult. Only a few pre-service teachers named and explained the engineering mode of experimentation (full credit) or explained the confirmation bias comprehensively (full credit).

Ad (ii) The assessment of correct student solutions included the assessment of a testable hypothesis (Item 2), a theoretically founded hypothesis (Item 3), and the systematic variation of variables (Item 8). The item difficulties of this group of items ranged from 0.15 to 1.10 logits.

Ad (iii) The assessment of criteria concerning the planning of a standardized and accurate procedure was difficult but less complicated than assessing student conceptions and correct student solutions. The item difficulty for Item 10, which required naming that several aspects lacked standardization, was 0.77 logits. The item difficulty for Item 14, which required naming and explaining that the measurement procedure was inaccurate, was 0.53 logits.

Ad (iv) In comparison, the assessment and correction of incorrect student solutions were relatively easy: The item difficulty of Item 4 dealing with the assessment of an untestable hypothesis was 1.11 logits below the item difficulty of Item 2 requiring the assessment of a testable hypothesis. The correction of an untestable hypothesis (Item 17) was even easier. The assessment (Item 7) and correction (Item 19)

of an unsystematic variation of variables was considerably easier than the assessment of the systematic variation of variables (Item 8). The logit difference in Items 7 and 8 was 1.49 (Figure 5). Further items required the assessment of incorrect student solutions, i.e., incomprehensive hypothesis formation (Item 1, −0.15 logits), imprecise design of experiment (Item 9, −0.09 logits), imprecise data analysis (Item 11, −0.02 logits), missing error analysis (Item 12, −0.30 logits), conclusion without a link to hypothesis (Item 13, −0.72 logits) and incorrect data analysis (Item 15, −0.80 logits) were relatively easy. The whole item group had item difficulties ranging from −1.34 to −0.02 logits.

There were three exceptions regarding this item category: Item 5 and 18 dealing with the assessment and correction of an unfounded hypothesis, i.e., a hypothesis without justification (Item 5, 1.43 logits; Item 18, 1.40 logits) and Item 20 dealing with the correction of a conclusion without a link to the hypothesis (1.01 logits). The difficulties of Item 5 and 18 could have been influenced by the task format that required assessing (Item 5) or correcting (Item 18) several mistakes for one task, namely the missing foundation of the hypothesis (Item 5 and 18) in addition to the missing testability of the hypothesis (Item 4 and 17). Item 20 required the formulation of a conclusion with reference to the hypothesis and its verification for full credit. Few pre-service teachers went beyond the correction of the content of the given hypothetical student answer (scored with partial credit) and verified the hypothesis. No statement regarding the missing verification of the hypothesis was required for full credit of the 1.73 logits easier Item 13 covering the assessment of a conclusion without a link to the hypothesis. These three items were exceptional cases as they required demanding information processing for generating additional solutions. The task format could have influenced the item difficulty. Therefore, we excluded them from the item category iv.

A one-way ANOVA revealed that the four item categories differed significantly ($F(3, 13) = 13.64$, $p < 0.001$). A post hoc Tukey HSD test indicated that the mean item difficulty for the items of category iv (item group iv) assessing incorrect student solutions (mean = −0.45, $SD = 0.42$) differed significantly from the three other item groups (vi versus i: mean = 1.60, $SD = 0.88$, $p = 0.001$; vi versus ii: mean = 0.77, $SD = 0.54$, $p = 0.009$; vi versus iii: mean = 0.65, $SD = 0.17$, $p = 0.048$). The other three item groups i, ii, and iii did not differ significantly in their mean item difficulty from each other ($p > 0.05$).

Regarding the three phases of experimentation, no significant differences between mean item difficulties of the three phases were found, considering all 20 items ($p > 0.05$; Figure 5).

Overall, relatively few pre-service teachers used certain scientific terms, such as engineering mode, confirmation bias, error analysis, and control group, to assess the hypothetical high school students' experimentation competences. Most of them described the criterion without naming these specific terms. For instance, regarding item group i, 76 pre-service teachers identified the idea of the underlying engineering mode of experimentation (evaluated by partial credit), only seven of them used the scientific term engineering mode. Forty-nine pre-service teachers assessed the confirmation bias; only one of them named the student misconception confirmation bias. Regarding item group iv, the idea of missing error analysis was perceived by 313 pre-service teachers and the term named by only 34 of them. In the assessment of the unsystematic variation of variables, only 94 of 456 pre-service teachers used the term control group.

Figure 5. Wright Map of the case-centered 1D IRT modeling of *knowledge of what to assess* regarding experimentation competences (without item steps) (↑ = the item difficulty is greater than presented).

4. Discussion

In the following section, the results are discussed in the order of the research questions focusing on dimensionality and test quality, validation, and strengths and weaknesses of pre-service biology teachers.

4.1. Dimensionality and Test Quality

Concerning dimensionality, one could argue for a 1D model or a 3D model considering the modeling with the partial credit model. Taking into account the construct *knowledge of what to assess* regarding experimentation competences, a 1D model is the preferred option because of considerations set out in the following: The 1D parsimonious approach regarding the competence construct in science education represented an advantage for empirical testing cf. [49] and for teaching and assessing taking into account the other competences to be learned by pre-service teachers. For instance, pre-service teachers also have to learn to analyze and plan lessons to foster high school students' experimentation competences [36] and in the frame of assessment literacy they have to acquire knowledge of assessment purposes, knowledge of assessment strategies, and knowledge of assessment interpretation and action-taking [16]. Thus, we prioritized a more manageable conceptualization as opposed to more differentiated analyses possible with a more complex competence construct for *knowledge of what to assess* cf. [49] (p. 63). The benefits of operating with a broader construct for practical usefulness outweighed, in this case, a more differentiated conceptualization. Thus, arguments for multi-dimensionality such as the given latent correlations receded into the background.

The phenomenon that empirical results regarding experimentation related knowledge did not provide a clear picture concerning dimensionality was not only given for the construct *knowledge of what to assess* in the group of pre-service teachers. For example, varying results regarding dimensionality occurred in the research on similar constructs such as *scientific inquiry*, *scientific reasoning*, and *experimentation competences* investigating high school students. We summarize this research and structure the summary by proceeding from the more advanced students to the less advanced students: (i) Research with 10th graders on *scientific inquiry* revealed high latent correlations between the scales question, hypothesis, design, and data analysis (0.80–0.95) [22] (p. 132). It turned out that neither a four-dimensional model (comprising question, hypothesis, design, and data analysis) nor a 3D model (comprising observing, comparing, experimenting) outweighed a 1D model of *scientific inquiry*, which was in line with the approach to operate with a manageable amount of competence models for teaching biology. (ii) An analysis of high school students' *scientific reasoning* of grade 5–10 found weak and intermediate latent correlations of 0.33–0.73 between the subscales question, hypothesis, planning, and interpretation [33] (p. 56ff.). The author argued for a four-dimensional model. For this age group, the students were in the phase of acquiring knowledge on the phases that make up *scientific reasoning*. (iii) For the construct *experimentation competences*, manifest correlations of 0.38–0.74 were found for grade five and 0.64–0.78 for grade six between the three subscales of the SDDS model of Klahr [20]: search hypotheses, test hypotheses and evaluate evidence [34] (p. 42). While the study of Wellnitz with 10th graders suggested a 1D model, the other studies with younger students pointed out that experimentation related constructs require at least a two-dimensional model [33,34]. The phenomenon could be explained by a more integrative and interwoven processing of specialized knowledge coming along with study progress cf. [50]. More generally speaking, the fact of low latent correlations between the dimensions of *knowledge of what to assess* regarding experimentation competences was not surprising regarding the educational target groups within our study. A remarkable percentage of pre-service teachers, i.e., the undergraduates (53% of the sample), had not had very much biology teacher education courses (mean number of courses completed = 1.2) until their participation in our study.

For pre-service biology teachers, the subscales of *knowledge of what to assess* regarding experimentation competences showed the lowest latent correlations between hypothesis formation and analysis of data (0.57). In contrast, these two subscales correlated highest in studies conducted with high school students (correlation: 0.73 and 0.78 [33] (p. 58) [34] (p. 42)). For these students, the correlations between hypotheses and interpretation or between search hypotheses and evaluate evidence are explained by a greater relevance of domain-specific knowledge and less relevance of methodological knowledge in comparison to the phase planning/testing hypothesis [34] (p. 45). In our study with pre-service biology teachers, hardly no additional knowledge of biological phenomena was required to solve the tasks. For instance, a scenario included the information that the amount of

released gas bubbles indicates the rate of photosynthesis of waterweed in the experiment. It was only required for pre-service teachers to know that oxygen is a product of photosynthesis to link a greater amount of gas bubbles to a greater rate of photosynthesis to interpret data given in the following scenario. Learning that oxygen is a product of photosynthesis is the content of school curricula for grade seven/eight [51]. Therefore, in our study, knowledge of biological phenomena should not have influenced the test results in contrast to the studies investigating high school students. The highest correlation of subscales of *knowledge of what to assess* regarding experimentation competences existed between the design of an experiment and analysis of data (0.80), which was analogous to findings of Wellnitz's study of *scientific inquiry* [22] (p. 132). This could result from a stronger focus on these phases in research studies cf. [52,53] and perhaps as a consequence of teaching at university.

Regarding test quality, the 20 final items of *knowledge of what to assess* regarding experimentation competences in biology had a satisfactory item fit and discrimination for the 1D as well as the 3D model—except for the discrimination of Item 6 (0.15) and Item 18 (0.21). The low discrimination of items 6 and 18 could be explained by their great difficulty [54]. Only a few students solved both items. According to the contents of these items, we were interested in the knowledge of the term engineering mode of experimentation and its description (6). Second, to correct a hypothesis that is unfounded turned out to be a challenge. In biology teaching, there is a lack of clear rules concerning how to justify hypotheses. Several approaches exist that range from not addressing the fact that a hypothesis should be well-founded to expecting that a reason is given for the hypothesis [55,56]. Up to now, the issue of backing up hypotheses is not focused coherently in textbooks for school or teacher education [55–57].

The accuracy of the estimated item difficulties of the IRT analyses was given by the high item separation reliability of 0.99. The EAP/PV value, indicating the accuracy of the estimated person abilities, of 0.60 was comparable to tests measuring similar constructs. For example, a study measuring *scientific inquiry* (observing, comparing, experimentation) with 116 items [22] (p. 129) reached an EAP/PV reliability of 0.59 for high school students. Thereby, the subscale experimentation (22 items) reached a reliability of 0.41 (EAP/PV) in a 3D model with observing (0.37, 18 items) and comparing (0.39, 10 items) [22] (p. 136f.). Similar results were reached for an instrument measuring *scientific reasoning* with 24 items (EAP/PV = 0.69 (study I) and 0.68 (II) [33] (p. 51, 53)) and an instrument measuring *diagnostic competence* for students' experimentation competences with 17 items (Cronbach's alpha = 0.50, [14] (p. 189)). The measurement of a construct with full content and a limited number of items is in line with reduced reliability [58]. Since our construct *knowledge of what to assess* regarding experimentation competences covers the three phases of experimentation, the broad approach is reflected in the reliability of the instrument.

Moreover, low variances and open-answer formats can contribute to lower reliability [58,59]. On the upside, open tasks can measure skills closer to real-life performance than multiple-choice items and provide additional information [60].

In our study, some items were too difficult. While providing valuable information about *knowledge of what to assess*, they were not beneficial for precise measurement. Excluding difficult items or collapsing item steps could improve the quality of the instrument. More items for low and intermediate person abilities would improve the accuracy of the measurement [45] (p. 125f.). The low variances could result from a relatively homogenous sample of test persons (i.e., pre-service biology teachers).

Furthermore, in 2014/2015, *knowledge of what to assess* regarding experimentation competences might have hardly been addressed in teacher education courses, which is line with the lack of connecting CK, PCK, and PK (pedagogical knowledge) in German teacher education in that time [61]. Only in the last four years have there been nationwide efforts to systematically link these three knowledge areas further to develop the quality of teacher education [61]. However, each university, funded by the German Federal Ministry of Education and Research within the "Qualitätsoffensive Lehrerbildung", could decide its priorities for further developing their teaching. Thus, only a few universities addressed linking CK and PCK concerning competences in science (i.e., Technische Universität Braunschweig).

The chosen test length seemed suitable for measurement. One indicator was the difficulties of the items in the last scenario: Item 15 focusing on incorrect data analysis was comparably easy. This indicated that no respondent fatigue occurred. In contrast, item 16 (item category i) focuses on the student conception confirmation bias, which could explain why this item is more difficult than item 15.

The measurement instrument could be applied to undergraduate and graduate students of biology education. Significant DIF in the 1D model could only be detected for one item (i.e., Item 7 dealing with the systematic variation of variables). This item is considerably easier for graduate students than for undergraduate students. The fact could be due to an imprecise measurement related to the phenomenon of very low item difficulty. In addition, the DIF could be plausibly explained by specific training of the control of variables in (a) session(s) of teacher education courses, which was likely because the systematic variation of variables was one of the highlighted issues in reputable textbooks for German biology teacher education (e.g., [18]). In sum, the instrument was suitable to get an insight into pre-service teachers' *knowledge of what to assess* regarding experimentation competences in biology.

4.2. Validation

Latent correlations between the three constructs *knowledge of what to assess* regarding experimentation competences, *diagnostic competence* [14], and *examining competence* [35] were examined. *Knowledge of what to assess* regarding experimentation competences correlated highest with *examining competence* (0.78). The high correlation indicated a shared knowledge base. Both tests required knowledge about criteria for hypothesis formation, design of an experiment, and the analysis of data. Unexpectedly, the latent correlation between *knowledge of what to assess* regarding experimentation competences and *diagnostic competence* was comparably low (0.37). This could result from the test design. To solve the *diagnostic competence* tasks, the criteria for experimentation did not have to be known by the pre-service teachers. They were given, and pre-service teachers only had to identify whether they were fulfilled or not.

Moreover, the three instruments placed different emphasis on the successive phases of experimentation. Our instrument placed equal emphasis on the three phases hypothesis formation (seven items, 35% of items), design of an experiment (seven items, 35% of items), and analysis of data (six items, 30% of items). The instrument *diagnostic competence* had 16.6% of items focusing on hypothesis formation and 16.6% on the analysis of data. The majority, 67% of the items, dealt with the design of an experiment (42%) and performance in the sense of documentation (25%). The instrument for *examining competence* included the phase question formation, considering all four phases equally with 25% of the items. Considering the item distribution to the phases of experimentation in the three instruments investigated, the instrument for *examining competence*, and our instrument had a more similar emphasis on the different phases than the instrument for *diagnostic competence* and our instrument. The results have to be treated carefully due to the available instruments for related constructs for validation whose reliabilities are improvable.

The finding that only advanced students' *self-efficacy beliefs* correlated with *knowledge of what to assess* regarding experimentation competences could be explained by a better understanding of the contents addressed in the *self-efficacy beliefs* instrument by advanced students. During their studies, they engage with these topics and, consequently, they could achieve a more accurate ability to report on their *self-efficacy* regarding these subscales. Correlations of *knowledge of what to assess* with *self-efficacy beliefs* regarding *planning and conducting lessons in consideration of research results on biology education* and *planning lessons in consideration of core concepts and competences for biology* [42] were an indicator for validity since both *self-efficacy* subscales comprised information relevant for the assessment of experimentation competences.

As assumed, the number of learning opportunities and the performance in high school biology as well as biology courses and biology teacher education courses at university (grades) correlated with person abilities. This finding indicated that the test measured knowledge and skills acquired at university. The average grade in courses in biology teacher education showed a higher correlation with

knowledge of what to assess regarding experimentation competences than the average grade in courses in biology at university, which could be explained by the higher portion of biology school curricula procedural competences and contents used in the present study. The biology teacher education curriculum reflects the previously mentioned school curricula requirements to a certain degree [12].

The comparison of student abilities of students at the undergraduate and graduate levels showed higher person abilities for students at the graduate level, which was in accordance with our hypothesis regarding research question two and thus an indicator for validity. It underlined that the instrument measured knowledge that could probably be acquired during biology teacher education.

4.3. Strengths and Weaknesses of Pre-Service Biology Teachers Regarding Knowledge of What to Assess Regarding Experimentation Competences

Person abilities of pre-service biology teachers of *knowledge of what to assess* did not differ significantly for the three phases of experimentation.

Studies of high school students' *scientific reasoning* and *scientific inquiry* found that the interpretation of data analysis was more straightforward than the formation of hypotheses ([33] (p. 63) (grade 5–10) [22] (p. 141) (grade 10)). The findings for the phase design of an experiment were diverse and ranged from most difficult in some studies ([34] (p. 41) (grade 5 and 6); [33] (p. 63) to easiest in another [22] (p. 141)). The operationalization of the constructs could influence the results.

The finding of similar difficulties of the assessment of the three phases in our project was in line with skills pre-service teachers were expected to possess or acquire in their education. Having to teach and assess the whole process of experimentation, no significant differences in difficulties regarding the three phases should occur. However, specific criteria for experimentation proved to be challenging to assess, such as the founded hypothesis (Item 3, 5, 18). This criterion might not have been trained explicitly and intensively at school and university, which made it difficult to solve the tasks of the test instrument.

The restricted knowledge of scientific terminology by pre-service biology teachers in our study was striking. It could have been caused by a certain lack of precise communicative skills in the teacher education curriculum [12] and thus probably limited course time spent on teaching and practicing scientific terms. Furthermore, the study provided hints that misconceptions concerning experimentation competences were hard to identify for pre-service teachers, which could be explained by the fact that experimentation competences can benefit from different sources, such as CK taught in natural science subjects as well as from PCK taught in teacher education courses. Instead, student misconceptions were mainly taught about in PCK related teacher education courses. Comparing the portions of CK and PCK in the biology teacher education curriculum for secondary school teachers (that made up the most significant part of our participants), the share of PCK was much smaller than the share of CK [12]. In addition, the assessment of correct (item category ii) and incorrect student solutions (item category iv) was differently demanding. Correct student solutions in our study were a lot more challenging to assess than incorrect student solutions. Analogous to more complex features of compensatory decision-making in comparison with non-compensatory decision-making [62], the consideration of positive as well as negative aspects in student performance for an assessment was more demanding and consequently more difficult than concentrating on a mistake or disadvantage only.

Moreover, the assessment of specific criteria concerning the planning of the standardization and accuracy of the performance and measurement (item category iii) was very demanding. Despite dealing with incorrect student solutions, Item 10 (lack of standardization) and Item 14 (inaccurate measurement procedure) were difficult. The results could be explained by the neglect of these criteria in the curricula [51]. This was also reflected in the findings of high school students' (grade 12) experimentation competences. Less than 22% of high school students considered when and how often to perform measurements during the planning of an experiment [63].

Thus, the present study gave insights into which aspects of PCK relevant knowledge and skills concerning *knowledge of what to assess* regarding experimentation competences are already well

taught and learned. In addition, it revealed the remaining challenges for further developing biology teacher education.

4.4. Limitations

The comparison of the 1D and 3D models of *knowledge of what to assess* regarding experimentation competences did not provide a clear result regarding the dimensionality of the construct. Higher latent correlations were expected for the 1D model. An analysis with more items per subscale—so that all parts of the three subscales are better covered—could shed a brighter light on the question of dimensionality in order to evaluate how far the assessment of the three phases of experimentation requires similar knowledge and skills.

Considering the SDDS model [20], our construct *knowledge of what to assess* regarding experimentation competences included the assessment of experimentation competences regarding the three phases hypothesis formation, design of an experiment, and analysis of data. The formation of questions and performance of experiments that are part of some models or conceptualizations of experimentation competences cf. [64] were not considered.

In the assessment tasks, we worked with descriptions of biology classroom scenarios close to reality. Instead of this, videos of students experimenting in the classroom could measure pre-service teachers' *knowledge of what to assess* closer to reality. More comprehensive tasks would reduce the number of tasks required. At the same time, it could increase the quality of information gained regarding the knowledge measured. However, a more realistic (and more complex) assessment situation could divert the focus from the experimentation competences, which has carefully been weighed against a more focused assessment situation with reduced complexity, as applied in our study.

5. Conclusions

Knowledge of what to assess regarding experimentation competences could be modeled and measured reliably and validly. The analyzed data comprised assessments of 495 pre-service teachers of seven German federal states and 18 universities. The database included pre-service biology teachers of different semesters, different study programs, and different school types. Thus, the following conclusions are drawn more generally for biology teacher education. Further studies could shed light on certain pre-service biology teacher subsamples more specifically.

We worked out criteria for the assessment of experimentation competences regarding hypothesis formation, design of an experiment, and analysis of data according to the SDDS model [20]. With this approach, we gained knowledge for evidence-based biology teacher education in the field of teaching experimentation competences. The assessment tasks regarding these experimental phases of the developed instrument and the scenario format can—with adaptations based on the evidence given in our study—be used for teacher education designing teaching and learning environments to foster teaching experimentation competences in pre-service teachers. Thus, using the seven scenarios and not exceeding 90 minutes in testing time was an adequate approach.

Our study gave insights into pre-service teachers' strengths and weaknesses in the assessment of experimentation competences. The difficulties could be explained about the tasks. Assessing student conceptions as well as correct student solutions, turned out to be more difficult than assessing incorrect student solutions most of the time. The only exceptions we found concern the planning of a standardized and accurate performance and measurement. Comparably few pre-service teachers mastered these requirements. The results suggest that even more attention could be paid in teacher education on student conceptions to enable relevant assessments to be able to foster student learning systematically. Moreover, the relevance of knowing and understanding PCK relevant scientific terms for precise assessments should be highlighted in biology teacher education.

Our study on *knowledge of what to assess* focused on one of the four areas of assessment knowledge and skills defined by Abell and Siegel [16]. Further research could examine the other areas of knowledge about the assessment of experimentation competences as well as the relationships among the knowledge

areas. Moreover, an examination of the relation of assessment literacy and the ability to plan lessons under consideration of students' experimentation competences, which is closely linked to "knowledge of assessment interpretation and action-taking", can give insights for improving teaching experimental lessons [16] (p. 215). For this purpose, we have an overlapping sample with a parallel study on teaching competences for experimental lessons, with one focus on the ability to plan lessons [65] that needs to be analyzed in the future.

The 1D model makes sense regarding the interdependent complex assessment of experimentation competences. Nevertheless, having a closer look at (i) the three phases of experimentation that have to be assessed and (ii) by grouping the items by the specific challenges that have to be overcome provides more in-depth insights into pre-service teachers' strengths and weaknesses. Thereby, the study clearly shows what teacher education already tackles to a good extent and what could be more addressed in the future to bring forward pre-service teachers' *knowledge of what to assess*: This helps to reflect and further develop current practices in biology teacher education in the field of improving student experimentation competences. At the same time, it motivates further research to improve biology teacher education to overcome assessment challenges and to foster assessment literacy.

Author Contributions: Conceptualization, C.J. and S.B.; Data curation, C.J.; Formal analysis, C.J.; Funding acquisition, S.B. and M.H.; Investigation, C.J., S.B., M.H. and C.H.C.; Methodology, C.J., S.B. and C.H.C.; Project administration, S.B.; Resources, S.B.; Supervision, S.B. and M.H.; Validation, C.J. and S.B.; Visualization, C.J. and S.B.; Writing—original draft, C.J.; Writing—review & editing, S.B. and M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research project was funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01PK11014B.

Acknowledgments: We thank the KoKoHs team, all pre-service biology teachers who participated in our study and the university staff who contributed to conduct the investigation.

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

Appendix A

Table A1. Item content and scoring, items in order of decreasing difficulty per scoring category (H: Hypothesis formation, D: Design and performance of the experiment, A: Analysis of data).

Item	Item 6	Item 4	Item 14	Item 15	Item 16	Item 10	Item 11	Item 13
Phase	D	H	D	A	A	D	A	
Criterion	Student conception: *Engineering mode*	*Untestable hypothesis*	*Inaccurate measurement procedure*	*Incorrect data analysis*	Student conception: *Confirmation bias*	*Lack of standardization*	*Imprecise data analysis*	*A conclusion without a link to the hypothesis*
The specific requirement for a full credit for the item	Naming and explaining				Explaining	Naming	Naming	
	the student conception "engineering mode"	that the hypothesis is untestable	that the measurement procedure is inaccurate	that the data analysis is incorrect	the student conception "confirmation bias"	that several aspects lack standardization	that not all findings are taken account of	that the conclusion is not linked to the hypothesis
Score 2	The criterion is named and explained				The criterion is explained or named completely		The criterion is named precisely	
Score 1	The criterion is named or for Item 4 and 6 explained				The criterion is explained or named in parts		The criterion is named imprecisely	
Score 0	The criterion is neither named nor explained							

Table A1. *Cont.*

Item	Item 17	Item 18	Item 19	Item 20
Phase	H		D	A
Criterion	*Untestable hypothesis*	*Unfounded hypothesis*	*Unsystematic variation of variables*	*A conclusion without a link to the hypothesis*
The specific requirement for a full credit for the item	Correcting the given insufficient student answer by stating			
	a testable hypothesis	a founded hypothesis	a systematic variation of variables	a conclusion that supports the hypothesis
Score 2	Student answer is corrected completely			
Score 1	Student answer is corrected in parts	/	/	Student answer is corrected in parts
Score 0	Student answer is not corrected			

Item	Item 3	Item 5	Item 2	Item 1	Item 8	Item 9	Item 7	Item 12
Phase	H		H		D			A
Criterion	*Founded hypothesis*	*Unfounded hypothesis*	*Testable hypothesis*	*Incomprehensive hypothesis formation*	*Systematic variation of variables*	*Imprecise design of experiment*	*Unsystematic variation of variables*	*Missing error analysis*
The specific requirement for a full credit for the item	Naming		Naming or describing					
	that the hypothesis is founded	that the hypothesis is unfounded	that the hypothesis is testable	that the hypothesis formation is incomprehensive	that the variation of variables is systematic	that the design of the experiment is imprecise	that the variation of variables is unsystematic	that error analysis is missing
Score 2	The criterion is named		The criterion is named or described					
Score 1	/							
Score 0	The criterion is neither named nor described							

References

1. Black, P.; Wiliam, D. Assessment and Classroom Learning. *Assess. Educ.* **1998**, *5*, 7–74. [CrossRef]
2. Abell, S.K.; Volkmann, M.J. *Seamless Assessment in Science: A Guide for Elementary and Middle School Teachers*; Heinemann: Portsmouth, NH, USA, 2006; ISBN 978-0-325-00769-4.
3. Hattie, J.; Jaeger, R. Assessment and Classroom Learning: A deductive approach. *Assess. Educ.* **1998**, *5*, 111–122. [CrossRef]
4. Kultusministerkonferenz, K.M.K. (Ed.) *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss*; Wolters Kluwer Deutschland GmbH: München, Germany, 2005.
5. National Research Council. *National Science Education Standards*; The National Academies Press: Washington, DC, USA, 1996; ISBN 978-0-309-05326-6.
6. Hammann, M. Kompetenzentwicklungsmodelle: Merkmale und ihre Bedeutung—Dargestellt anhand von Kompetenzen beim Experimentieren. *MNU* **2004**, *57*, 196–203.
7. Hammann, M.; Phan, T.T.H.; Ehmer, M.; Bayrhuber, H. Fehlerfrei Experimentieren. *MNU* **2006**, *59*, 292–299.
8. Xu, Y.; Brown, G.T.L. Teacher assessment literacy in practice: A reconceptualization. *Teach. Teach. Educ.* **2016**, *58*, 149–162. [CrossRef]
9. Klieme, E.; Avenarius, H.; Blum, W.; Döbrich, P.; Gruber, H.; Prenzel, M.; Reiss, K.; Riquarts, K.; Rost, J.; Tenorth, H.-E.; et al. *The Development of National Education Standards: An Expertise*; Bundesministerium für Bildung und Forschung: Berlin, Germany, 2004.
10. Winterton, J.; Delamare-Le Deist, F.; Stringfellow, E. *Typology of Knowledge, Skills, and Competences: Clarification of the Concept and Prototype*; Office for Official Publications of the European Communities: Luxembourg, 2006; ISBN 92-896-0427-1.
11. Méhaut, P.; Winch, C. The European Qualification Framework: Skills, Competences or Knowledge? *Eur. Educ. Res. J.* **2012**, *11*, 369–381. [CrossRef]
12. der Kultusministerkonferenz, B. (Ed.) Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung. Available online: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf (accessed on 16 March 2020).
13. Stiggins, R. Assessment Literacy. *Phi Delta Kappan* **1991**, *72*, 534–539.
14. Dübbelde, G. Diagnostische Kompetenzen angehender Biologie-Lehrkräfte im Bereich der naturwissenschaftlichen Erkenntnisgewinnung. Ph.D. Thesis, Universität Kassel, Kassel, Germany, 2013. Available online: https://kobra.uni-kassel.de/handle/123456789/2013122044701 (accessed on 16 March 2020).
15. Magnusson, S.; Krajcik, J.; Borko, H. Nature, Sources and Development of Pedagogical Content Knowledge for Science Teaching. In *Examining Pedagogical Content Knowledge*; Gess-Newsome, J., Lederman, N.G., Eds.; Springer: Dordrecht, The Netherlands, 1999; pp. 95–132. ISBN 978-0-7923-5903-6.
16. Abell, S.K.; Siegel, M.A. Assessment Literacy: What Science Teachers Need to Know and Be Able to Do. In *The Professional Knowledge Base of Science Teaching*; Corrigan, D., Dillon, J., Gunstone, R., Eds.; Springer: Dordrecht, The Netherlands, 2011; pp. 205–221. ISBN 978-90-481-3926-2.
17. OECD. PISA for Development Science Framework. In *PISA for Development Assessment and Analytical Framework: Reading, Mathematics and Science*; OECD Publishing: Paris, France, 2018; pp. 71–97. [CrossRef]
18. Schulz, A.; Wirtz, M.; Starauschek, E. Das Experiment in den Naturwissenschaften. In *Experimentieren im Mathematisch-Naturwissenschaftlichen Unterricht*; Rieß, W., Wirtz, M., Barzel, B., Schulz, A., Eds.; Waxmann: Münster, Germany, 2012; pp. 15–18.
19. Klautke, S. Ist das Experimentieren im Biologieunterricht noch zeitgemäß? *MNU* **1997**, *50*, 323–329.
20. Klahr, D. *Exploring Science: The Cognition and Development of Discovery Processes*; The MIT Press: Cambridge, MA, USA, 2000.
21. Li, J.; Klahr, D. The Psychology of Scientific Thinking: Implications for Science Teaching and Learning. In *Teaching Science in the 21st Century*; Rhoton, J., Shane, P., Eds.; NSTA Press: Arlington, VA, USA, 2006; pp. 307–328.
22. Wellnitz, N. *Kompetenzstruktur und -Niveaus von Methoden Naturwissenschaftlicher Erkenntnisgewinnung*; Logos: Berlin, Germany, 2012.

23. Ehmer, M. Förderung von kognitiven Fähigkeiten beim Experimentieren im Biologieunterricht der 6. Klasse: Eine Untersuchung zur Wirksamkeit von methodischem, epistemologischem und negativem Wissen. Ph.D. Thesis, Christian-Albrechts-Universität Kiel, Kiel, Germany, 2008. Available online: https://macau.uni-kiel.de/servlets/MCRFileNodeServlet/dissertation_derivate_00002469/diss_ehmer.pdf (accessed on 16 March 2020).
24. Koslowski, B. *Theory and Evidence: The Development of Scientific Reasoning*; The MIT Press: Cambridge, MA, USA, 1996.
25. Mayer, J.; Ziemek, H.-P. Offenes Experimentieren: Forschendes Lernen im Biologieunterricht. *Unterr. Biol.* **2006**, *317*, 4–12.
26. Krüger, D. Bezaubernde Biologie—Mit Hypothesen der Lösung auf der Spur. *MNU* **2009**, *62*, 41–46.
27. Schauble, L.; Klopfer, E.; Raghaven, K. Students' Transition from an Engineering Model to a Science Model of Experimentation. *J. Res. Sci. Teach.* **1991**, *9*, 859–882. [CrossRef]
28. Köhler, K. Welche fachgemäßen Arbeitsweisen werden im Biologieunterricht eingesetzt? In *Biologie Didaktik. Praxishandbuch für die Sekundarstufe I und II*; Spörhase-Eichmann, U., Ruppert, W., Eds.; Cornelsen: Berlin, Germany, 2004; pp. 146–159.
29. Straube, P. *Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-) Studierenden im Fach Physik*; Logos: Berlin, Germany, 2016.
30. Bouffard-Bouchard, T.; Parent, S.; Larivee, S. Influence of Self-Efficacy on Self-Regulation and Performance among Junior and Senior High-School Age Students. *Int. J. Behav. Dev.* **1991**, *14*, 153–164. [CrossRef]
31. Tschannen-Moran, M.; Woolfolk Hoy, A.; Hoy, W.K. Teacher Efficacy: Its Meaning and Measure. *Rev. Educ. Res.* **1998**, *68*, 202–248. [CrossRef]
32. Saklofske, D.; Michaluk, B.; Randhawa, B. Teachers' Efficacy and Teaching Behaviors. *Psychol. Rep.* **1988**, *63*, 407–414. [CrossRef]
33. Grube, C.R. Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung: Untersuchung der Struktur und Entwicklung des wissenschaftlichen Denkens bei Schülerinnen und Schülern der Sekundarstufe I. Ph.D. Thesis, Universität Kassel, Kassel, Germany, 2010. Available online: https://kobra.uni-kassel.de/handle/123456789/2011041537247 (accessed on 16 March 2020).
34. Hammann, M.; Phan, T.T.H.; Bayrhuber, H. Experimentieren als Problemlösen: Lässt sich das SDDS-Modell nutzen, um unterschiedliche Dimensionen beim Experimentieren zu messen? *Z. Erzieh.* **2007**, *8*, 33–49.
35. Krüger, D.; Upmeier zu Belzen, A.; Nordmeier, V.; Tiemann, R.; Hartmann, S.; Mathesius, S.; Stiller, J.; Straube, P. Kooperation der Projekte Ko-WADiS und ExMo. Unpublished.
36. Bögeholz, S.; Joachim, C.; Hasse, S.; Hammann, M. Kompetenzen von (angehenden) Biologielehrkräften zur Beurteilung von Experimentierkompetenzen. *Unterrichtswissenschaft* **2016**, *44*, 40–54.
37. List, M.K. Testbearbeitungsverhalten in Leistungstests: Modellierung von Testabbruch und Leistungsabfall. Ph.D. Thesis, Christian-Albrechts-Universität Kiel, Kiel, Germany, 2018. Available online: https://macau.uni-kiel.de/servlets/MCRFileNodeServlet/dissertation_derivate_00007735/diss_mk_list_testbearbeitungsverhalten_in_leistungstests.pdf (accessed on 16 March 2020).
38. Kultusministerium, N. (Ed.) Kerncurriculum für das Gymnasium Schuljahrgänge 5 -10: Naturwissenschaften. 2007. Available online: http://db2.nibis.de/1db/cuvo/datei/kc_gym_nws_07_nib.pdf (accessed on 16 March 2020).
39. Mayring, P. *Qualitative Inhaltsanalyse: Grundlagen und Techniken*; Beltz: Weinheim, Germany, 2010.
40. Donner, A.; Rotondi, M.A. Sample Size Requirements for Interval Estimation of the Kappa Statistic for Interobserver Agreement Studies with a Binary Outcome and Multiple Raters. *Int. J. Biostat.* **2010**, *6*. [CrossRef]
41. De Swert, K. Calculating Inter-Coder Reliability in Media Content Analysis Using Krippendorff's Alpha. Available online: https://www.polcomm.org/wp-content/uploads/ICR01022012.pdf (accessed on 16 March 2020).
42. Mahler, H. Selbstwirksamkeitserwartungen angehender Biologielehrkräfte—Entwicklung eines Messinstrumentes. Master's Thesis, Georg-August-Universität Göttingen, Göttingen, Germany, 2014. Unpublished.
43. Masters, G.N. A Rasch model for partial credit scoring. *Psychometrika* **1982**, *47*, 149–174. [CrossRef]
44. Wu, M.L.; Adams, R.J.; Wilson, M.R.; Haldane, S.A. *ACER ConQuest Version 2.0: Generalised Item Response Modelling Software*; Australian Council for Educational Research: Camberwell, Victoria, Australia, 2007.

45. Boone, W.J.; Staver, J.R.; Yale, M.S. *Rasch Analysis in the Human Sciences*; Springer: Dordrecht, The Netherlands, 2014.
46. Bond, T.G.; Fox, C.M. *Applying the Rasch Model*; Routledge: New York, NY, USA, 2015.
47. OECD. *PISA 2006 Technical Report*; OECD: Paris, France, 2009.
48. Pohl, S.; Carstensen, C.H. *NEPS Technical Report—Scaling the Data of the Competence Tests (NEPS Working Paper No. 14)*; Otto-Friedrich-Universität, Nationales Bildungspanel: Bamberg, Germany, 2012.
49. Schecker, H.; Parchmann, I. Modellierung naturwissenschaftlicher Kompetenz. *ZfDN* **2006**, *12*, 45–66.
50. Velten, S.; Nitzschke, A.; Nickolaus, R.; Walker, F. Die Fachkompetenzstruktur von Technikern für Elektrotechnik und Einflussfaktoren auf ihre Kompetenzentwicklung. *J. Technol. Educ.* **2018**, *6*, 201–222.
51. Kultusministerium, N. (Ed.) Kerncurriculum für das Gymnasium Schuljahrgänge 5–10: Naturwissenschaften. 2015. Available online: https://db2.nibis.de/1db/cuvo/datei/nw_gym_si_kc_druck.pdf (accessed on 16 March 2020).
52. Völzke, K.; Arnold, J.; Kremer, K. Denken und Verstehen beim naturwissenschaftlichen Problemlösen. Eine explorative Studie. *Z. Interpret. Schul Unterr.* **2013**, *2*, 58–86. [CrossRef]
53. Chen, Z.; Klahr, D. All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. *Child Dev.* **1999**, *70*, 1098–1120. [CrossRef] [PubMed]
54. University of Washington. Understanding Item Analyses. Available online: https://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/ (accessed on 16 March 2020).
55. Baack, K.; Steinert, K. *Natura 7/8 Biologie für Gymnasien, Niedersachsen*; Klett: Stuttgart, Germany, 2015.
56. Hammann, M. Experimentieren. In *Biologie-Methodik. Handbuch für die Sekundarstufe I und II*; Spörhase, U., Ruppert, W., Eds.; Cornelsen: Berlin, Germany, 2014; pp. 102–106.
57. Bayrhuber, H.; Hammann, M. (Eds.) *Linder Biologie: Abi-Aufgabentrainer, Wissen Anwenden und Kompetenzen Einüben*; Schroedel: Braunschweig, Germany, 2013.
58. Bühner, M. *Einführung in die Test- und Fragebogenkonstruktion*; Pearson: Hallbergmoos, Germany, 2011.
59. Stecher, B.M.; Klein, S.P. The Cost of Science Performance Assessments in Large-Scale Testing Programs. *Educ. Eval. Policy Anal.* **1997**, *10*, 1–14. [CrossRef]
60. Shavelson, R.F. *Measuring College Learning Responsibly: Accountability in a New Era*; Stanford University Press: Stanford, CA, USA, 2009.
61. Bundesministerium für Bildung und Forschung. Qualitätsoffensive Lehrerbildung. Available online: https://www.qualitaetsoffensive-lehrerbildung.de/de/fachwissenschaften-fachdidaktik-und-bildungswissenschaften-1803.html (accessed on 10 February 2020).
62. Eggert, S.; Bögeholz, S. Students' Use of Decision-Making Strategies With Regard to Socioscientific Issues: An Application of the Rasch Partial Credit Model. *Sci. Educ.* **2010**, *94*, 230–258. [CrossRef]
63. Arnold, J.; Kremer, K.; Mayer, J. Wissenschaftliches Denken beim Experimentieren – Kompetenzdiagnose in der Sekundarstufe II. *Erkenn. Biol.* **2012**, *11*, 7–20.
64. Mayer, J.; Grube, C.; Möller, A. Kompetenzmodell naturwissenschaftlicher Erkenntnisgewinnung. In *Lehr- und Lernforschung in der Biologiedidaktik (Band 3)*; Harms, U., Sandmann, A., Eds.; StudienVerlag: Innsbruck, Austria, 2008; pp. 63–78.
65. Hasse, S.; Joachim, C.; Bögeholz, S.; Hammann, M. Assessing teaching and assessment competences of biology teacher trainees: Lessons from item development. *Int. J. Educ. Math. Sci. Technol.* **2014**, *2*, 191–205. [CrossRef]