education sciences

MDPI

Article

Construction and Evaluation of an Instrument to Measure Content Knowledge in Biology: The CK-IBI

Jörg Großschedl [1,*], Daniela Mahler [2] and Ute Harms [2]

[1] Institute for Biology Education; University of Cologne, Herbert-Lewin-Straße 10, 50931 Köln, Germany
[2] Leibniz Institute for Science and Mathematics Education (IPN) at Kiel University; Olshausenstraße 62, 24118 Kiel, Germany; mahler@ipn.uni-kiel.de (D.M.); harms@ipn.uni-kiel.de (U.H.)
* Correspondence: j.grossschedl@uni-koeln.de; Tel.: +49-221-470-7375

Received: 13 July 2018; Accepted: 5 September 2018; Published: 11 September 2018

Abstract: The teaching process is well described as an interaction between teacher, student, and content. Thus, it seems obvious that teachers must know the content to help students to learn it. Instruments have been developed to measure teachers' content knowledge (CK) in biology, but few of them have been provided to the scientific community. Furthermore, most of them have a topic-specific approach, so there is a need for a more comprehensive measure. In efforts to meet this need we have developed an instrument called the CK in biology inventory (CK-IBI), which has a broader scope than previously published instruments and covers knowledge of five biological disciplines (i.e., ecology, evolution, genetics and microbiology, morphology, and physiology). More than 700 pre-service biology teachers were enrolled to participate in tests to assess the instrument's objectivity, reliability, and validity in two cross-sectional evaluations. Item and scale analyses as well as validity checks indicate that the final version of the CK-IBI (37 items; Cronbach's $\alpha = 0.83$) can be scored objectively, is unidimensional, reliable, and validly measures pre-service biology teachers' CK. As the instrument was used in a German context, it has been translated into English to enable its scrutiny and use by international communities.

Keywords: content knowledge; instrument development; assessment; validity; pre-service biology teachers

1. Introduction

1.1. Construction and Evaluation of an Instrument to Measure Content Knowledge in Biology

The teaching process is well described as an interaction between teacher, student, and subject-matter (i.e., the content), because as noted by Ball [1], "Teachers cannot help children learn things they themselves do not understand" (p. 5) [2–4]. Furthermore, policy documents regard teachers' content knowledge (CK) as a significant predictor of students' learning [5,6]. However, verification of this view and analysis of associated factors are hampered by a conspicuous absence of empirical evidence in the literature [7–10]. A few scholars have found indications that teachers' CK has considerable effects on students' achievement [11,12]. However, several reviews and meta-analyses indicate that it has little or even no predictive power for achievement and learning [3,13]. Furthermore, few studies have provided clues regarding factors that could influence the strength of the relationships between teachers' CK and students' achievement and learning. Notably, a review by Darling-Hammond [14] suggests that CK fosters teaching effectiveness up to a certain threshold level, beyond which its influence tapers off. Thus, if the average CK of a particular sample exceeds the threshold level, a possible influence of teachers' CK would be masked. If so, the range of test subjects' CK would influence the degree

to which variations in it affect students' learning, and the instruments used to assess it must cover appropriate ranges of CK in order to detect possible relationships.

While a direct effect of CK on students' learning is rarely observed, CK may have indirect effects by influencing various relevant activities and attributes of teachers, for example, their lesson planning, teaching behavior, and motivational orientations. Accordingly, several authors have found that CK improves student participation by enhancing lesson planning and teaching behavior [13,15]. Furthermore, Carlsen [16,17] found that knowledgeable teachers ask fewer questions than their less knowledgeable colleagues, but elicit more questions from their students. Moreover, knowledgeable teachers seem to be able to provide diversification rather than uniformity [18,19]. In an illustrative examination of how much teachers' CK affects their lesson planning and simulated teaching, Hashweh [2] asked biology and physics teachers to plan lessons using both biology and physics textbooks. So, each participant acted as an expert in one subject and a novice in another. Hashweh [2] found that experts made many modifications. They restructured the chapter structure of the textbook, ignored an irrelevant theme, and made important additions or deleted details, whereas novice teachers adhered closely to the content and structure of the textbook. Concerning simulated teaching, expert teachers were able to actively involve the students, elicit more sophisticated questions, and detect student preconceptions, whereas novice teachers focused on recall, reinforced preconceptions, or incorrectly criticized correct answers.

Regarding motivational orientations of teachers, CK reportedly increases the preference for teaching a particular topic [20] and the self-efficacy beliefs for teaching [21]. Self-efficacy beliefs describe a person's conviction that he or she can successfully execute a certain behavior required to produce desired outcomes [22]. As individuals' abilities affect their willingness to initiate coping behavior and expend effort on a particular task [22], teachers' self-efficacy should be positively related to their likelihood to implement instructional innovations, use adequate teaching strategies, and encourage students' autonomy [23,24]. Similarly, self-efficacy beliefs are predictors of students' learning, according to several studies [25–27].

A related construct, pedagogical content knowledge (PCK), seems to influence students' learning more directly [28]. Pedagogical content knowledge was initially described as an "amalgam of content and pedagogy" (p. 8) [29] forming the type of knowledge that makes subject matter comprehensible for students. Scholars assume that CK is a prerequisite for PCK development (e.g., [30,31]), and corroborative research shows that teacher education programs with a content focus improve the PCK of pre-service teachers who have had few formal opportunities to develop PCK [28,32].

Given the clear importance of CK, and uncertainties regarding the factors influencing its relationships with students' achievements and learning, there is a clear need for objective, reliable and sensitive instruments to measure teachers' CK. Accordingly, various authors have attempted to develop instruments for assessing the CK of pre-service and in-service teachers of science subjects and mathematics (e.g., [28,32–37]). However, few of these instruments have been made available to the public, which clearly restricts the interpretability, credibility, and comparability of the results. Thus, some scholars have called for greater transparency regarding the genesis and collection of the data, for example by publishing the instruments [38,39]. This is also important for addressing major problems in psychological and related sciences associated with the reproducibility of empirical findings [40].

Hence, major aims of the project "Measuring the professional knowledge of pre-service mathematics and science teachers" (German acronym: KiL) were to address such problems. Specific aims of part of the project reported in this paper were to develop and disseminate an instrument for assessing pre-service biology teachers' CK. The resulting instrument, called the CK in biology inventory (CK-IBI), presented and discussed here, is based on international findings and normative settings of the German standards for teacher education [6]. The standards describe particular contents of teacher education in rather abstract terms, so curricula of 16 representative German universities were subjected to detailed analysis in efforts to ensure the instrument's curricular validity [41,42]. The instrument has

been developed for application in Germany and has been translated for an English speaking audience. Although the instrument covers all of the current basic biological disciplines, further studies should be conducted to clarify whether possible language- or culture-dependent differences between countries restrict the validity of the translated CK-IBI.

1.2. Conceptualization of CK

Cochran and Jones [43] distinguish four elements of CK sensu lato. One is CK sensu stricto, knowledge of the major facts and concepts of a particular domain, e.g., microbiology. Another is substantive knowledge, which covers the domain's explanatory structures or paradigms. Shulman [29] describes this type of knowledge as knowledge of "the variety of ways in which the basic concepts and principles of the discipline (i.e., the domain. "Discipline" and "domain" are used interchangeably in this article.) are organized to incorporate its facts" (p. 9). In addition, Cochran and Jones [43] distinguish knowledge of how validity or invalidity is established within the domain (i.e., syntactic knowledge) and the teachers' beliefs about the subject matter. When designing the CK-IBI we judged that incorporating items intended to probe the third and fourth elements would make it too extensive. Thus, our instrument was (and is) intended to cover CK (sensu stricto) and substantive knowledge. For convenience, these two elements are jointly referred to as CK in this article, in accordance with the CK conceptualization presented by Hashweh [2], who defines it as knowledge of the content and its organization, "such as knowledge of discipline conceptual schemes, and more specific knowledge, such as knowledge of details of a particular topic" (p. 110). This knowledge is characterized by detailed topic knowledge, knowledge of other discipline concepts, knowledge of higher order principles, and knowledge of ways of connecting the topic to other entities in the discipline [2].

1.3. Assessment of CK

In previous research both qualitative and quantitative approaches have been applied for assessing teachers' CK. Qualitative procedures are valuable for obtaining deep insights into teachers' minds, for example through interviews (e.g., [44,45]) or analyzing lesson plans [45]. However, they cannot be practically applied in surveys of views or abilities of large numbers of subjects, as planned in this study.

In addition to these methods, there are various methods for assessing the structural nature of CK, involving (for example) use of concept mapping tasks [2,46], sorting tasks [2], and diagnostic games-like instruments [46]. In studies involving large samples, quantitative procedures (e.g., paper-and-pencil tests) have proven utility for assessing teachers' CK. In such tests scholars usually distinguish between open-ended items (e.g., [47,48]) and closed-ended items (e.g., [10,49,50]). Open-ended items can provide richer insights into subjects' thinking processes and understanding than closed-ended items, because responses reflect their personal knowledge and logic, rather than the test designers' logic implicitly underlying prefabricated responses in closed-ended items [51]. However, substantial effort is inevitably required from both test subjects and evaluators, who respectively provide and process responses to open-ended items (e.g., [52,53]). Closed-ended items (such as multiple-choice items) are advantageous in this respect as they can be rapidly scored, thereby facilitating assessment of numerous subjects [52,54]. In addition, they minimize subjectivity in evaluations of participants' responses [52,53], and allow far more items to be answered in a given period, thus potentially enabling exploration of broader CK ranges [52,53]. Furthermore, although some researchers claim that open-ended items have greater validity [28,51], other authors have found no differences in their validity in some cases [55]. A highly relevant finding for this study is that suitably formulated open-ended and closed-ended items can provide highly similar indications of participants' skills, knowledge, and abilities (e.g., [53,56,57]). Thus, for the study presented here we developed paper-and-pencil tests including both closed- and open-ended items to evaluate a broad range of biology teachers' CK in practical test times and acquire information about each individual participant's logic.

1.4. Hypotheses

Two data collections were conducted in two consecutive years (hereafter Evaluation 1 and 2, described below) to evaluate the validity of the CK-IBI. Both data collections included tests of the instrument with large samples of pre-service biology teachers, analyses of the acquired data (item analyses, scale analyses, and validity tests) and consequent adjustment of the instrument. We identified a variety of criteria and constructs, stated hypotheses referring to the relationship between these criteria/constructs and CK-IBI scores and tested hypotheses to investigate the criterion and construct validity of CK-IBI scores. Hereafter, we justify the selection of criteria/constructs and describe the corresponding hypotheses.

- *Teacher education program*. Pre-service teachers choose between two secondary teacher education programs (both require the study of two teaching subjects) which provide a teaching certificate for schools qualifying their students for an academic (grade 5–12 [or 13]; academic track) or nonacademic career (grade 5–9 [or 10]; nonacademic track). There are strong indications that the type of teacher education program pre-service teachers attend influences their performance [37,58,59], partly due to associated variations in the number of subject-related courses they take [60]. Pre-service teachers of the academic track reportedly perform better in CK tests than their nonacademic colleagues [48,61,62]. Thus, we hypothesize that CK scores for participants of the academic track would be higher than those of their colleagues of the nonacademic track.
- *Period of time spent in higher education*. During their 3.5 to 5 years of higher education, pre-service teachers in Germany get lectures to develop professional knowledge, that is, CK, PCK, and pedagogical/psychological knowledge (PPK). Beyond that, they have instructional practice at schools, lasting up to five months in total [63]. Research confirms that pre-service teachers' professional knowledge arise during the course of higher education [61,64]. As higher education in Germany is structured in semesters—the semester is one of the two periods of time that a year at university is divided into—we expect a positive correlation between CK scores and semester (Since there is no standard order of CK contents across universities in Germany, for the sake of simplicity this hypothesis assumes that pre-service teachers are best prepared in the end of higher education, although content thought in the first semester may have been forgotten in the end).
- *Academic success*. The high school grade point average (GPA) is one of the most important criteria for selecting higher education candidates in Germany [65]. There are varying opinions regarding what it measures, e.g., cognitive abilities [66,67] or academic achievement [68]. A meta-analysis by Baron-Boldt [69] showed that GPA is a valid predictor of academic success (finding a correlation between GPA and academic success in university with $r = 0.46$), and other authors regard GPA as one of the best available predictors of academic success [70,71]. Thus, unsurprisingly given the association between GPA and CK, several authors have found moderate correlations between GPA and measured CK of pre-service teachers of physics (e.g., [48]) and mathematics [72,73]. Therefore, we expected to find negative correlations between GPA and CK subscale scores.
- *Cognitive abilities*. Inferences about peoples' cognitive abilities are commonly derived from their formal reasoning abilities. This construct—defined as the "basic intellectual processes of manipulating abstractions, rules, generalizations, and logical relationships" [74] (p. 583)—is a viable predictor of learning progress and performance according to various studies (e.g., [75]). Three sub-constructs of formal reasoning ability can be distinguished: verbal, nonverbal figural, and numerical reasoning (the abilities to solve text-based, geometric, and quantitative problems, respectively) [76]. As verbal abilities seem most relevant for a primarily language-based instrument, we expect CK scores to be positively related to verbal reasoning abilities, but not to the other subscales.
- *Knowledge of the nature of science (NOS)*. Knowledge of NOS refers to individuals' conceptions of the values and assumptions underlying scientific understanding and methodology: "an individual's

beliefs concerning whether or not scientific knowledge is amoral, tentative, empirically based, a product of human creativity, or parsimonious reflect [sic] that individual's conception of the nature of science" [77] (p. 331). As knowledge of NOS is sometimes assumed to be an integral facet of CK [32], we expected to find a positive correlation between CK and knowledge of NOS scores.

- *PCK.* Research has shown that biology teachers' CK and PCK are highly correlated but distinct domains of knowledge [78], in accordance with findings regarding the knowledge of teachers of various other subjects, physics, and mathematics for example [48,79,80]. We expect CK scores and PCK scores to be highly correlated, but CK and PCK to be empirically separable constructs.
- *PPK.* Pedagogical knowledge was initially defined as knowledge of the "broad principles and strategies of classroom management and organization" [81] (p. 8), which is independent of the subject matter. Tamir [82] extended this definition by identifying four elements of pedagogical knowledge: knowledge of "instructional strategies for teaching", "students' understanding", "classroom management", and "assessment". Voss [83] recently further extended the boundaries of pedagogical knowledge into PPK, by including psychological aspects related to the classroom and heterogeneity of individual students. Großschedl and colleagues [32] found that pre-service biology teachers' CK and PPK are moderately correlated domains of knowledge, while other researchers have found a substantial correlation among a sample of pre-service physics teachers and a weak correlation among a sample of pre-service mathematics teachers [33]. We hypothesize that CK and PPK scores would be positively correlated, but expect CK and PPK to be empirically separable constructs.
- *Opportunities to learn.* It is well known that the curriculum of educational institutions as well as the intensity of CK, PCK, and PPK contents in teacher education influence students' performance [3]. Thus, CK, PCK, and PPK contents considered in participants' previous teacher education were captured as indicators of their opportunities to learn (in addition to type of teacher education program and number of semesters). Regarding the contents considered in previous teacher education as indicators of the realized curriculum, we expect to find a positive correlation between coverage of CK contents in previous teacher education and the participants' CK scores. Given that PCK is defined as an "amalgam of content and pedagogy" [81], we also expect to detect positive (but weaker) correlations between their CK scores and the PCK/PPK contents considered, due to reciprocal effects between PCK and CK/PPK.
- *Interest.* Individual interest is defined as a relatively enduring preference for particular topics, subject areas, or activities [84]. Pohlmann and Möller [85] have shown that "subject-specific interest" is a positive predictor of pre-service teachers' self-efficacy, working commitment, and task orientation (i.e., pursuit of increasing competence), all of which are positively related to academic performance [86–90]. Thus, there are positive correlations between interest and achievement, with published r values ranging between 0.05–0.26 according to a review by Fishman and Pasanella [91]. A meta-analysis by Schiefele et al. [92] corroborated these results, finding mean correlations between the two constructs of $r = 0.31$, averaged over various subject areas, and $r = 0.16$ in the subject area of biology. As interest is reportedly a highly content-specific motivational characteristic [84,93], we expect CK-IBI scores to be significantly related to interest in the subject, but not correlated, or even negatively correlated, to interest in pedagogy/psychology and interest in the pedagogy of subject matter.
- *Self-concept.* Shavelson [94] defined self-concept as "a person's perception of himself" (p. 411), arising from his set of attitudes, beliefs, and knowledge about his personal characteristics and attributes [95,96]. The general self-concept can be subdivided into academic and non-academic self-concepts. A large body of research on academic self-concept has revealed positive relations between students' academic self-concept and their performance (e.g., [97,98]). Recently, Paulick et al. [99] showed that pre-service teachers' academic self-concept is empirically separable into CK-, PCK-, and PPK-related components. As self-concept and achievement have reciprocal

effects [100,101], we expect to find a stronger positive correlation between CK scores and CK self-concept than correlations between CK scores and PCK/PPK self-concepts.

2. Evaluation 1

In this phase of the study a preliminary version of the CK-IBI instrument was evaluated, as previously described for an instrument designed to measure PCK (the PCK-IBI) by Großschedl and colleagues [8]. This version included items probing CK related to five biological disciplines (i.e., ecology, evolution, genetics and microbiology, morphology, and physiology) forming five subscales. The following sections describe the sample and test methods applied, the relationships between CK and other constructs and criteria used for validation purposes. Results of item and scale analyses are then presented, and finally the validity of the subscales was checked by correlational analyses, which considered teacher education program, the period of time spent in university studies, high school grade point average (GPA), and cognitive abilities as predictors of subscale scores. After describing the respective criteria/constructs, we make unambiguous statement about its hypothesized relationship towards CK.

2.1. Materials and Methods

2.1.1. Sample and Procedure

The sample used in Evaluation 1 consisted of 263 German pre-service biology teachers enrolled in 11 universities throughout Germany (66.9% academic track, 33.1% nonacademic track). The participants' average age was 22.8 years (SD = 2.5; 21.1% male): 22.6 years (SD = 2.2; range = 19–32 years; 25.5%) in the academic track group and 23.1 years (SD = 2.8, range = 20–34 years; 14.0% male) in the nonacademic track group. On average, they had attended 4.8 semesters in higher education (SD = 2.4), those in the academic track group 5.1 on average (SD = 2.5; range 2–12 semesters) and those in the nonacademic track group 4.1 on average (SD = 2.3; range 1–10 semesters). Slightly more than a third (36%) studied biology and a second science subject (32.4 and 41.1% of those in the academic and nonacademic tracks, respectively), while the rest studied a second non-science subject.

The tests lasted four hours, including two 15-minute breaks, conducted in lecture halls of the participants' respective universities. The breaks were implemented, as research has shown that continuous testing without breaks increases subjective fatigue [102]. However, even testing periods lasting four hours seem to have no negative effect on test performance [103], and therefore should not bias our results. Before the first break, they filled in a demographic questionnaire and addressed questions in an instrument designed to measure cognitive abilities (described below). After this break, three booklets were randomly distributed to them, including items designed to probe CK of ecology and physiology (booklet A), genetics, microbiology, and evolution (booklet B), and morphology (booklet C). Other instruments that are not considered here were also included. After completing their assignments participants received incentive of 40 Euros (approximately US$ 54 at the time of the survey).

2.1.2. Operationalization of CK as a Dependent Variable

Content knowledge was assessed using 72 items (nine short answer items, 63 closed-ended items) focusing on the participants' knowledge of ecology (24 items), evolution (11 items), genetics and microbiology (11 items), morphology (10 items), and physiology (16 items) (Figure 1). The items were developed and judged by professors of the respective biological disciplines, with support of a 22-page manual clarifying the theoretical conceptualization of CK. The manual provided examples of items, together with essential information regarding item formats and construction, and listed relevant contents for teacher education. As already mentioned, the selection of csontents was based on an in-depth analysis of the curricula of 16 representative German universities and the German national teacher education standards [6]. For each item a coding scheme was provided by the developers.

In attempts to ensure face validity, all items were systematically judged and revised by the authors and other researchers. Items were dichotomously scored (0 = wrong answer vs. 1 = correct answer) or polytomously scored (0 = wrong answer; 1 = answer half correct [partial credit]; 2 = correct answer).

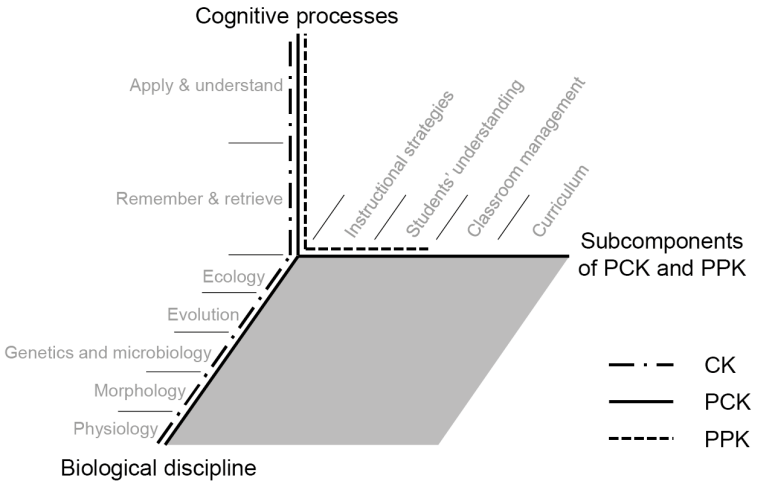

Figure 1. Three-dimensional model of professional knowledge used for item development in the KiL project.

2.1.3. Independent Variables

Teacher education program. Participant's type of teacher education program was captured by a single multiple-choice item ("Which type of school do you aspire to teach in?"), with seven answer alternatives, each indicating a particular school type of the academic track (code = 1) or nonacademic track (code = 0).

Cognitive abilities. We used three subscales of the cognitive abilities test by Heller and Perleth [104] to assess verbal reasoning abilities (20 items; $\alpha = 0.58$, $M = 12.49$, SD = 2.87), nonverbal figural reasoning abilities (25 items; $\alpha = 0.77$, $M = 17.60$, SD = 4.15) and numerical reasoning abilities (20 items; $\alpha = 0.83$, $M = 15.13$, SD = 3.88).

Academic success. Grade point average was captured using a single item, with GPA ranging from 1 (good performance) to 4 (poor performance).

2.2. Results

2.2.1. Statistical Item Analyses

We aimed to cover each biological discipline using the eight best performing items, providing they had discrimination indices >0.15 and item difficulties in the range 0.20 to 0.80. However, after removing 39 items that did not meet these criteria (see Supplemental Material A), the only subscale for which eight items remained was ecology. So, for the other subscales we used all the remaining items. We then averaged scores for the retained items associated with each subscale to obtain measures of CK of ecology (eight items; $\alpha = 0.56$, $M = 4.98$, SD = 2.16), evolution (six items; $\alpha = 0.48$, $M = 6.15$, SD = 2.37), genetics and microbiology (seven items; $\alpha = 0.68$, $M = 3.96$, SD = 2.21), morphology (five items; $\alpha = 0.40$, $M = 2.00$, SD = 1.52), and physiology (seven items; $\alpha = 0.59$, $M = 3.76$, SD = 2.15).

2.2.2. Criterion Validity

We calculated product moment correlations between our subscales and three criteria (participants' type of teacher education program, number of semesters, and GPA) to assess criterion validity of our instrument (Table 1). We expected and observed positive correlations between our subscales ($r = 0.16$ to 0.43) and the type of teacher education program, indicating that participants aspiring to a teaching career in the academic track outperformed participants of the nonacademic track. Furthermore, we

expected positive correlations between our subscales and number of semesters, which was observed for all subscales (r = 0.11 to 0.27) except ecology (r = −0.06). Finally, we expected and corroborated our hypothesis that students' GPA is negatively related to subscale scores (r = −0.11 to −0.38). The results widely support the criterion validity of the CK measure.

Table 1. Product moment correlations between content knowledge (CK) subscales and predictors in Evaluation 1.

Subject Matter	Teacher Education Program	Semester	Grade Point Average (GPA)	Verbal Reasoning	Nonverbal Figural Reasoning	Numerical Reasoning
Ecology (n = 89)	0.35 ***	−0.06	−0.23 *	0.28 **	−0.13	−0.14
Evolution (n = 85)	0.43 ***	0.16	−0.38 ***	0.27 *	0.09	0.27 *
Genetics and microbiology (n = 85)	0.32 **	0.27 *	−0.29 **	0.14	−0.05	0.17
Morphology (n = 89)	0.16	0.11	−0.21 *	0.11	0.22 *	0.12
Physiology (n = 89)	0.21 *	0.13	−0.11	0.26*	−0.10	−0.06

Note. Teacher education program (0 = nonacademic track, 1 = academic track); * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

2.2.3. Construct Validity

As expected and shown in Table 1, correlational analysis applied to assess our constructs' validity detected positive correlations between the CK subscale scores and verbal reasoning abilities (r = 0.11 to 0.28). Correlations between CK subscale scores and both nonverbal figural (r = −0.13 to 0.22) and numerical reasoning abilities (r = −0.14 to 0.27) were lower and even negative in some cases. However, none of the correlations were strong, indicating that our subscales do not merely reflect formal reasoning ability.

3. Evaluation 2

In this phase of the study we aimed to evaluate the revised version of the CK instrument, and identify a parsimonious and efficient set of items. To enable robust assessment of the instrument's validity and more refined predictions of participants' CK performance, the questionnaire was extended to capture not only participants' teacher education program and number of semesters but also their opportunities to learn CK, performance in other domains of professional knowledge, individual interest, and self-concept.

3.1. Materials and Methods

3.1.1. Sample and Procedure

A total of 432 pre-service biology teachers enrolled in 12 universities throughout Germany were recruited: 79.8% and 20.2% aspiring to teaching careers in the academic and nonacademic tracks, respectively. The average age of participants was 23.4 years (SD = 2.8; 19.8% male); 23.3 years (SD = 2.8; range 18–43 years; 21.9% male) in the academic-track group and 23.7 years in the nonacademic-track group (SD = 2.82; range 19–35; 11.5% male). The participants had attended 5.9 semesters in higher education on average (SD = 2.8); those in the academic-track group 5.8 semesters on average (SD = 2.8; range 2–17 semesters), and those in the nonacademic-track group 6.0 semesters on average (SD = 2.9; range 1–14 semesters). Almost half (43.9%) studied biology and a second science subject (46.4% in the academic track group, 34.5% in the nonacademic track group), while the others studied biology and a second non-science subject.

The tests in this evaluation were administered in the same manner as those in Evaluation 1 in terms of duration (including breaks), local conditions, and monetary reimbursement. Before the break participants filled out a questionnaire consisting of questions about learning opportunities, interest, and self-concept. After the break, they completed a questionnaire that included the CK instrument and others designed to measure CK, NOS, and PPK.

3.1.2. Operationalization of PCK as a Dependent Variable

A revised form of the CK instrument with modifications based on results of Evaluation 1 was tested in Evaluation 2 (see Supplemental Material B). This version included 40 items (nine short answer items and 31 open-ended items), in sets of eight covering each biological discipline. Of the items used in the version tested in Evaluation 1, 33 were retained unchanged and 39 were discarded because of item interdependencies or failure to meet difficulty or discrimination criteria. The other seven were redeveloped following the developmental procedure applied in Evaluation 1.

3.1.3. Independent Variables

Opportunities to learn. Of the three measures used to assess CK, PCK, and PPK contents in the participants' previous education, two were applied to measure content-related knowledge (CK and PCK), by inviting responses to statements about them ("Please tick how intensively the following CK/PCK-related contents have been covered in your university studies to date.") on a 5-point Likert-scale (1 = Not Covered, 2 = Rarely Covered, 3 = Moderately Covered, 4 = Excellently Covered). Cronbach's α values for CK and PCK items were 0.86 (15 items; M = 42.89, SD = 0.33) and 0.92 (nine items, M = 24.25, SD = 6.74), respectively. One measure focused on PPK-related content ("Have you attended a university course that addresses this content?"), using a 2-point scale (0 = I have not attended; 1 = I have attended). This measure consisted of 33 items (α = 0.91, M = 14.71, SD = 7.66).

PCK. PCK was assessed using 36 items of the PCK-IBI instrument [8]. This includes 16 open-ended items, six short answer items, and eight closed-ended items (six items consist of two or three subitems with different item formats; subitems were merged because of subitem interdependencies). Some probe participants' knowledge of instructional strategies, including both representation of subject matter and responses to specific learning difficulties (here: knowledge of instructional strategies for teaching; 18 items). Others probe knowledge of students' conceptions and preconceptions (here: knowledge of students' understanding; 18 items; see Figure 1). Two items were deleted because they did not meet acceptance criteria (item difficulty 0.22 to 0.84 and discrimination indices > 0.19). After removing these items, we averaged responses, resulting in a measure of PCK (34 items; α = 0.77, M = 27.51, SD = 7.52), with expected a posteriori/plausible value (EAP/PV) and weighted likelihood estimate (WLE) reliability values of 0.66 and 0.72, respectively [8]. The PCK-IBI is a mixture of dichotomously scored (0 = wrong answer vs. 1 = correct answer) and polytomously scored items (0 = wrong answer; 1 = answer half correct [partial credit]; 2 = correct answer). A coding scheme was used to judge the pre-service teachers' answers.

NOS. NOS was measured using 23 Likert-type items (1 = Strongly Disagree; 2 = Disagree; 3 = Uncertain or Not Sure; 4 = Agree More Than Disagree; 5 = Strongly Agree) of the "Student Understanding of Science and Scientific Inquiry (SUSSI)" [105]. As both positively and negatively phrased items were included, the responses were recoded so the scores were all positively correlated with understanding of science and scientific inquiry. The Cronbach's α value for the knowledge of NOS measure was 0.80 (M = 84.06, SD = 9.56).

PPK. PPK was assessed using 67 items (12 open-ended, four short answer and 51 closed-ended items). These items were developed by experienced educational researchers, based on the national teacher education standards of Germany [106] in an analogous manner to the CK and PCK items. They focused on pre-service teachers' knowledge of "instructional strategies for teaching" (14 items), "students' understanding" (19 items), "classroom management" (21 items), and "assessment" (13 items). Twelve items were removed because they had low discrimination power. The Cronbach's alpha value of the resulting scale was 0.92 (55 items; M = 82.60, SD = 29.38), indicating good internal consistency.

Interest. We used six items of the Study Interest Questionnaire (SIQ) [107] to measure pre-service biology teachers' interest in the three domains of university studies: the subject (source of CK), pedagogy/psychology (source of PPK), and pedagogy of subject matter (source of PCK; German term Fachdidaktik) [108] (p. 656). The participants were asked to assess the degree to which each of these items (e.g., "Dealing with the contents and issues of this study area is one of my favorite

activities") applies to the three sources of professional knowledge (subject, pedagogy/psychology, and the pedagogy of subject matter) on 4-point scales from "does not apply at all" (1) to "fully applies" (4). We averaged scores for items associated with each subscale, resulting in measures for interest in the subject (α = 0.80, M = 20.38, SD = 3.05), pedagogy/psychology (α = 0.88, M = 14.56, SD = 4.40), and pedagogy of subject matter (α = 0.82, M = 15.56, SD = 3.68).

Self-concept. To measure pre-service teachers' self-concept, we used items of two subscales (professional competence and methodological competence) of the Berlin Evaluation Instrument for self-evaluated student competencies (BEvaKomp) [109]. The instrument measures students' CK-, PCK-, and PPK-related self-concept components. The participants were asked to assess the degree to which each of these items, e.g., "I can see the connections and inconsistencies in the subject area of CK (vs. PCK vs. PPK)", applies to the three domains of professional knowledge (CK, PCK, and PPK) on 4-point scales ranging from "does not apply at all" (1) to "fully applies" (4). Cronbach's α values for CK-, PCK-, and PPK-related self-concept components were 0.80 (M = 27.98, SD = 4.23), 0.89 (M = 24.19, SD = 5.57), and 0.88 (M = 22.51, SD = 5.28), respectively, indicating that the subscales had acceptable internal consistency.

3.1.4. Statistical Analysis

Rasch analysis. In order to obtain unconfounded psychometric measurements the procedures used must capture the same ability or attribute (i.e., latent trait) [110,111]. Hence, operationalized variables used in any instrument must each capture a single latent trait to generate valid overall scores [112]. A number of procedures have been developed to test such "unidimensionality" [113], and a frequently used option is to fit the data to a unidimensional measurement model, e.g., the Rasch model [114]. Discrepancies between modelled and empirical data are expressed using descriptive statistics, such as INFIT and OUTFIT, which are mean-square residuals ranging from 0 to infinity, but with expected values of 1 (if the null hypothesis of unidimensionality is not violated). Indices smaller than 1 indicate higher than expected predictability, which may be due to redundancy in the data (i.e., overfit) [115,116]. In Classical Test Theory low mean-square residuals are regarded as desirable, and according to Item Response Theory "they do no harm", despite indicating some redundancy in responses [116] (p. 370). The cited authors do not present any strict thresholds for acceptable OUTFIT and INFIT values, but claim that the values should ideally be somewhere between 0.5 to 1.5. However, in this study we used a tighter range of acceptability: from 0.8 to 1.2. In addition to calculating OUTFIT and INFIT values for the selected items, we applied t-value-tests to identify values that significantly differed from 1. OUTFIT means outlier-sensitive fit, so OUTFIT statistics are highly sensitive to unexpected observations, that is, responses to items with difficulties far away from persons focal ability (i.e., items that respondents find relatively very easy or very hard). In contrast, INFIT means inlier-sensitive fit, and INFIT statistics are more sensitive to discrepancies in patterns of responses to items targeting the persons' focal ability [115,117].

We used ACER ConQuest software (version 1.0.0.1) [118] to analyze the acquired data. As the measures included dichotomously and polytomously scored items, both person ability and item difficulty were estimated using Masters' [119] partial credit model (PCM). This is an extension of the simple logistic (Rasch) model that enables analysis of cognitive items scored in more than two ordered categories, with differing measurement scales, and can provide estimates of threshold parameters for each item, even if the thresholds vary among items [118,120]. We used the WLE method to estimate Person ability as it is less biased than maximum likelihood estimation, and provides best point estimates of individual ability [121]. The person ability scores were used for further calculations implemented in SPSS®. The measurement accuracy was assessed by calculating EAP/PV and WLE reliability values [118].

Dimensionality analysis of the CK-IBI. Multidimensional Rasch analysis was applied to analyze the dimensionality of the CK-IBI. Since our instrument is intended to cover five biological disciplines, initially a five-dimensional model was fitted to the data, assuming that knowledge of ecology, evolution,

genetics & microbiology, morphology, and physiology form separate dimensions. Then we compared the model's fitting parameters to corresponding parameters of a one-dimensional model, which implicitly assumes that CK scores reflect a single latent trait. To identify the model providing the best fit to the data we calculated deviance factors (inversely reflecting the degree to which the data fit underlying assumptions). To assess the significance of differences between the models' deviance factors we applied χ^2-tests, in which the degrees of freedom depend on the number of estimated parameters [122].

In addition, we compared the five- and one-dimensional models using Akaike's Information Criterion (AIC = deviance + 2 np; [123]) and Bayes' Information Criterion (BIC = deviance + [lnN] 2 np; [124]). These criteria do not allow significance tests, but offer the possibility to take into account models' parsimony. Generally, the lower the coefficient, the better the fit between the model and the data [124, 125]. The AIC is superior when the number of possible response patterns greatly exceeds the sample size, while the BIC is superior in opposite cases [126].

Analysis of the CK, PCK, and PPK measures' factor structure. We used confirmatory factor analysis (CFA), which can be applied to examine the relationship between any set of observed variables (e.g., set of items) and set of continuous non-observed variables, to analyze the construct validity of our CK, PCK, and PPK measures. Bentler [122] recommends a minimum ratio of 5:1 between sample size and number of free parameters, but we could not meet this recommendation due to the high number of items. However, using scores for all of the items would lead to a ratio of nearly 1:1, so we used scores for parcels of items associated with subscales as manifest indicators (e.g., giving 15 rather than 126 factor loadings) of latent variables [80,127].

We expected a three-factor model including single latent traits for CK, PCK, and PPK to provide adequate explanation of the variation in responses to our measures. We allocated items associated with each of the three factors to five parcels, thereby reducing the number of factor loadings from 126 to 15 and obtaining a sample size to number of free parameters ratio of 9:1. To estimate the parcel scores we used the WLE (Weighted Maximum Likelihood Estimate) method, which provides best point estimates [121]. As the participants were enrolled in 12 universities, we used "Type = complex" when conducting the CFA to consider the nested structure of the data, with factor variance set to 1 to fix the metric of the latent variables. Distributions of our parcel scores did not meet normality requirements, so we applied full information maximum likelihood estimation with robust standard errors in MPlus 5.21 [128].

Differential item functioning (DIF). A test instrument should always be "fair" to all of the test subjects, in the sense of functioning equivalently towards all groups. So, the difficulty of all the included items should be solely governed by the construct the instrument is intended to measure, and not influenced by any irrelevant or extraneous factors. Otherwise, representatives of different groups with the same ability in terms of that construct may have differing probabilities of answering an item correctly. This is referred to as differential item functioning (DIF) [129–131]. Hence, as any group of subjects may include representative of numerous subgroups (differing, for instance, in race, gender, age, etc.), it is important to identify factors that may be most relevant for testing an instrument. We selected two factors: the participants' track (academic versus nonacademic) and second teaching subject (science (physics or chemistry) versus a non-science subject).

Several techniques have been developed to detect DIF [129,131]. We chose to use a model, based on item response theory, implemented in ACER ConQuest software (version 1.0.0.1) [118]. The model included interactions between item difficulty and both factors (track and second teaching subject, respectively), in order to capture variations in responses due to differences in item difficulties among the corresponding subgroups (academic versus nonacademic groups, and those taking another science subject versus another non-science subject), on the same scale [132]. The statistical significance of differences between groups in item difficulties depends on sample size, so we used effect sizes suggested by the Educational Testing Service [133]. Following their recommendations, we distinguished three categories: A, B, and C for differences in item difficulty between two groups that

are non-significant (<0.43 logits), significant and indicative of a slight to moderate effect (0.43 to 0.64 logits), and significant but indicative of a moderate to large effect (> 0.64), respectively [130,133,134].

3.2. Results

3.2.1. Statistical Item Analyses

Item analysis was conducted in two steps. First, three items were removed because their item difficulty values were outside the pre-defined acceptable range of 0.20 to 0.80, resulting in a set of 37 items (see Supplemental Material B). Then, the dimensionality of the CK-IBI was investigated by Rasch analysis. We fitted a five-dimensional PCM to the data, each dimension representing one of the discipline-based subscales: ecology, evolution, genetics and microbiology, morphology and physiology. To explore the empirical separability of the subscales, we also fitted a one-dimensional model to the data. Results of the Rasch analysis show that the one-dimensional model fits the data better than the five-dimensional model. The information-based criteria are lower for this model (AIC = 17061.87, BIC = 17285.50) than for the five-dimensional model (AIC = 17146.43, BIC = 17426.99). Furthermore, a χ^2-test shows that the one-dimensional model significantly outperforms the five-dimensional model ($\chi^2(14) = 56.56$, $p < 0.001$). Thus, biology teachers' CK represents a single latent trait. In addition, all items adequately discriminated between persons, having discrimination indices between 0.20 and 0.51 (Supplemental Material A). Finally, we averaged responses, resulting in a measure of CK (37 items; $\alpha = 0.83$, $M = 27.82$, SD = 8.13; seven short answer items; 30 closed-ended items), with EAP/PV and WLE reliability values of 0.74 and 0.75, respectively.

The OUTFIT and INFIT statistics (all 0.8–1.2) showed that all of the items adequately fitted Rasch model expectations, indicating that the CK-IBI provides acceptable unidimensionality (see Supplemental Material A). Rasch analysis locates item difficulty values and person ability scores on an interval scale with a common metric [135]. Hence, the suitability of any instrument for exploring abilities or traits of a given population can be displayed in a Wright map. The map we obtained using this approach shows that the mean and standard deviation of the items' difficulty correspond to the ability of the sample (see Figure 2).

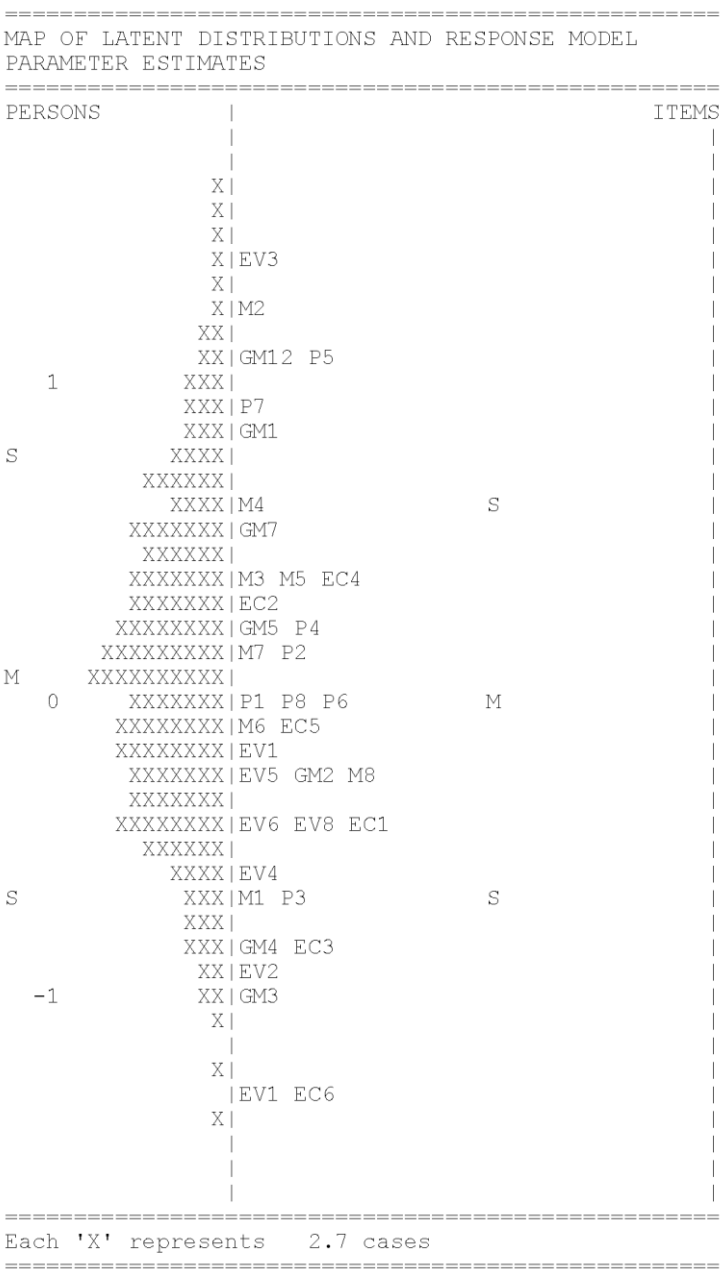

Figure 2. Wright map showing difficulty of the items on the same measurement scale as the ability of the participants. Means (M) and standard deviations (S) for ability and difficulty of the items are shown.

3.2.2. Latent Structure of Professional Knowledge

To investigate the construct validity of the professional knowledge measures, we investigated whether professional knowledge is represented by three distinguishable domains: CK, PCK, and PPK [8]. For this purpose, we subjected data obtained from 430 pre-service biology teachers enrolled in teacher education programs at 12 German universities to CFA. We hypothesized that a three-factor model would adequately explain pre-service teachers' responses to our measures. To test this hypothesis, we examined the model's goodness of fit to the acquired data using several approaches. A χ^2-test showed there was no significant deviation between observed and predicted covariance matrices ($\chi^2(87) = 109.15$, $p = 0.06$), indicating good model fit. We also calculated three statistics recommended for judging model fits in one-time analyses: the comparative fit index (CFI), Tucker–Lewis fit index (TLI), and root-mean-square error of approximation (RMSEA) [136]. Both CFI and TLI compare the model of interest (here the three factor model) with a "baseline" model in which latent variables are

assumed to be uncorrelated. The CFI and TLI values of our model are 0.98, comfortably exceeding the general threshold (0.95) for acceptable model fit [136,137]. At 0.02, the RMSEA (square root of the average of the covariance residuals) was also substantially lower than the general threshold for acceptability, 0.05 [138]. However, we detected comparatively strong latent correlations between CK and PCK ($r = 0.82$), and between CK and PPK ($r = 0.51$, $p < 0.001$ in both cases), indicating that high performance in CK is accompanied by high performance in PCK and PPK. As all of the applied tests indicated that our model has good fit to the acquired data, no post-hoc modifications were applied.

3.2.3. Criterion Validity

Product moment correlation coefficients of two criteria selected to assess criterion validity—"type of teaching program" (track) and "number of semesters"—show that both are significantly and positive correlated with CK scores ($r = 0.29$ and 0.26, respectively; $p < 0.001$ in both cases). These results strongly support the measure's criterion validity.

3.2.4. Construct Validity

Construct validity was checked by correlating CK scores with four categories of measures, capturing: (1) opportunities to learn, (2) performance in further domains of professional knowledge, (3) individual interest, and (4) self-concept. First, CK scores were correlated with subscale scores reflecting how much typical CK, PCK, and PPK contents were considered in participants' previous teacher education. The results show that CK scores are more strongly correlated with opportunities to develop CK ($r = 0.26$, $p < 0.001$) than with opportunities to develop PCK ($r = 0.11$, $p < 0.05$) and PPK ($r = 0.07$, $p = 0.07$), suggesting that our instrument reflects CK more strongly than PCK or PPK (Table 2). Second, correlations between CK scores and performance in other domains of professional knowledge were examined. As expected, we found significant and positive correlations between CK scores and knowledge of NOS, PCK and PPK scores ($r = 0.31$ to 0.61, $p < 0.001$ in all cases; Table 2), corroborating the validity of our CK measure. Third, we found a positive correlation between CK scores and interest in CK ($r = 0.14$, $p < 0.01$), but negative correlations between CK scores and interest in both PCK ($r = -0.06$, $p = 0.11$) and PPK ($r = -0.17$, $p < 0.001$) (Table 2). Fourth, we found a stronger correlation between CK scores and CK-related self-concept ($r = 0.28$, $p < 0.001$) than between CK scores and both PCK-related and PPK-related self-concept ($r = 0.12$, $p < 0.01$ and $r = -0.05$, $p = 0.15$, respectively). These findings further corroborate the validity of the CK-IBI.

Table 2. Descriptive statistics for predictor variables and product moment correlations between CK scores and predictors in Evaluation 2 ($n = 431$).

	Predictor Variable	*M*	SD	CK	
				r	*p*
Opportunities to learn	Track	–	–	0.29	<0.001
	Semester	5.87	2.81	0.24	<0.001
	CK	2.86	0.49	0.26	<0.001
	PCK	2.69	0.75	0.11	<0.05
	PPK	0.45	0.23	0.07	0.07
Performance	NOS	3.65	0.42	0.31	<0.001
	PCK	27.51	7.52	0.61	<0.001
	PPK	82.60	29.38	0.42	<0.001
Interest	CK	3.39	0.51	0.14	<0.01
	PCK	2.59	0.61	−0.06	0.11
	PPK	2.43	0.73	−0.17	<0.001
Self-concept	CK	2.85	0.46	0.28	<0.001
	PCK	2.58	0.57	0.12	<0.01
	PPK	2.46	0.50	−0.05	0.15

Note. CK, PCK, and PPK = content knowledge, pedagogical content knowledge, and pedagogical and psychological knowledge, respectively.

3.2.5. DIF

To analyze "teacher education program" and "second teaching subject", DIF was calculated for each item. While the range was 0.01 to 0.86 logits for "teacher education program", it was 0.00 to 0.68 logits for "second teaching subject". Negligible DIF (category A) was found for 35 items with two items showing moderate to large DIF (category C; see Supplemental Material A) for "teacher education program". For "second teaching subject" 33 items for category A, one item for category C (see Supplemental Material A) and three items showing slight moderate DIF (category B) were found.

4. Discussion

The aim of the cross-sectional KiL study (2011–2013) was to develop an objective, reliable, and valid instrument to measure pre-service biology teachers CK. To meet this aim, a paper-and-pencil test format including both closed- and open-ended items was selected and a manual was produced to assist formulation of items in the instrument (called the CK-IBI) in efforts to maximize its objectivity. Then the instrument was evaluated and refined in two developmental phases (designated Evaluations 1 and 2) in which its reliability and validity were tested and the items were adjusted. Our findings indicate that we successfully developed an instrument that objectively, reliably and validly measures pre-service biology teachers' CK in a reasonable amount of time (45–60 min). Research addressing cognitive fatigue during testing has shown that participants are able to work on cognitively demanding tests over suchlike testing periods without negative fatigue-related disturbance [102,103,139]. Ackerman and colleagues [103], for example, found that four hours of continuous testing increase subjective fatigue, but have no negative effect on test performance.

4.1. Evaluation 1

Evaluation 1 focused on a preliminary version of the CK-IBI. According to the results of the statistical item analysis, the preliminary version of the CK-IBI was revised to provide parsimonious and efficient subscales. The analyses broadly confirm the criterion validity of the CK-IBI. The results show that—as expected and reported in other studies [32,48,61,62]—pre-service biology teachers from the academic track group outperform their fellow students from the nonacademic group. This indicates that the CK-IBI can discriminate between different groups of prospective biology teachers. Moreover, the GPA score, a construct which is strongly related to academic achievement [69–71] and CK performance [48,72,73] was highly related to the CK scores. However, surprisingly we found no significant correlation between participants' number of semesters in higher education and CK scores related to most of the subscales of the CK-IBI. Only their "microbiology and genetics" CK scores were positively related to the number of semesters, indicating that performance in this content area increases with time spent in the teacher education program. A possible explanation for this unexpected result is that the universities that our participants attend deal with specific content areas at different stages of their teacher education. Moreover, most of the single subscales in the preliminary version were not satisfactorily reliable, and an exception was "microbiology and genetics" (which correlated with the number of semesters as hypothesized). As expected, we found a positive relationship between verbal reasoning abilities and CK, together with weaker relationships (some negative) between the other subscales of the cognitive abilities test and CK, supporting the construct validity of the CK-IBI.

4.2. Evaluation 2

The preliminary item set was revised and several items were developed to improve the CK-IBI according to results of Evaluation 1. Rasch analysis-based comparison of a one-dimensional model to a model treating CK as consisting of the five dimensions analogous to the targeted content-areas confirmed that the revised CK-IBI reliably measures a single latent trait. This result is important for the interpretation of total scores [110,112]. To assess test fairness, we investigated the degree (if any) of our items' DIF with respect to "type of teacher education program" (academic track versus nonacademic

track) and second teaching subject (science versus non-science). The results showed that two items discriminated against participants who aspire to a career in academic track schools. Both of these items concern typical lower grade topics (the human eye and vertebrate classes). Hence, the DIF may be due to a stronger focus on such topics in teacher education programs that certify pre-service biology teachers for a career in nonacademic track schools than in their academic track-oriented counterparts. In addition, two items concerning anthocyanins and ATP discriminated against participants lacking a second science subject, presumably because these are chemistry-related topics. However, three items concerning typical biological topics (e.g., the digestive tract of ruminants), discriminated against participants taking a second science subject. We have no explanation for this finding. Due to the low amount of items showing DIF, we decided to keep these items in the CK-IBI. However, researchers should consider removing these items when subsamples should be compared concerning the type of teacher education program or second teaching subject.

In further assessments of the revised instrument, as in Evaluation 1, we found expected correlations between CK scores and both type of teacher education program and number of semesters that support its criterion validity. Similarly, we found that CK scores were more strongly correlated to constructs associated with CK than to constructs associated with PCK or PPK, supporting its construct validity. Findings show that opportunities to learn that are highly related to CK, CK-related interest, and CK-related self-concept were all more strongly correlated with pre-service biology teachers' CK than the constructs related to PCK or PPK. Another construct that is highly related to CK is knowledge of NOS (e.g., [32,77]), thus the expected and confirmed strong correlation between these two constructs further support the instrument's validity. Examination of the latent factor structure also revealed that the CK-IBI measures CK as a distinct and separable domain, in accordance with previous findings that CK is separable from other domains of professional knowledge (e.g., [29,78,140–142]).

Although CK, PCK, and PPK are empirically separable, we found a strong correlation between the pre-service biology teachers' CK and PCK scores, and a moderate correlation between their CK and PPK scores. This pattern is consistent with theoretical expectations as PPK is defined as a domain of teachers' professional competence that is not related to a specific content, whereas both CK and PCK are related to the content (e.g., [29,143]). The strong correlation between CK and PCK scores is also consistent with previous findings regarding the professional knowledge of pre-service physics teachers [48] and pre-service mathematics teachers [79], who reported correlations with $r_{latent} = 0.68$ and 0.79, respectively ($p < 0.001$ in both cases). A strong correlation between CK and PCK performance of in-service mathematics teachers has also been found [80]. However, Jüttner and Neuhaus [144] found a weaker correlation between in-service biology teachers' CK and PCK performance ($r_{latent} = 0.22$, $p < 0.01$), possibly because the items they developed and applied to measure PCK included necessary information about the content. Previous analyses of the relationship between CK and PPK have found a moderate correlation [32] or weak correlation [33].

In summary, the results of the two cross-sectional studies support the objectivity, reliability, and validity of the CK-IBI. In addition, as well as the expected correlation between CK scores and CK-related opportunities, we detected correlations between CK scores and both PCK- and PPK-related opportunities to learn. These opportunities to learn also seem to provide opportunities to develop CK. This result does not conflict with the validity of the instrument, but rather confirms that CK, PCK, and PPK are related domains of professional knowledge (e.g., [32,33,48,78,80,144]). Hence, CK, PCK, and PPK should not be regarded as isolated domains, for instance when considering possible strategies to improve teacher education or characteristics of a successful teacher.

4.3. Limitations

Our results confirm the objectivity, reliability, and validity of the CK-IBI. However, the following limitations should be noted and discussed. Pre-service biology teachers' voluntarily participated in both evaluations. Thus, it is possible that the results are biased due to the motivation to participate, as those who participate in a voluntary study may be more likely to be enthusiastic about the topic of

the study (here biological CK) and have higher relevant self-efficacy beliefs. However, the remuneration for participation may have attenuated this bias. A further concern is related to the test fairness of some items. We identified DIF associated with several items and reasons for the DIF in some cases, which may assist users considering whether to retain or remove those items (e.g., DIF of some samples was related to type of teacher education program, and thus would be irrelevant for testing a sample, e.g., consisting solely of prospective academic track teachers).

Due to the assignment of the content areas to different test booklets in study 1, we did not acquire scores for all items from every participant. However, as the aim of this study was to develop a valid instrument, it was essential to use a sufficient number of items and the applied booklet design helped us to do so in an acceptable amount of test time.

4.4. Implications

4.4.1. Implications for Further Research

The results of our study broadly confirm the validity of the CK-IBI. Nevertheless, additional research could further confirm its validity and/or identify ways to enhance it. One potentially informative approach is cross-validation with other instruments based on paper-and-pencil tests, or other methods developed to measure biology teachers' CK, e.g., interviews or concept mapping. Our research design prevented inferences regarding the CK-IBI's predictive validity, so it would be interesting to assess its validity for predicting students' performance, and thus the relevance of its CK measurements for teaching. This would require its application to a sample of in-service teachers, after checking the suitability of the CK-IBI for testing samples of in-service teachers. It should be mentioned that finding no effect of CK on students' performance would not necessarily refute the instrument's validity, and would have potential implications for the importance of CK for successful teaching [7,9,12].

The challenges facing biology teachers continuously change, due (inter alia) to new biological research findings, and consequent incorporation of new topics such as epigenetics or genetic engineering in teacher education programs and school curricula. Thus, the instrument could be extended or amended by adding or changing items in accordance with developments in biology education.

Beyond that, future research should answer the question how to implement the CK-IBI into a survey best. One of the questions our research left open is, whether demographic items should be placed at the beginning of a survey or at the end, i.e., before or after administering the CK-IBI. We placed demographic items first, since research has shown that this increases response rates for demographic items without having negative effects on response rates for the rest of a survey [145]. In contrast, placing demographic items at the end reduces response rates on demographic items without an effect on response rates for the rest. However, we cannot exclude that placing demographic items first leads pre-service teachers to respond differently to the CK-IBI as if they were not primed with their demographic characteristics (e.g., track). This effect is known as *stereotype threat*, which is "... a situational threat that diminishes performance, originating from a negative stereotype about one's own social group" [146] (p. 300). Future research has to clarify, whether negative stereotypes could arise from track membership (academic vs. nonacademic). This would be plausible, as track membership determines the number of subject-related courses pre-service teachers attend [60] and research has shown that pre-service teachers of the academic track outperform their colleagues of the nonacademic track in matters of performance and motivational characteristics [147]. Thus, the difference between pre-service teachers CK-IBI scores observed in our study at least in part could be attributed to the priming of pre-service teachers with the demographic question referring to the teacher education program. In order to avoid activating negative stereotypes about test performance, some researchers [148] recommend to place questions about the demographic group (e.g., track membership) at the end of a survey and to place performance measures at the beginning. However,

unpublished results show that pre-service teachers which identify with teacher-related stereotypes reach higher CK-IBI scores than their counterparts without such stereotypes [149].

4.4.2. Implications for Teacher Education

In addition to confirming the quality of the developed instrument, the results of this study have implications for strategies to improve pre-service biology teacher education. Our findings reveal that CK is a distinct domain of biology teachers' professional knowledge, but related to PCK and PPK. This indicates that acquisition of knowledge in these three domains does not proceed in an isolated manner. This conclusion is supported by our finding that opportunities to learn related to PCK and PPK also provide opportunities for CK development. Therefore, we infer that integration of the different domains of professional knowledge provides beneficial opportunities for their development. Due to the strong relations between CK and PCK, and between PCK and PPK, joint consideration in teacher education seems plausible. However, there are also interactions between CK and PPK during teaching; PPK enables teachers to use available time optimally to promote student learning related to a specific content [143].

4.4.3. Implications for the Further Application of the CK-IBI

The overarching aim of this study was to provide an objective, reliable, and valid instrument to measure pre-service biology teachers' CK. The results of this study indicate that we achieved this goal and that our instrument is ready for application by other researchers in other projects. The CK-IBI was originally developed in German language and was translated to enable its use by an international audience. We believe that it could be applied in several contexts, including those outlined below.

4.5. Applications

4.5.1. Application in Further Research

The CK-IBI could assist efforts to address various research questions related to CK. Notably, its use in a longitudinal study could help to elucidate how biology teachers' CK develops during teacher education and during teachers' occupational life. This would also help to identify specific learning opportunities for biology teachers to develop CK. A further interesting research question concerns the relevance of CK for students' performance. Available empirical indications of its importance are mixed [7,9,12], and application of the CK-IBI could help clarification of this issue. However, as already stated, further validation of the instrument for use with in-service samples would be required before using the CK-IBI in research outside university-level teacher education settings. The CK-IBI could also help cross-validation of other developed instruments. Moreover, both our instrument and the test development procedure (involving curriculum analysis, item development by experts and two rounds of evaluation and adjustment to ensure objectivity, validity, and reliability) could serve as role-models for other test-development studies. We would appreciate the application of the CK-IBI by international groups, as this would provide information about its reliability and validity for non-German samples. The current version of the CK-IBI appears to be a good starting point for adaptation to meet different demands in different educational systems, as illustrated by application of an instrument developed by Jüttner and Neuhaus [34] to a US sample, although its development involved detailed consideration of the curricula of German universities.

4.5.2. Application in University-Level Teacher Education

The CK-IBI could be used by teachers in university-level teacher education courses to measure pre-service biology teachers' CK. Moreover, pre-service biology teachers could use the CK-IBI to test their own CK in order to optimize their learning progress.

4.5.3. Application in School

The CK-IBI is also a useful instrument for in-service teachers to self-evaluate their CK in order to identify their specific needs for CK-related professional development.

4.6. Conclusions

We have developed one of the first (to our knowledge) objective, reliable, and valid instruments for assessing pre-service biology teachers' CK that has been provided to the scientific community. The analysis of the curricula of universities all over Germany, the participation of experts in the item development, and two-phase evaluation of the instrument's quality, helped us in reaching this goal. We are optimistic that the CK-IBI will be used by a wide audience and applied in both various research fields and other contexts. As it covers all the basic biological disciplines we assume that it will also be applicable in studies on teacher education in other countries. The CK-IBI is currently being applied in the framework of a study examining the development of German pre-service biology teachers' CK during university-level teacher education.

Supplementary Materials: The following are available online at http://www.mdpi.com/2227-7102/8/3/145/s1, Supplemental_Material_A: Item statistics, Supplemental_Material_B: CK in biology inventory (CK-IBI).

Author Contributions: Conceptualization, J.G. and U.H.; Methodology, J.G. and D.M.; Validation, J.G., U.H. and D.M.; Formal Analysis, J.G.; Investigation, J.G., D.M. and U.H.; Writing-Original Draft Preparation, J.G. and D.M.; Writing-Review and Editing, J.G.; Visualization, J.G.; Supervision, J.G. and U.H.; Project Administration, U.H.; Funding Acquisition, U.H.

Funding: This research was funded by the German Leibniz Association (grant number SAW-2011-IPN-2).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Ball, S.J. *Politics and Policy Making in Education: Explorations in Policy Sociology*; Routledge: London, UK, 2012.
2. Hashweh, M.Z. Effects of Subject-matter Knowledge in the Teaching of Biology and Physics. *Teach. Teach. Educ.* **1987**, *3*, 109–120. [CrossRef]
3. Hattie, J. *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement*; Routledge: London, UK, 2009.
4. Ferguson, P.; Womack, S.T. The Impact of Subject Matter and Education Coursework on Teaching Performance. *J. Teach. Educ.* **1993**, *44*, 55–63. [CrossRef]
5. American Council on Education. Touching the Future: Final Report: Presidents' Task Force on Teacher Education. Available online: https://www.acenet.edu/news-room/Documents/Touching-the-Future-Final-Report-2002.pdf (accessed on 1 May 2018).
6. Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany. Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung. Available online: http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf (accessed on 1 May 2018).
7. Carlisle, J.F.; Correnti, R.; Phelps, G.; Zeng, J. Exploration of the Contribution of Teachers' Knowledge about Reading to Their Students' Improvement in Reading. *Read. Writ.* **2009**, *22*, 457–486. [CrossRef]
8. Großschedl, J.; Welter, V.; Harms, U. A New Instrument for Measuring Pre-service Biology Teachers' Pedagogical Content Knowledge: The PCK-IBI. *J. Res. Sci. Teach.* (accepted).
9. Lange, K.; Ohle, A.; Kleickmann, T.; Kauertz, A.; Möller, K.; Fischer, H.E. Zur Bedeutung von Fachwissen und fachdidaktischem Wissen für Lernfortschritte von Grundschülerinnen und Grundschülern im naturwissenschaftlichen Sachunterricht. *Zeitschrift für Grundschulforschung* **2015**, *8*, 23–38.
10. Sadler, P.M.; Sonnert, G.; Coyle, H.P.; Cook-Smith, N.; Miller, J.L. The Influence of Teachers' Knowledge on Student Learning in Middle school Physical Science Classrooms. *Am. Educ. Res. J.* **2013**, *50*, 1020–1049. [CrossRef]

11. Heller, J.I.; Daehler, K.R.; Wong, N.; Shinohara, M.; Miratrix, L. Differential Effects of Three Professional Development Models on Teacher Knowledge and Student Achievement in Elementary Science. *J. Res. Sci. Teach.* **2012**, *49*, 333–362. [CrossRef]
12. Ohle, A.; Fischer, H.E.; Kauertz, A. Der Einfluss des physikalischen Fachwissens von Primarstufenlehrkräften auf Unterrichtsgestaltung und Schülerleistung [Primary School Teachers' Content Knowledge in Physics and Its Impact on Teaching and Students' Achievement]. *ZfDN* **2011**, *17*, 357–389.
13. Byrne, C.J. Teacher Knowledge and Teacher Effectiveness: A Literature Review, Theoretical Analysis, and Discussion of Research Strategy. In Proceedings of the 14th Annual Convention of the Northeastern Educational Research Association, Ellenville, NY, USA, October 1983.
14. Darling-Hammond, L. Teacher Quality and Student Achievement: A Review of State Policy Evidence. *Educ. Policy Anal. Arch.* **2000**, *8*, 1–44. [CrossRef]
15. Leinhardt, G.; Smith, D.A. Expertise in Mathematics Instruction: Subject Matter Knowledge. *J. Educ. Psychol.* **1985**, *77*, 247–271. [CrossRef]
16. Carlsen, W.S. Why Do You Ask? The Effects of Science Teacher Subject-matter Knowledge on Teacher Questioning and Classroom Discourse. In Proceedings of the Annual Meeting of the American Educational Research Association, Washington, DC, USA, 20–24 April 1987.
17. Carlsen, W.S. Effects of New Biology Teachers' Subject-matter Knowledge on Curricular Planning. *Sci. Educ.* **1991**, *75*, 631–647. [CrossRef]
18. Dewey, J. *The Sources of a Science of Education*; Horace Liveright: New York, USA, 1929.
19. Hill, H.C.; Rowan, B.; Ball, D.L. Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement. *Am. Educ. Res. J.* **2005**, *42*, 371–406. [CrossRef]
20. Großschedl, J.; Konnemann, C.; Basel, N. Pre-service Biology Teachers' Acceptance of Evolutionary Theory and Their Preference for Its Teaching. *Evol. Educ. Outreach* **2014**, *7*, 1–16. [CrossRef]
21. Riese, J.; Reinhold, P. Empirische Erkenntnisse zur Struktur professionelller Handlungskompetenz angehender Physiklehrkräfte [Empirical Findings Regarding the Structure of Future Physics Teachers' Competence Regarding Professional Action]. *ZfDN* **2010**, *16*, 167–187.
22. Bandura, A. Self-efficacy: Toward a Unifying Theory of Behavioral Change. *Adv. Behav. Res. Ther.* **1978**, *1*, 139–161. [CrossRef]
23. Cousins, J.B.; Walker, C.A. Predictors of Educators' Valuing of Systemic Inquiry in Schools. Available online: https://evaluationcanada.ca/system/files/cjpe-entries/15--0--025.pdf (accessed on 1 May 2018).
24. Guskey, T. Teacher Efficacy, Self-concept, and Attitudes toward the Implementation of Instructional Innovation. *Teach. Teach. Educ.* **1988**, *4*, 63–69. [CrossRef]
25. Bandura, A. Perceived Self-efficacy in Cognitive Development and Functioning. *Educ. Psychol.* **1993**, *28*, 117–148. [CrossRef]
26. Muijs, D.; Reynolds, D. *Effective Teaching: Research and Practice*; Paul Chapman: London, UK, 2001.
27. Tschannen-Moran, M.; Woolfok-Hoy, A.; Hoy, W.K. Teacher Efficacy: Its Meaning and Measure. *Rev. Educ. Res.* **1998**, *68*, 202–248. [CrossRef]
28. Baumert, J.; Kunter, M.; Blum, W.; Brunner, M.; Voss, T.; Jordan, A.; Klusmann, U.; Krauss, S.; Neubrand, M.; Tsai, Y.-M. Teachers' Mathematical Knowledge, Cognitive Activation in the Classroom, and Student Progress. *Am. Educ. Res. J.* **2010**, *47*, 133–180. [CrossRef]
29. Shulman, L.S. Those Who Understand: Knowledge growth in Teaching. *Educ. Res.* **1986**, *15*, 4–14. [CrossRef]
30. Ball, D.L.; Hill, H.C.; Bass, H. Knowing Mathematics for Teaching. Who Knows Mathematics Well Enough to Teach Third Grade, and How Can We Decide? *Am. Educ.* **2005**, *29*, 14–46.
31. Ma, L. *Knowing and Teaching Elementary Mathematics: Teachers' Understanding of Fundamental Mathematics in China and the United States*; Erlbaum: Hillsdale, NJ, USA, 1999.
32. Großschedl, J.; Harms, U.; Kleickmann, T.; Glowinski, I. Preservice Biology Teachers' Professional Knowledge: Structure and Learning Opportunities. *J. Sci. Teach. Educ.* **2015**, *26*, 291–318. [CrossRef]
33. Buchholtz, N.; Kaiser, G.; Blömeke, S. Die Erhebung mathematikdidaktischen Wissens—Konzeptualisierung einer komplexen Domäne [Measuring Pedagogical Content Knowledge in Mathematics—Conceptualizing a Complex Domain]. *J. Math. Didaktik* **2013**, *35*, 101–128. [CrossRef]
34. Jüttner, M.; Neuhaus, B.J. Validation of a Paper-and-pencil Test Instrument Measuring Biology Teachers' Pedagogical Content Knowledge by Using Think-aloud Interviews. *J. Educ. Train. Stud.* **2013**, *1*, 113–125. [CrossRef]

35. Kirschner, S.; Borowski, A.; Fischer, H.E. Das Professionswissen von Physiklehrkräften: Ergebnisse der Hauptstudie. In *Konzepte fachdidaktischer Strukturierung für den Unterricht, Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Oldenburg 2011*; Bernhold, S., Ed.; Lit: Oldenburg, Germany; pp. 209–211.
36. Schmelzing, S.; van Driel, J.H.; Jüttner, M.; Brandenbusch, S.; Sandmann, A.; Neuhaus, B.J. Development, Evaluation, and Validation of a Paper-and-pencil Test for Measuring Two Components of Biology Teachers' Pedagogical Content Knowledge Concerning the "Cardiovascular System". *Int. J. Sci. Math. Educ.* **2013**, *11*, 1369–1390. [CrossRef]
37. Tatto, M.T.; Senk, S. The Mathematics Education of Future Primary and Secondary Teachers: Methods from the Teacher Education and Development Study in Mathematics. *J. Teach. Educ.* **2011**, *62*, 121–137. [CrossRef]
38. Lupia, A.; Alter, G. Data Access and Research Transparency in the Quantitative Tradition. *PS-Polit. Sci. Polit.* **2013**, *47*, 54–59. [CrossRef]
39. Lupia, A.; Elman, C. Openness in Political Science: Data Access and Research Transparency. *PS-Polit. Sci. Polit.* **2014**, *47*, 19–42. [CrossRef]
40. Open Science Collaboration. Estimating the Reproducibility of Psychological Science. *Science* **2015**, *349*, 4716–4718. [CrossRef] [PubMed]
41. Großschedl, J.; Harms, U.; Glowinski, I.; Waldmann, M. Erfassung des Professionswissens angehender Biologielehrkräfte: Das KiL-Projekt. *MNU* **2014**, *67*, 457–462.
42. Mehrens, W.A.; Phillips, S.E. Sensitivity of Item Difficulties to Curricular Validity. *J. Educ. Meas.* **1987**, *24*, 357–370. [CrossRef]
43. Cochran, K.F.; Jones, L.L. The Subject Matter Knowledge of Preservice Science Teachers. In *International Handbook of Science Education*; Fraser, B.J., Tobin, K.G., Eds.; Kluwer Academic Publishers: Dortrecht, The Netherlands, 1998; pp. 707–718.
44. Arzi, H.; White, R. Change in Teachers' Knowledge of Subject Matter: A 17-year Longitudinal Study. *Sci. Educ.* **2008**, *92*, 221–251. [CrossRef]
45. Käpylä, M.; Heikkinen, J.-P.; Asunta, T. Influence of Content Knowledge on Pedagogical Content Knowledge: The Case of Teaching Photosynthesis and Plant Growth. *Int. J. Sci. Educ.* **2009**, *31*, 1395–1415. [CrossRef]
46. Douvdevany, O.; Dreyfus, A.; Jungwirth, E. Diagnostic Instruments for Determining Junior High-school Science Teachers' Understanding of Functional Relationships Within the 'Living Cell'. *Int. J. Sci. Educ.* **1997**, *19*, 593–606. [CrossRef]
47. Guyton, E.; Farokhi, E. Relationships among Academic Performance, Basic Skills, Subject Matter Knowledge, and Teaching Skills of Teacher Education Graduates. *J. Teach. Educ.* **1987**, *38*, 37–42. [CrossRef]
48. Riese, J.; Reinhold, P. Die professionelle Kompetenz angehender Physiklehrkräfte in verschiedenen Ausbildungsformen—Empirische Hinweise für eine Verbesserung des Lehramtsstudiums [The Professional Competencies of Trainee Teachers in Physics in Different Educational Programs—Empirical Findings for the Improvement of Teacher Education Programs]. *ZfE* **2012**, *15*, 111–143.
49. Akyol, G.; Tekkaya, C.; Sungur, S.; Traynor, A. Modeling the Interrelationships among Pre-service science Teachers' Understanding and Acceptance of Evolution, Their Views on Nature of Science and Self-efficacy Beliefs Regarding Teaching Evolution. *J. Sci. Teach. Educ.* **2012**, *23*, 937–957. [CrossRef]
50. Phelps, G.; Schilling, S. Developing Measures of Content Knowledge for Teaching Reading. *Elem. Sch. J.* **2004**, *105*, 31–48.
51. Schoenfeld, A.H. The Complexities of Assessing Teacher knowledge. *Meas. Interdisciplin. Res. Perspect.* **2007**, *5*, 198–204. [CrossRef]
52. Lipton, A.; Huxham, G.J. Comparison of Multiple-choice and Essay Testing in Preclinical Physiology. *Br. J. Med. Educ.* **1970**, *4*, 228–238. [CrossRef] [PubMed]
53. Walstad, W.; Becker, W. Achievement Differences on Multiple-choice and Essay Tests in Economics. *Am. Econ. Rev.* **1994**, *84*, 193–196.
54. Bacon, D.R. Assessing Learning Outcomes: A Comparison of Multiple-choice and Short Answer Questions in a Marketing Context. *J. Mark. Educ.* **2003**, *25*, 31–36. [CrossRef]
55. Bridgeman, B.; Lewis, C. The Relationship of Essay and Multiple-choice Scores with Grades in College Courses. *J. Educ. Meas.* **1994**, *31*, 37–50. [CrossRef]
56. Bennett, R.E.; Rock, D.A.; Wang, M. Equivalence of Free-response and Multiple-choice Items. *J. Educ. Meas.* **1991**, *28*, 77–92. [CrossRef]

57. Wainer, H.; Thissen, D. Combining Multiple-choice and Constructed-response Test Scores: Toward a Marxist Theory of Test Construction. *Appl. Psychol. Meas.* **1993**, *6*, 103–118. [CrossRef]
58. Blömeke, S.; Delaney, S. Assessment of Teacher Knowledge across Countries: A Review of the State of Research. *ZDM-Math. Educ.* **2012**, *44*, 223–247. [CrossRef]
59. Wang, J.; Lin, E. Comparative Studies on US and Chinese Mathematics Learning and the Implications for Standards-based Mathematics Teaching Reform. *Educ. Res.* **2005**, *34*, 3–13. [CrossRef]
60. Blömeke, S.; Kaiser, G.; Döhrmann, M.; Lehmann, R. Mathematisches und mathematikdidaktisches Wissen angehender Sekundarstufen-I-Lehrkräfte im internationalen Vergleich. In *TEDS-M 2008: Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für Die Sekundarstufe I im Internationalen Vergleich*; Blömeke, S., Kaiser, G., Lehmann, R., Eds.; Waxmann: Münster, Germany, 2010; pp. 197–238.
61. Kleickmann, T.; Richter, D.; Kunter, M.; Elsner, J.; Besser, M.; Krauss, S.; Baumert, J. Teachers' Content Knowledge and Pedagogical Content Knowledge: The Role of Structural Differences in Teacher Education. *J. Teach. Educ.* **2013**, *64*, 90–106. [CrossRef]
62. Schmidt, W.H.; Tatto, M.T.; Bankow, K.; Blömeke, S. *The Preparation Gap: Teacher Education for Middle School Mathematics in Six Countries*; MSU Center for Research in Mathematics and Science Education: East Lansing, MI, USA, 2007.
63. Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany. The Education System in the Federal Republic of Germany 2011/2012: A Description of the Responsibilities, Structures and Developments in Education Policy for the Exchange of Information in Europe. Available online: http://www.kmk.org/fileadmin/doc/Dokumentation/Bildungswesen_en_pdfs/teachers.pdf (accessed on 1 May 2018).
64. Großschedl, J.; Neubrand, C.; Kirchner, A.; Oppermann, L.; Basel, N.; Gantner, S. Entwicklung und Validierung eines Testinstruments zur Erfassung des evolutionsbezogenen Professionswissens von Lehramtsstudierenden (ProWiE). *ZfDN* **2015**, *21*, 173–185. [CrossRef]
65. Heine, C.; Briedis, K.; Didi, H.J.; Haase, K.; Trost, G. *Bestandsaufnahme von Auswahl- und Eignungsfeststellungsverfahren beim Hochschulzugang in Deutschland und ausgewählten Ländern*; HIS-Hochschul-Informations-System-GmbH: Hannover, Germany, 2006.
66. Anderson, J.R.; Lebière, C. *The Atomic Components of Thought*; Erlbaum: Mahwah, NJ, USA, 1998.
67. Rindermann, H.; Oubaid, V. Auswahl von Studienanfängern: Vorschläge für ein zuverlässiges Verfahren. *Forschung und Lehre* **1999**, *41*, 589–592.
68. Rohde, T.E.; Thompson, L.A. Predicting Academic Achievement with Cognitive Ability. *Intelligence* **2007**, *35*, 83–92. [CrossRef]
69. Baron-Boldt, J. *Die Validität von Schulabschlussnoten für Die Prognose von Ausbildungs-und Studienerfolg*; Peter Lang: Frankfurt, Germany, 1989.
70. Moser, K. Alternativen zur Abiturnote? Sinn und Unsinn neuer eignungsdiagnostischer Verfahren. *Forschung und Lehre* **2007**, *8*, 474–476.
71. Tarazona, M. Berechtigte Hoffnung auf bessere Studierende durch hochschuleigene Studierendenauswahl? Eine Analyse der Erfahrungen mit Auswahlverfahren in der Hochschulzulassung. *Beiträge zur Hochschulforschung* **2006**, *28*, 68–89.
72. Blömeke, S.; Kaiser, G.; Lehmann, R. *Professionelle Kompetenz Angehender Lehrerinnen und Lehrer: Wissen, Überzeugungen und Lerngelegenheiten Deutscher Mathematikstudierender und—Referendare: Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung*; Waxmann: Münster, Germany, 2008.
73. Kleickmann, T.; Anders, Y. Learning at University. In *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers: Results from the COACTIV Project*; Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., Neubrand, M., Eds.; Springer: New York, USA, 2013; pp. 321–332.
74. Carrol, J.B. *Human Cognitive Abilities: A Survey of Factor-analytic Studies*; Cambridge University Press: Cambridge, UK, 1993.
75. Hattie, J.A.C.; Hansford, B.C. Self Measures and Achievement: Comparing a Traditional Review of Literature with Meta-analysis. *Aust. J. Educ.* **1982**, *26*, 71–75. [CrossRef]
76. Heitmann, P.; Hecht, M.; Schwanewedel, J.; Schipolowski, S. Students' Argumentative Writing Skills in Science and First-language Education: Commonalities and Differences. *Int. J. Sci. Educ.* **2014**, *36*, 3148–3170. [CrossRef]

77. Ledermann, N.G. Students' and Teachers' Conceptions of the Nature of Science: A Review of the Research. *J. Res. Sci. Teach.* **1992**, *29*, 331–359. [CrossRef]
78. Großschedl, J.; Mahler, D.; Kleickmann, T.; Harms, U. Content-related Knowledge of Biology Teachers from Secondary Schools: Structure and Learning Opportunities. *Int. J. Sci. Educ.* **2014**, *36*, 2335–2366. [CrossRef]
79. Blömeke, S.; Suhl, U. Modeling Teacher Competencies: Using Different IRT-scales to Diagnose Strengths and Weaknesses of German Teacher Trainees in an International Comparison. *Z. Erziehungswiss* **2010**, *13*, 473–505. [CrossRef]
80. Krauss, S.; Brunner, M.; Kunter, M.; Baumert, J.; Blum, W.; Neubrand, M.; Jordan, A. Pedagogical Content Knowledge and Content Knowledge of Secondary Mathematics Teacher. *J. Educ. Psychol.* **2008**, *100*, 716–725. [CrossRef]
81. Shulman, L.S. Knowledge and Teaching: Foundations of the New Reform. *Harv. Educ. Rev.* **1987**, *57*, 1–22. [CrossRef]
82. Tamir, P. Subject Matter and Related Pedagogical Knowledge in Teacher Education. *Teach. Teach. Educ.* **1988**, *4*, 99–110. [CrossRef]
83. Voss, T.; Kunter, M.; Baumert, J. Assessing Teacher Candidates' General Pedagogical/Psychological Knowledge: Test Construction and Validation. *J. Educ. Psychol.* **2011**, *103*, 952–969. [CrossRef]
84. Schiefele, U. Interest, Learning, and Motivation. *Educ. Psychol.* **1991**, *26*, 299–323. [CrossRef]
85. Pohlmann, B.; Möller, J. Fragebogen zur Erfassung der Motivation für die Wahl des Lehramtsstudiums (FEMOLA). *Zeitschrift für Pädagogische Psychologie* **2010**, *24*, 73–84. [CrossRef]
86. Carrol, A.; Houghton, S.; Wood, R.; Unsworth, K.; Hattie, J.; Gordon, L.; Bower, J. Self-efficacy and Academic Achievement in Australian High School Students: The Mediating Effects of Academic Aspirations and Delinquency. *J. Adolesc.* **2009**, *32*, 797–817. [CrossRef] [PubMed]
87. Multon, K.D.; Brown, S.D.; Lent, R.W. Relation of self-Efficacy Beliefs to Academic Outcomes: A Meta-analytic Investigation. *J. Couns. Psychol.* **1991**, *38*, 30–38. [CrossRef]
88. Robbins, S.; Lauver, K.; Le, H.; Davis, D.; Langley, R.; Carlstrom, A. Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-analysis. *Psychol. Bull.* **2004**, *130*, 261–288. [CrossRef] [PubMed]
89. Sheard, M. Hardiness Commitment, Gender, and Age Differentiate University Academic Performance. *Br. J. Educ. Psychol.* **2009**, *79*, 189–204. [CrossRef] [PubMed]
90. Sideridis, G.D. Goal Orientation, Academic Achievement, and Depression: Evidence in Favor of a Revised Goal Theory Framework. *J. Educ. Psychol.* **2005**, *97*, 366–375. [CrossRef]
91. Fishman, J.A.; Pasanella, A.K. College Admission-selection Studies. *Rev. Educ. Res.* **1960**, *30*, 298–310. [CrossRef]
92. Schiefele, U.; Krapp, A.; Winteler, A. Interest as a Predictor of Academic Achievement: A Meta-analysis of Research. In *The Role of Interest in Learning and Development*; Renniger, K.A., Hidi, S., Krapp, A., Eds.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1992; pp. 183–212.
93. Lowman, R.L. Interests. In *Corsini Encyclopedia of Psychology*; Wiley: Hobroken, NJ, USA, 2010.
94. Shavelson, R.J.; Hubner, J.J.; Stanton, G.C. Self-concept: Validation of Construct Interpretations. *Rev. Educ. Res.* **1976**, *46*, 407–441. [CrossRef]
95. Ghazvini, S.D. Relationships between Academic Self-concept and Academic Performance in High School Students. *Procedia-Soc. Behav. Sci.* **2011**, *15*, 1034–1039. [CrossRef]
96. Purkey, W.W. *Self-Concept and School Achievement*; Prentice Hall: Englewood Cliffs, NJ, USA, 1970.
97. Arens, A.K.; Yeung, A.S.; Craven, R.G.; Hasselhorn, M. The Twofold Multidimensionality of Academic Self-concept: Domain Specificity and Separation between Competence and Affect Components. *J. Educ. Psychol.* **2011**, *103*, 970–981. [CrossRef]
98. Marsh, H.W.; Martin, A.J. Academic Self-concept and Academic Achievement: RELATIONS and Causal Ordering. *Br. J. Educ. Psychol.* **2011**, *81*, 59–77. [CrossRef] [PubMed]
99. Paulick, I.; Großschedl, J.; Harms, U.; Möller, J. Preservice Teachers' Professional Knowledge and Its Relation to Academic Self-concept. *J. Teach. Educ.* **2016**, *67*, 173–182. [CrossRef]
100. Marsh, H.W.; Craven, R.G. Reciprocal Effects of Self-concept and Performance from a Multidimensional Perspective: Beyond Seductive Pleasure and Unidimensional Perspectives. *Perspect. Psychol. Sci.* **2006**, *1*, 133–163. [CrossRef] [PubMed]
101. Retelsdorf, J.; Köller, O.; Möller, J. Reading Achievement and Reading Self-concept: Testing the Reciprocal Effects Model. *Learn. Instr.* **2014**, *29*, 21–30. [CrossRef]

102. Ackerman, P.L.; Kanfer, R. Test Length and Cognitive Fatigue: An Empirical Examination of Effects on Performance and Test-taker Reactions. *J. Exp. Psychol. Appl.* **2009**, *15*, 163–181. [CrossRef] [PubMed]
103. Ackerman, P.L.; Kanfer, R.; Shapiro, S.W.; Newton, S.; Beier, M.E. Cognitive Fatigue during Testing: An Examination of Trait, Time-on-task, and Strategy Influences. *Hum. Perform.* **2010**, *23*, 381–402. [CrossRef]
104. Heller, K.A.; Perletz, C. *Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement*; Beltz-Testgesellschaft: Göttingen, Germany, 2000.
105. Liang, L.L.; Chen, S.; Chen, X.; Kaya, O.N.; Adams, A.D.; Macklin, M.; Ebenezer, J. Student Understanding of Science and Scientific Inquiry (SUSSI): Revision and Further Validation of an Assessment Instrument. In Proceedings of the Annual Conference of the National Association for Research in Science Teaching (NARST), San Francisco, CA, USA, 3–6 April 2006.
106. Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany. Standards für die Lehrerbildung: Bildungswissenschaften. Available online: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf (accessed on 17 January 2018).
107. Krapp, A.; Schiefele, U.; Wild, K.P.; Winteler, A. The Study Interest Questionnaire (SIQ). *Diagnostica* **1993**, *39*, 335–351.
108. Van Driel, J.H.; Berry, A. The Teacher Education Knowledge Base: Pedagogical Content Knowledge. In *International Encyclopedia of Education*, 3rd ed.; McGraw, B., Peterson, P.L., Baker, E., Eds.; Elsevier: Oxford, UK, 2010; pp. 656–661.
109. Braun, E.; Gusy, B.; Leidner, B.; Hannover, B. The Berlin Evaluation Instrument for Self-evaluated Student Competences (BEvaKomp). *Diagnostica* **2008**, *54*, 30–42. [CrossRef]
110. Hattie, J. Methodology review: Assessing Unidimensionality of Tests and Items. *Appl. Psychol. Meas.* **1985**, *9*, 139–164. [CrossRef]
111. Segars, A. Assessing the Unidimensionality of Measurement: A Paradigm and Illustration within the Context of Information Systems Research. *Omega* **1997**, *25*, 107–121. [CrossRef]
112. Hagell, P. Testing Rating Scale Unidimensionality Using the Principal Component Analysis (PCA)/t-Test Protocol with the Rasch Model: The Primacy of Theory over Statistics. *Open J. Stat.* **2014**, *4*, 456–465. [CrossRef]
113. Tennant, A.; Pallant, J.F. Unidimensionality Matters! (A Tale of Two Smiths?). *Rasch Meas. Trans.* **2006**, *20*, 1048–1051.
114. Wright, B.D.; Linacre, J.M. Observations are Always Ordinal; Measurements, However, Must be Interval. *Arch. Phys. Med. Rehabil.* **1989**, *70*, 857–860. [PubMed]
115. Linacre, J.M. What Do Infit and Outfit, Mean-square and Standardized Mean? *Rasch Meas. Trans.* **2002**, *3*, 878.
116. Wright, B.D.; Linacre, J.M. Reasonable Mean-square Fit Values. *Rasch Meas. Trans.* **1994**, *8*, 370–371.
117. Wright, B.D.; Masters, G.N. Computation of OUTFIT and INFIT Statistics. *Rasch Meas. Trans.* **1990**, *3*, 84–85.
118. Wu, M.L.; Adams, R.J.; Wilson, M.R.; Haldane, S.A. *ACER ConQuest Version 2: Generalised Item Response Modelling Software*; Australian Council for Educational Research: Camberwell, Australian, 2007.
119. Masters, G.N. A Rasch Model for Partial Credit Scoring. *Psychometrika* **1982**, *47*, 149–174. [CrossRef]
120. Wright, B.D.; Mok, M. Rasch Models Overview. *J. Appl. Meas.* **2000**, *1*, 83–106. [PubMed]
121. Warm, T.A. Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika* **1989**, *54*, 427–450. [CrossRef]
122. Bentler, P.M. Comparative Fit Indexes in Structural Models. *Psychol. Bull.* **1990**, *107*, 238–246. [CrossRef] [PubMed]
123. Akaike, H. Likelihood of a Model and Information Criteria. *J. Econ.* **1981**, *16*, 3–14. [CrossRef]
124. Wilson, M.; De Boeck, P.; Carstensen, C. Explanatory Item Response Models: A Brief Introduction. In *Assessment of Competencies in Educational Contexts: State of the Art and Future Prospects*; Hartig, J., Klieme, E., Leutner, D.E., Eds.; Hogrefe and Huber: Göttingen, Germany, 2008; pp. 91–120.
125. Schermelleh-Engel, K.; Mossbrugger, H. Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-fit Measures. *Meth. Psychol. Res. Online* **2003**, *8*, 23–74.
126. Rost, J. *Lehrbuch Testtheorie—Testkonstruktion*, 2nd ed.; Verlag Hans Huber: Bern, Switzerland, 2004.
127. Little, T.D.; Cunningham, W.A.; Shahar, G.; Widaman, K.F. To Parcel or not to Parcel: Exploring the Question, Weighing the Merits. *Struct. Equ. Model.* **2002**, *9*, 151–173. [CrossRef]
128. Muthén, L.K.; Muthén, B.O. *Mplus User's Guide*, 6th ed.; Muthén and Muthén: Los Angeles, CA, USA, 2007.
129. Penfield, R.D.; Lam, T.C.M. Assessing Differential Item Functioning in Performance Assessment: Review and Recommendation. *Educ. Meas.* **2000**, *19*, 5–15. [CrossRef]

130. Penfield, R.D. An Approach for Categorizing DIF in Polytomous Items. *Appl. Psychol. Meas.* **2007**, *20*, 335–355. [CrossRef]
131. Zumbo, B.D. *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and LIKERT-Type (Ordinal) Item Scores*; Directorate of Human Resources Research and Evaluation, Department of National Defense: Ottawa, ON, Canada, 1999.
132. Wetzel, E.; Böhnke, J.R.; Carstensen, C.H.; Ziegler, M.; Ostendorf, F. Do Individual Response Styles Matter? Assessing Differential Item Functioning for Men and Women in the NEO-PI-R. *J. Individ. Differ.* **2013**, *34*, 69–81. [CrossRef]
133. Zieky, M. Practical Questions in the Use of DIF Statistics in Test Development. In *Differential Item Functioning, Holland*; Paul, W., Wainer, H., Eds.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1993; pp. 337–347.
134. Linacre, J.M.; Wright, B.D. Mantel-Haenszel DIF and PROX are Equivalent. *Rasch Meas. Trans.* **1989**, *3*, 52–53.
135. Smith, E.V. Metric Development and Score Reporting in Rasch Measurement. *J. Appl. Meas.* **2000**, *1*, 303–326. [PubMed]
136. Schreiber, J.B.; Stage, F.K.; King, J.; Nora, A.; Barlow, E.A. Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *J. Educ. Res.* **2006**, *99*, 323–337. [CrossRef]
137. Hu, L.; Bentler, P.M. Evaluating Model Fit. In *Structural Equation Modeling: Issues, Concepts, and Applications*; Hoyle, R., Ed.; Sage: Newbury Park, CA, USA, 1995; pp. 76–99.
138. Steiger, J.H. Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivar. Behav. Res.* **1990**, *25*, 173–180. [CrossRef] [PubMed]
139. Thorndike, E.L. Mental Fatigue. *J. Educ. Psychol.* **1911**, *2*, 61–80. [CrossRef]
140. Grossman, P.L. *The Making of a Teacher: Teacher Knowledge and Teacher Education*; Teacher College Press: New York, NY, USA, 1990.
141. Jüttner, M.; Boone, W.; Park, S.; Neuhaus, B.J. Development and Use of a Test Instrument to Measure Biology Teachers' Content Knowledge (CK) and Pedagogical Content Knowledge (PCK). *Educ. Assess. Eval. Account.* **2013**, *25*, 45–67. [CrossRef]
142. 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.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1999; pp. 95–132.
143. Voss, T.; Kunter, M. Teachers' General Pedagogical/Psychological Knowledge. In *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers. Results from the COACTIV Project*; Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., Neubrand, M., Eds.; Springer: New York, NY, USA, 2013; pp. 207–228.
144. Jüttner, M.; Neuhaus, B.J. Development of Items for a Pedagogical Content Knowledge-test Based on Empirical Analysis of Pupils' Errors. *Int. J. Sci. Educ.* **2012**, *34*, 1125–1143. [CrossRef]
145. Teclaw, R.; Price, M.C.; Osatuke, K. Demographic Question Placement: Effect on Item Response Rates and Means of a Veterans Health Administration Survey. *J. Bus. Psychol.* **2012**, *27*, 281–290. [CrossRef]
146. Ihme, T.A.; Möller, J. "He Who Can, Does; He Who Cannot, Teaches?": Stereotype Threat and Preservice Teachers. *J. Educ. Psychol.* **2015**, *107*, 300–308. [CrossRef]
147. Neugebauer. Wer entscheidet sich für ein Lehramtsstudium - und warum? Eine empirische Überprüfung der These von der Negativselektion in den Lehrerberuf (Who Chooses to Study Education—and Why? An Empirical Examination of the Thesis of Negative Selection into the Teaching Profession). *Z. Erziehungswiss* **2013**, *16*, 157–184.
148. Jordan, A.H.; Lovett, B.J. Stereotype Threat and Test Performance: A Primer for School Psychologists. *J. Sch. Psychol.* **2007**, *45*, 45–59. [CrossRef]
149. Hansen, J. *Spielt die Identifikation mit Stereotypen im Studium des Sekundarstufenlehramts Biologie eine Rolle? (Does Identification with Stereotypes in University Studies of Pre-Service Biology Teachers Matter?)*; Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik an der Christian-Albrechts-Universität zu Kiel: Kiel, Germany, 2017.