sustainability

MDPI

Article

Measuring System Competence in Education for Sustainable Development

Nina Roczen [1,*], Frank Fischer [2], Janis Fögele [3], Johannes Hartig [1] and Rainer Mehren [4]

[1] Department of Teacher and Teaching Quality, DIPF | Leibniz Institute for Research and Information in Education, 60323 Frankfurt, Germany; hartig@dipf.de
[2] Department of Geography, Justus Liebig University Gießen, 35390 Gießen, Germany; Frank.Fischer@geogr.uni-giessen.de
[3] Institute of Geography, University of Hildesheim, 31141 Hildesheim, Germany; foegele@uni-hildesheim.de
[4] Institute for Geography Education, University of Münster, 48149 Münster, Germany; rainer.mehren@uni-muenster.de
* Correspondence: roczen@dipf.de; Tel.: +49-69-24708-153

Abstract: This paper presents the development of an instrument for the assessment of system competence in the field of Education for Sustainable Development (ESD). Based on an already existing, more complex model of system competence for the school subject geography, we have developed a test that refers to central themes and principles of ESD using exclusively closed answer formats. Building on the results of cognitive laboratories and expert feedback from various fields, the instrument was (further) developed in an iterative process of feedback and revision. We conducted a quantitative pilot study with N = 366 8th and 9th grade students. The results indicate that the development of our system competence test was successful—the overall test yielded a high reliability and only very few items were not working as intended. Furthermore, the difficulties of the items were appropriate for the ability levels of the students and the results of a confirmatory factor analysis (CFA) suggest that the newly developed test measures system competence with one dimension. As the test is compact, easy to interpret, and yet reliable, it is particularly suitable for monitoring purposes in the field of ESD.

Keywords: Education for Sustainable Development; system competence; assessment; monitoring; indicator; sustainable development goals (SDGs)

check for updates

Citation: Roczen, N.; Fischer, F.; Fögele, J.; Hartig, J.; Mehren, R. Measuring System Competence in Education for Sustainable Development. *Sustainability* **2021**, *13*, 4932. https://doi.org/10.3390/su13094932

Academic Editor: David González-Gómez

Received: 10 March 2021
Accepted: 22 April 2021
Published: 28 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Challenges in the context of sustainability are often characterized by a high degree of factual complexity. This complexity results, among other things, from a multitude of strongly interlinked elements, a dynamic characterized by time-delayed changes or acceleration, non-transparency due to a lack of information, a multitude of sometimes conflicting goals, and the scale levels from local to global [1]. Numerous empirical studies show that dealing with complexity and the search for solutions [2–4] often do not do justice to this complexity and that monocausal thinking prevails instead. This often leads to solutions which later turn out to be the problems of tomorrow—which proves that this statement from the famous 1972 report "The limits of growth" has lost none of its relevance [5]. Instead of proceeding according to the trial and error principle, students need to be instructed to make more cognitive decisions per action, which means running through causes and their causes, or effects and effects of the effects in their mind [6]. In this sense, a so-called system thinker takes superordinate principles of systems into account in a cognitive analysis and mental representation of systems. This principle-led perspective provides a deeper understanding of the internal and external interplay and complexity of systems, which may prevent human interference in such systems having unpredictable and unwanted adverse effects [7].

Consequently, systems thinking is regarded as central to Education for Sustainable Development (ESD). This central role is demonstrated, for example, in a large meta-study on core competencies in the context of sustainability [8], in which these are summarized as "the abilities to recognize and understand relationships; to analyse complex systems; to think of how systems are embedded within different domains and different scales; and to deal with uncertainty" [9] (p. 10). On the basis of an international Delphi study, Rieckmann [10] also identifies systems thinking as crucial for ESD in order to understand central problems of world society and to be able to act in terms of sustainable development. Moreover, numerous central reference documents, such as those of the UN [9], identify systems thinking as a key competence of sustainability.

For some time now, there have been efforts to establish the promotion of ESD at the international level, e.g., with the UN Decade for ESD (2005–2014; [11]), the Global Action Programme on Education for Sustainable Development (2015–2019; [12]) and eventually the current Agenda 2030 [13], in the framework of which a separate sub-goal for ESD was formulated (sub-goal 4.7). With these programs, the need for large-scale monitoring has also increased in order to be able to capture the status of achievement of ESD-related goals. Such international monitoring is associated with various requirements for indicators for the achievement of the various (sub-) goals, for example, the use of the indicators should be easy and time- and cost-efficient [14,15].

For system competence as one of the key competences in the field of ESD, this means that indicators which meet those requirements are needed. However, existing instruments are complex to evaluate; for example, they require experts to code open text answers.

In this article, we present the development process of a competence-based assessment tool, which is suitable for providing data for national and international monitoring of SDG 4.7, but which can also be used to evaluate single measures for the promotion of system competence in the field of ESD.

1.1. The Concept of Systems Thinking in Social Ecology

Systems thinking generally refers to the mental representation of a system extracted from the real world, which can then be used to analyze the complexity of situations and, thus, to achieve an understanding of spatial and temporal relations [16].

In order to develop a competency test that takes into account the complex interactions between natural and social (sub-) systems, we have based our development on the conceptualization of systems thinking of social ecology [17]. In contrast to purely natural-science-based approaches, which consider social influences as external disturbances (and social-science-based approaches, which define natural influences as external disturbances; [18]), the approach of social ecology allows an integrative human-environment-system approach and is, thus, equally applicable to "pure" natural and social systems as well as to human-environment systems [19]. The following sub-dimensions of systems thinking can be distinguished:

1. *System structure*: System structure refers to the knowledge of relevant system elements as well as the interconnection of these elements.
2. *System boundary*: System boundary refers to the ability to perceive a system as a specific sub-part of the world and to distinguish it from its surroundings.
3. *System interaction*: System interaction describes the understanding that several factors can be cause and effect at the same time, for direct and indirect interactions as well as for simple and multiple feedback loops. It also includes an awareness of the openness of self-sustaining (autopoietic) systems.
4. *System dynamics*: System dynamics is characterized by an understanding of the irreversibility of development processes and changes. The initial state in a complex system can usually not be restored in a completely identical way. Furthermore, in addition to linear developmental progressions, exponential and logistic developments must also be considered [20–23].

5. *System emergence*: System emergence refers to the understanding that the interaction of system components at one level can give rise to new properties at the next higher level.
6. *System prognosis*: System prognosis means the ability to derive prognoses while taking into account possible consequences and (complex) effects of the system behavior. This also includes the awareness of the limited predictability of system changes.
7. *System regulation*: System regulation refers to the ability to develop systematic measures in the sense of adaptive management, based on analyses of (complex) effects. Regulation requires both a reduction in complexity and a continuous consideration of system dynamics.

1.2. System Competence

Based on the conceptualization of systems thinking in social ecology described above, system competence can be defined as the abilities and skills to recognize, describe, and model a complex socio-ecological area of reality as a system, to describe and model the system's structure, dynamics and boundaries, and based on this, to make prognoses and take measures for system use and regulation [24]. Based on this conceptualization of systems thinking and on existing theoretically derived models of system competence in the literature [3,20,23,25], a normative competence level model for geography education was developed. This so-called GeoSysko competency model provides the basis for the development of the instrument for monitoring system competence in ESD presented in this paper [24].

The GeoSysko test models system competence in the area of geography two-dimensionally: (1) *System organization and behavior* refers to the ability to recognize, describe, and model a complex socio-ecological area of reality as a system. (2) *System-adequate intention to act* means the ability to make predictions and take actions for system use and regulation based on that modeling. To vary the difficulty of the test items measuring these two sub-competencies, the following difficulty-generating characteristics were manipulated during item development: (a) The number of elements and relations of a system that have to be considered, (b) the degree of interconnectedness of elements that needs to be captured, and (c) the degree to which system-specific properties (e.g., understanding of emergence, limited predictability, etc.) have to be applied. For example, to solve an easy item measuring the dimension "System organization and behavior", only a low number of elements and relations have to be identified and only isolated, monocausal relations have to be recognized. Knowledge of system-specific properties is barely required.

Psychometric testing of the GeoSysko model confirmed the two dimensions. Competence levels were empirically confirmed by regression analyses predicting item difficulties on the basis of the three difficulty-generating characteristics (for detailed information on the competence model and the empirical validation, see [7]).

1.3. Challenges of Measuring System Competence

Starting from the GeoSysko model, we wanted to develop a new competence measure that is not only suitable for ESD across different school subjects but also suitable for large-scale monitoring purposes. In doing so, the following challenges arise.

1.3.1. Challenges of Measuring System Competence in General

A key challenge in measuring system competence is that it is highly correlated with domain-specific knowledge [26]. If a participant has only limited knowledge about the subject, it is difficult for him or her to model a complex system. Therefore, many studies either choose upstream interventions to provide the information necessary to complete the task to every participant (e.g., [27]) or to integrate that information into the test stem (e.g., [7]). The latter also facilitates the international use of such tests in the sense of international SDG monitoring, since less consideration must be given to the country-specific content of school curricula. With the latter method, however, it must be accepted that the developed test also measures reading competence to a considerable extent.

Another issue in the context of validity focuses on ensuring that system competence is adequately represented in the test formats as a construct. Numerous recent studies, therefore, base task development on competence models [7,28–30]. These competency models largely converge in content [24], and are well grounded in systems theory. The formats of the test items also show an influence [31] so that they are usually varied. Test formats that are particularly relevant for measuring system competence are progression or stock/flow diagrams (e.g., for measuring understanding of system behavior) and, most importantly, concept maps (e.g., for measuring understanding of complex system organization).

1.3.2. Challenges in Measuring System Competence in ESD

System competence is considered to be domain-specific, whereas ESD is regarded as a teaching principle and cross-curricular task for many subjects. Existing instruments in the area of system competence typically focus on one or, at most, two subjects. Extensive empirical research is available, particularly in the framework of the subjects biology, geography, geosciences, physics, and chemistry.

In social sciences such as politics or economic education, which are also central to a systemic understanding of sustainability issues, less empirical research is available. The focus on the subject leads to two limitations that undermine the requirements associated with ESD. The focus on individual school subjects implies that contents and topics of the tests are usually taken from a specific subject curriculum. The selected media in the test stem and items also reflect the subject (e.g., climate diagrams in geography).

On the other hand, these tests also represent a conceptual understanding of the subject with a special emphasis on corresponding key concepts (e.g., energy in physics). The interdisciplinary task of ESD, however, is characterized by its own key concepts or principles for dealing with the learning subjects [32]—different aspects of multi-perspectivity play a central role for system competence in the field of ESD: Considering social, ecological, economic, and political or cultural dimensions, systematically switching between local and global approaches, and taking into account the temporal perspective of intergenerational justice [33]. The awareness of conflicting goals resulting from this consideration of different perspectives, the need to deal with non-knowledge due to the complexity involved, and the anticipation of consequences (of consequences...) in the planning of regulatory measures are further principles. At the same time, these principles are associated with challenges that are specific to the development of measurement instruments for large-scale monitoring purposes (see below).

1.3.3. Challenges of System Competence Monitoring in the Field of ESD

Large-scale international educational monitoring places very specific demands on the indicators to be used. For example, they must be “manageable” [34], “easy”, “robust” [14], or “timely” [15]. For an appropriate measurement instrument in this context, this means that it should be as simple as possible, easy to analyze, and economical in its application. However, some of the principles of ESD, such as dealing with high complexity and not-knowing, make the implementation of the general requirements for indicators particularly challenging. As mentioned above, the format of concept maps is particularly suitable to capture system competence [35]. However, the use of usually (partially) open concept maps proves to be difficult. These are usually analyzed by calculating a structure index (e.g., [7]) or by comparison with expert maps (e.g., [36]). Both procedures are very time consuming and require coders who are experienced in terms of system competence. Therefore, they are less suitable for large-scale assessment and educational monitoring. However, the avoidance of these comparatively more elaborate evaluation techniques carries the risk that central components of the construct, such as the handling of conflicting goals or not-knowing, might no longer be adequately reflected.

1.4. Aims for the Test Development

In this project, we aim to develop and psychometrically confirm a test for assessing students' system competence in the context of ESD. In doing so, we have also taken into account the challenges mentioned above. In the development process, we have pursued the following goals:

The test instrument should be applicable both in practice, for example to evaluate the effectiveness of individual ESD measures, but also and in particular for international monitoring in the field of ESD (for example to monitor the progress regarding the Agenda 2030 sub-goal 4.7). Consequently, the test instrument to be developed should be compact and easy to use. Specifically, we aim to achieve this by measuring system competence with only one dimension and by using only closed test formats. In addition, the application of item response theory (IRT) scaling methods allows us to obtain test scores on a common scale for students who complete different subsets of the total test. The corresponding objective is to reliably measure system competence even with a reduced set of items. In order to reliably measure the level of system competence of all individuals in the target sample, we aim to ensure that the difficulty of the test items approximately covers the range of competence levels of 15-year-olds.

As ESD represents an interdisciplinary task [37], the test instruments should be applicable to the entire domain of ESD and not be limited to individual school subject-specific approaches. The scope of the test content should be ensured by drawing on key reference documents such as the sustainable development goals [9]. The central conceptual teaching principles of ESD (e.g., the consideration of different perspectives such as local and global or intergenerational) should also be taken into account.

1.5. Research Questions

In accordance with the objectives described above, we have developed a test for measuring system competence in ESD based on the precursor model for geography. We describe the development process in Section 2. In our pilot study, we administered the instrument to a sample of 15-year-old students (grades 8 and 9). For a psychometric test of the instrument, we investigated the following research questions.

(1) Does the test reliably measure system competence?
(2) Do the difficulties of the test items cover the range of the students' competence levels in grades 8 and 9?
(3) Is it possible to measure system competence with one dimension?

2. Test Development

Step 1: Selection of the Interdisciplinary Topics for the Test Stems

The first step was to select the topics for the test stems. Frameworks and conceptual work in the field of ESD identify typical areas within the human-environment system. For example, the orientation framework for the learning area of global development in the context of ESD states that both the subject-specific and the interdisciplinary topics should "reflect the multidimensionality of the idea of sustainable development", take up "globalization and global development processes", and "at the same time allow for a reference to everyday life and a global view of the world" [33] (p. 96). Common to the various frameworks in the literature is the requirement for those topics to address "as many aspects of the ESD construct as possible and to gain insight into the interconnectedness" [38] (p. 26). Finally, the SDGs established in the framework of Agenda 2030 for sustainable development define seventeen fields of action and topics in the human-environment system. From the comparison of the German "Cross-Curricular Framework in the Context of Education for Sustainable Development" [33], the thematic priorities set in Agenda 21, and the analysis of topics in (inter)national curricula across several subjects [39], the following six particularly typical fields of ESD—which can also be found in the SDGs in a comparable form [9]—can be distilled: (1) climate and climate change, (2) natural resources (especially water and soil),

(3) urbanization or settlement development, (4) production, consumption, and (alternative) energy(s), (5) poverty and justice, and (6) migration. ESD as a teaching principle, with the issues mentioned here as examples, creates learning opportunities for teaching a number of subjects [39]. A test stem with items was developed for each of these six fields. Within those fields, our tasks refer to the following subject areas: climate change and coral reefs, electronic waste, megacities, beef consumption and the rainforest, textile production, and overtourism (test item examples are given in Appendix A).

Step 2: Identification of ESD-Specific Conceptual Principles in the Context of Systems Thinking

The above-mentioned principles of dealing with learning subjects in the field of ESD such as considering multiple perspectives or dealing with uncertainty [32] were given special consideration in the context of the present test development. These conceptual principles as well as other basic principles of item development (explanation of the competency model, explanation of different task formats, . . .) were presented in a manual for task development.

Step 3: Task Development

On the basis of the topics and the manual for task development, six tasks with one test stem and 10 to 12 test items each were developed. The test items targeted the three different competence levels. The test items had different forms (text, impact diagrams, etc.), while only closed item formats were included in order to ensure the desired compactness of the test.

The conceptualization of the new tasks was based on an iterative task development process [21], in which the authors of this paper were divided into two teams. Based on a manual including the theoretical conceptualization of system competence, the principles of ESD didactics and specific topics, the first team created task prototypes. Feedback from the second team (as well as from external experts in the field of ESD/system competence) was then used to further optimize the tasks. We repeated this process several times. This comparatively complex procedure offers decisive advantages: Firstly, not only the tasks are gradually improved, the feedback and adjustments are also used to refine the theoretical conceptualization of the target competence. Secondly, experiences (e.g., with the predecessor model GeoSysko; [24]) show that by this procedure a very large proportion of the developed items can be retained for the final version of the test.

Step 4: Cognitive Laboratories

The developed tasks went through a cognitive laboratory procedure [40] in order to identify problems in task construction or test procedure at an early stage. In a first step (“cognitive walk-through”), the task developers went through the individual items and specified how a potential respondent is likely to interpret the task and which mental operations should be required to answer it. By naming the processes and steps necessary for the solution of an item, indications of problematic items and possibilities to avoid comprehension difficulties can be identified. This first “cognitive walk-through” also served to better structure the interview guidelines for the subsequent second step, when the items were tested with students. The results regarding text comprehension were generally satisfactory. Some wordings were simplified and single words were replaced. In the case of essential technical terms, individual solutions were developed.

In the second step (“cognitive laboratories”), the tasks were completed by five students from the target population. The task of the interviewers leading the test sessions was to encourage the students to think aloud during the task processing. Furthermore, the students went through probing (students were asked to explain their answers), paraphrasing (understanding of the item/text), and confidence rating (degree of confidence regarding the answer) after they had completed the tasks [41]. This procedure provided information on the text comprehension, the students’ solution strategies, and difficulties. Based on the guidelines developed in the first step, the interviewers also addressed students with regard to individual operations that might be problematic. In addition, the system competence test requires students to meet a minimum standard in understanding concept maps. Therefore, an introductory explanation of concept maps including an exercise preceded the assessment instrument and was also tested in this phase. A central finding from this phase was that sequences of terms in which individual connections are not consistent and which areonly coherent in terms of the complete sequence, lead to difficulties in understanding. Accordingly, we decided that all individual connections between two concepts within a larger sequence of connected concepts must always be comprehensible in terms of language and content. For example, the chain of action "CO_2 in the air → increases → air temperature → acidifies → ocean" was changed to "CO_2 in the air → causes → increase in air temperature → leads to → ocean acidification". Again, after this step of the cognitive laboratory procedure, technical terms and respective descriptions were adjusted accordingly to the difficulties revealed by the paraphrasing and probing procedure.

Step 5: Feedback from Experts

In parallel to the cognitive laboratories, we asked teachers to assess our tasks with regard to their appropriateness in terms of subject content, educational aims, and language for the 8^{th} and 9^{th} grade. The review of the tasks by teachers provided, among other things, important insights with regard to incorrect or incomplete student conceptions. Based on this, we clarified the language in various cases. Teachers also provided information on subject-specific practices, such as the omission of abbreviations (e.g., "CO_2") for chemical formulas in texts.

Furthermore, an expert rating was conducted with external experts in the field of system competence from German and Swiss universities, who assessed to what extent system competence was adequately operationalized in the information text and the items of the test booklets. In addition, they evaluated whether, from a theoretical point of view, the selected item formats were suitable for testing system competence. After this expert rating, we only optimized individual items marginally.

3. Methods

In the following, we will describe the sample, the test instrument, the data collection procedure, and the analysis of the data.

3.1. Sample

The sample of the pilot study included 366 students from three schools of the type "Gesamtschule" (comprehensive school) one school of the type "Realschule" (secondary-level school), one school of the type "Hauptschule" (secondary-level school), and two schools of the type "Gymnasium" (grammar school) in the German Federal States of Bavaria, Hesse, and North Rhine-Westphalia.

The students were 8th and 9th graders. Altogether 45.66% (n = 168) were female and 54.34% (n = 198) male, while the average age was 14.25 (SD = 0.87) and 87.6% of them indicated that they "always" or "mostly" spoke German at home.

3.2. Instruments

The instrument was structured as follows: Firstly, the students were asked about demographic data such as age, gender (open-response), and language at home (whether German was spoken "never", "sometimes", "mostly", or "always" at home). Subsequently, the tasks targeting system competence were presented. Each student was given one of six different test booklets each containing three of the six tasks. A balanced design was used to distribute the tasks to the booklets, ensuring that each task was presented at each of the three positions within the booklets and that each combination of tasks was presented at least once (three combinations appeared twice). After the system competence tasks, students answered some additional questionnaire items that were tested in the pilot study but are not part of the system competence assessment.

3.3. Procedure

All students and their parents gave written informed consent to take part in the study. The test was administered in the classrooms, following a standardized instruction on how to complete the test booklets and an explanation of concept maps by external test administrators who visited school during a 90-min period. After the introduction, the test booklets were distributed randomly. The students worked on average about 60 to 70 min on the test booklets.

3.4. Analyses

To obtain scores for scaling, we proceeded as follows: Initially, raw scores were assigned for each item. Raw scores were based on the degree of agreement with the perfectly correct pattern. In a simple multiple choice item, a student received the raw score 1 for the correct answer and the raw score 0 for the wrong answers. For more complex items, a student received as many raw score points as sub-tasks within that item were solved correctly. For example, for the items in which arrowheads must be drawn to complete a concept map, a student received as many raw score points as he or she had correctly completed the single arrows. The raw point values were then converted into test scores. For the less complex item formats (all forms of simple multiple choice items, detecting mistakes within concept map representations of systems, assigning elements of a system into the right order), students received a test score of 1 for a completely correct response and 0 for responses containing mistakes (i.e., with raw scores lower than the maximum). For more complex formats in which relatively independent correct partial solutions are possible (adding arrowheads or "+" and "−" to a concept map, several multiple choice elements within one test item) students received a score of 2 for a completely correct response (i.e., the maximum number of raw score points). For responses with few mistakes (the exact number depending on the achievable raw score), they received a partial credit score of 1.

The scaling of the test was performed in the framework of item response theory (IRT; e.g., [42]) using the R package Test Analysis Modules (TAM) Version 3.5-19 [43,44]. The one-dimensional Rasch model was specified for dichotomous items and the partial credit model for polytomous items.

To assess the quality of the single items, we inspected item difficulties and infit (weighted mean square) item fit statistics from the IRT calibration. Additionally, we considered corrected item-score correlations (excluding the item of interest from the score) and correlations of items and with an additional corrected score excluding all items from the same task.

The IRT scaling procedure provides both estimates of students' competencies and item difficulties (or difficulties of item steps for partial credit items) on one common scale. This allowed us to inspect the congruence of both distributions, that is, we could explore whether the item difficulty range fits the range of students' competence levels or whether the items are either too difficult or too easy or whether the range of item difficulties is too broad or too narrow for our target population. We complemented this graphic-descriptive comparison of the two distributions by testing the arithmetic means (t-test) and variances (F-test) for differences. The analysis also yielded estimates for the reliability of the overall test.

Additionally, we inspected local item dependencies (Q3 statistics from the Partial Credit Model) to get a first indication of whether our test captures system competence with only one dimension. The Q3 statistics [45] provide information on correlations between individual test items after controlling for the individual level of system competence. Correlations above $r = 0.2$ [46] are an indication that the correlations between items are not only due to the dimension targeted by the test. In other words, those residual correlations are an indication of violations of unidimensionality and possibly indicate multidimensionality.

To obtain further indications of the dimensionality of our test, we also directly tested the dimensionality confirmatorily. Due to the limited size of the sample, we conducted a unidimensional confirmatory factor analysis (CFA) with aggregated scores for each of the six tasks instead of using individual item scores (item parceling; e.g., [47]). CFA was conducted using the R package lavaan (version 0.6-7) [48].

4. Results

In the following, we present the results of our pilot study, i.e., the results on the reliability, the alignment between students' system competence levels and the difficulty of the test items, as well as on the dimensionality of our competence test.

4.1. Item Discrimination and Reliability

Based on item-total correlations, at least seven of the 73 items were identified that do not satisfactorily distinguish between students with high and low system competence; they had correlations below 0.2. An additional eight items had values below 0.3. Correlations corrected by excluding all items from the same task from the score were marginally lower than the traditional corrected item-total correlations and provided very consistent information. Based on the infit from IRT scaling, fewer items were flagged as misfitting. Only one of 91 threshold parameters had an infit higher than 1.3, and an additional four had values higher than 1.25. The five items associated with high infit values were all included in the seven items with the lowest item-total correlations. Although at least the items with the lowest item-total correlations have to be revised or dropped from the test when compiling a version for further use, they were included in the analyses presented here. WLE reliability for the test with all items and each student responding to half of the tasks was 0.89 and EAP reliability was 0.90. Based on the Spearman-Brown prophecy formula, WLE reliability can be expected to still reach a value of 0.84 if students would respond to two of the six tasks, and of 0.73 if scores were based on a single task.

4.2. Matching of the Distributions of Person Abilities and Item Difficulties

Classical item difficulties (on a metric with P_i from 0 to 1) showed that items were overall of average difficulty ($\overline{P} = 0.53$), with only few very difficult (four with $P_i < 0.20$) and few very easy items (four with $P_i > 0.80$). Item difficulty parameters δ_i on the joint IRT scale for items and persons makes the fit of items with persons even clearer. The Wright map in Figure 1 contrasts the distribution of individual abilities (left side of the figure) with the difficulties of the items or the individual item thresholds for partial credit items (right side of the figure; each item or item threshold for partial credit items is represented by one asterisk while their positions indicate the respective difficulties). If the distribution of difficulties were lower than the distribution of personal skills, the items would be too easy for the sample, if it was higher, they would be too difficult. If the distribution of persons was broader than the distribution of difficulties, this would mean a lack of items to reliably assess persons with particularly high or low system competence.

A visual comparison of the two distributions already suggests that they correspond both in regard to position and variance. This impression can also be corroborated by means of inferential statistics: while the mean value of student competencies is fixed at zero, the mean value of difficulties across all thresholds is $\overline{\delta} = -0.076$, which is not significantly different from zero ($t = -0.61; df = 89; p = 0.55$). A comparison of the variances also shows that they do not differ significantly from one another ($var(\theta) = 1.23$; $var(\delta) = 1.42$; $F_{89,365} = 1.15$; $p = 0.371$).

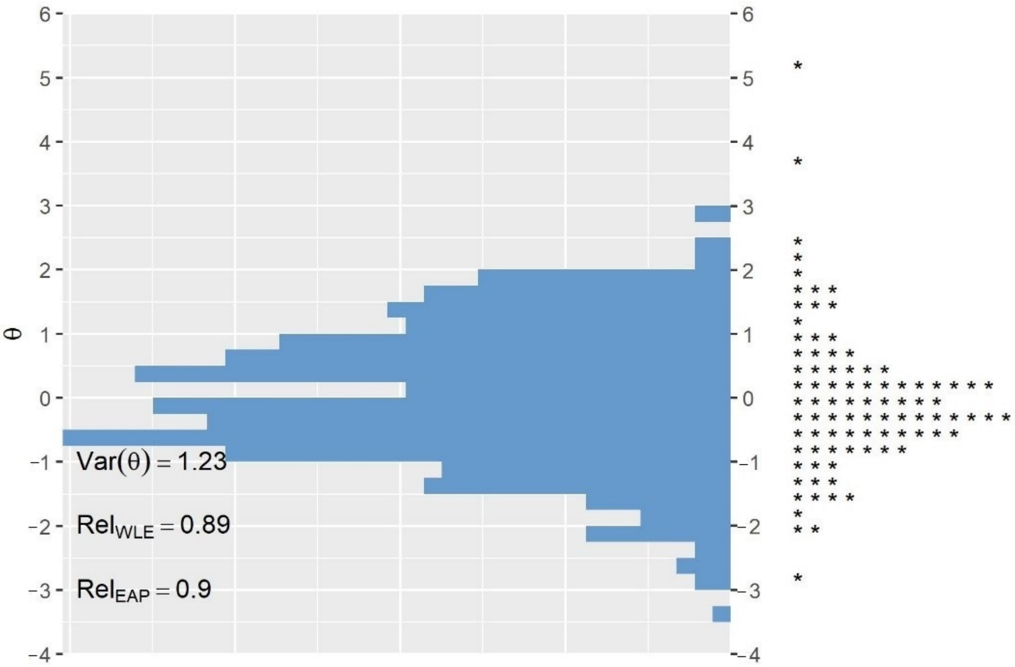


Figure 1. Distribution of the students' system competence θ and of the item difficulty parameters δ (each item or item threshold for partial credit items is represented by one asterisk).

4.3. Dimensionality

Figure 2 shows the distribution of the resulting average residual correlations between items. In total, 114 item pairs (4.3%) had Q3 values above the cutoff of 0.2 [46]. The mean of the Q3 values was slightly negative (−0.03), which is to be expected when local independence holds [45]. Overall, the Q3 values do not provide evidence that the correlations are due to further dimensions in addition to the competence addressed by our test.

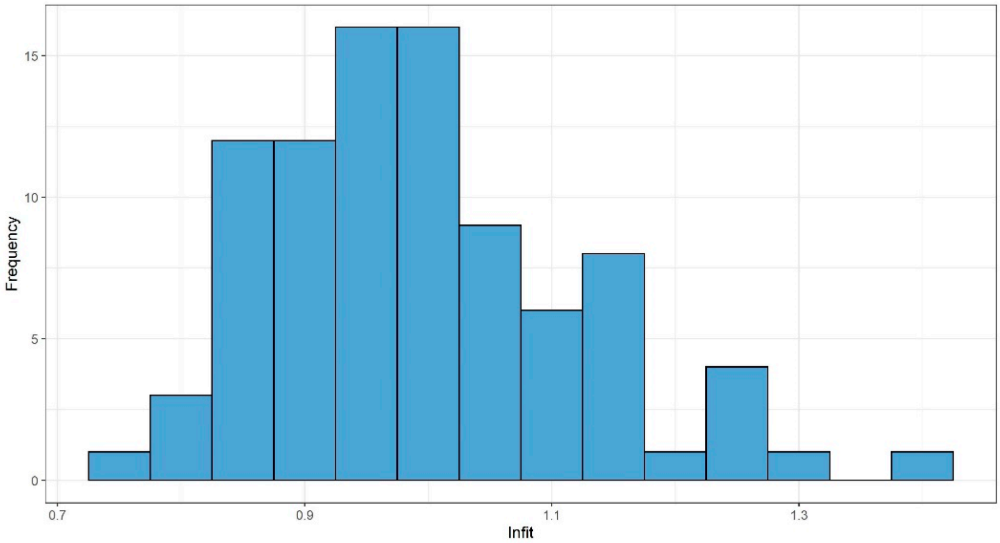


Figure 2. Distribution of the Q3 statistics.

The results of the one-dimensional confirmatory factor analysis with scores from the six tasks (see Table 1) point in the same direction: the good fit suggests the plausibility of the correlations between the test items being caused by only one dimension, namely, system competence.

Table 1. One-dimensional confirmatory factor analysis with scores from the six tasks: standardized factor loadings.

Task	λ	SE
Electronic waste	0.84	0.032
Climate change and coral reefs	0.82	0.035
Megacities	0.85	0.033
Beef consumption and the rainforest	0.79	0.039
Textile production	0.80	0.036
Overtourism	0.87	0.032

Note: Model Fit—2 = 26.57, df = 9, p = 0.002, CFI = 0.966, TLI = 0.943, RMSEA = 0.073, SRMR = 0.055.

5. Discussion

The aim of the present study was to develop an instrument for the assessment of system competence in the field of ESD which is not only reliable and valid, but also economical and easy to use and interpret, so that it can also be used for large-scale monitoring purposes in that field. To achieve this goal, we have developed a new test instrument that is not based on the content and principles of individual school subjects unlike existing tests or models addressing systems competence (see, e.g., Schecker and colleagues for physics [49], Mambrey and colleagues for biology [50], or Mehren and colleagues for geography [7]). Instead, it focuses on central themes and principles of ESD and, therefore, can be used independently of individual school subjects. Furthermore, the measurement instrument presented in this article is based exclusively on closed item formats in order to meet the requirement of a simple and economical implementation in the context of ESD monitoring. Existing instruments in contrast require elaborate coding and evaluation procedures [51,52].

Building on the results of cognitive laboratories and expert feedback from various fields, the instrument was (further) developed in an iterative process of feedback and revision. Subsequently, we conducted a quantitative pilot study to test the quality of the individual items, the reliability of the test as a whole, the adequacy of item difficulties in regard to the person abilities, and to collect first indications of the test's dimensionality.

Thedevelopment of our system competence test was successful—the test yieldeda high reliability, and only a small proportion of items showed a lack of fit with the (partial credit) Rasch model and, related to this, a lack of capacity to distinguish high and low competent persons. The result that a satisfactory reliability is still achieved with even only one or two tasks per student is particularly pleasing with regard to the goal of using the test for monitoring purposes in the field of ESD. Furthermore, there was a good match of the distributions of item difficulties and the students' competence levels. First analyses of dimensionality using Q3 statistics and a CFA with the tasks as indicators suggest that the test measures a one-dimensional construct, namely, system competence.

Despite these encouraging results, there are limitations to be considered and further analysis is needed. The fact that our instrument can now be used easily and economically (instead of, for example, using expert ratings of student responses) could come at a price: the instrument must be critically reviewed to determine whether important elements of the understanding of complex socio-ecological systems, such as the ambiguity of right and wrong or the limited predictability, can be adequately captured using closed item formats.

For future analyses with data from a larger calibration sample, in-depth analyses based on individual items and not only on scores per task are required to confirm the one-dimensionality. In addition, the setting of competence levels is needed to allow a criterion-oriented interpretation of test results. Furthermore, we will compare alternative scoring possibilities and investigate how well the empirically found difficulties can be predicted by difficulty-generating characteristics that were manipulated during item development.

6. Conclusions

In this paper, we have presented the development and psychometric evaluation of a system competence test for the field of ESD. The test is characterized by the following properties: It is compact, easy to interpret, and yet reliable, and, thus, suitable to be included into international educational monitoring reports. Particularly, it can serve as an outcome indicator for SDG 4.7. However, it is also suitable for evaluating ESD measures. With these characteristics, we hope to contribute to a more effective promotion and assessment of system competence within the framework of ESD.

Author Contributions: Conceptualization, F.F., J.F., J.H., R.M. and N.R.; methodology, J.H. and N.R.; formal analysis, J.H.; investigation, F.F., J.F., J.H., R.M. and N.R.; writing—original draft preparation, F.F., J.F., R.M. and N.R.; writing—review and editing, F.F., J.F., J.H., R.M. and N.R.; visualization, J.H.; supervision, J.H. and R.M.; project administration, J.H., R.M. and N.R.; funding acquisition, J.F., J.H., R.M. and N.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the German Federal Ministry of Education and Research (BMBF), grant number 01JJ17113A/ 01JJ17113B".

Institutional Review Board Statement: Not applicable.

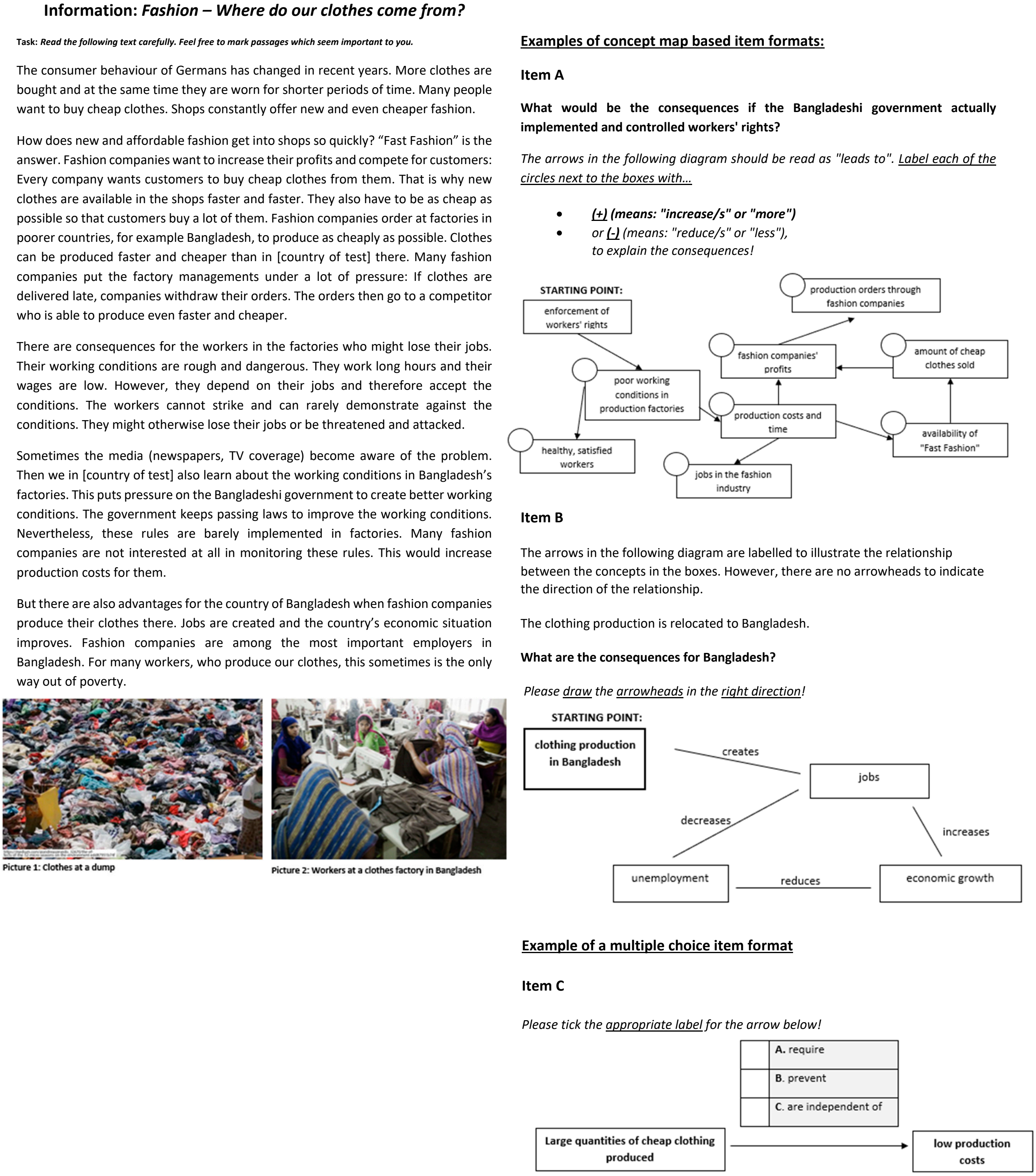
Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: After finalization of the test instrument, the data will be made publicly available via the German Network for Educational Research Data—VerbundFDB (https://www.forschungsdaten-bildung.de/ueber-verbund?la=en, accessed on 10 March 2021).

Acknowledgments: We wish to thank Astrid Henkel for her language support, Maria Noorani-Yazdanabad and Silan Gülmez for their support with item development and data collection.

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, or in the decision to publish the results.

Appendix A



Information: *Fashion – Where do our clothes come from?*

Task: ***Read the following text carefully. Feel free to mark passages which seem important to you.***

The consumer behaviour of Germans has changed in recent years. More clothes are bought and at the same time they are worn for shorter periods of time. Many people want to buy cheap clothes. Shops constantly offer new and even cheaper fashion.

How does new and affordable fashion get into shops so quickly? "Fast Fashion" is the answer. Fashion companies want to increase their profits and compete for customers: Every company wants customers to buy cheap clothes from them. That is why new clothes are available in the shops faster and faster. They also have to be as cheap as possible so that customers buy a lot of them. Fashion companies order at factories in poorer countries, for example Bangladesh, to produce as cheaply as possible. Clothes can be produced faster and cheaper than in [country of test] there. Many fashion companies put the factory managements under a lot of pressure: If clothes are delivered late, companies withdraw their orders. The orders then go to a competitor who is able to produce even faster and cheaper.

There are consequences for the workers in the factories who might lose their jobs. Their working conditions are rough and dangerous. They work long hours and their wages are low. However, they depend on their jobs and therefore accept the conditions. The workers cannot strike and can rarely demonstrate against the conditions. They might otherwise lose their jobs or be threatened and attacked.

Sometimes the media (newspapers, TV coverage) become aware of the problem. Then we in [country of test] also learn about the working conditions in Bangladesh's factories. This puts pressure on the Bangladeshi government to create better working conditions. The government keeps passing laws to improve the working conditions. Nevertheless, these rules are barely implemented in factories. Many fashion companies are not interested at all in monitoring these rules. This would increase production costs for them.

But there are also advantages for the country of Bangladesh when fashion companies produce their clothes there. Jobs are created and the country's economic situation improves. Fashion companies are among the most important employers in Bangladesh. For many workers, who produce our clothes, this sometimes is the only way out of poverty.

Picture 1: Clothes at a dump

Picture 2: Workers at a clothes factory in Bangladesh

Examples of concept map based item formats:

Item A

What would be the consequences if the Bangladeshi government actually implemented and controlled workers' rights?

The arrows in the following diagram should be read as "leads to". Label each of the circles next to the boxes with…

- ***(+) (means: "increase/s" or "more")***
- *or (-) (means: "reduce/s" or "less"), to explain the consequences!*

STARTING POINT:

enforcement of workers' rights

poor working conditions in production factories

healthy, satisfied workers

production costs and time

jobs in the fashion industry

fashion companies' profits

production orders through fashion companies

amount of cheap clothes sold

availability of "Fast Fashion"

Item B

The arrows in the following diagram are labelled to illustrate the relationship between the concepts in the boxes. However, there are no arrowheads to indicate the direction of the relationship.

The clothing production is relocated to Bangladesh.

What are the consequences for Bangladesh?

Please draw the arrowheads in the right direction!

STARTING POINT:

clothing production in Bangladesh

creates

jobs

decreases

increases

unemployment

reduces

economic growth

Example of a multiple choice item format

Item C

Please tick the appropriate label for the arrow below!

□	**A.** require
□	**B.** prevent
□	**C.** are independent of

Large quantities of cheap clothing produced → **low production costs**

Figure A1. Test item examples from the task "textile production". Note: Each of the six tasks (i.e., climate change and coral reefs, electronic waste, megacities, beef consumption and the rainforest, textile production, and overtourism) consists of an information sheet ("test stem"—text, pictures, graphics) and 10–12 items. The information sheet basically contains all the factual information that is necessary for processing. The test is at the moment only available in German. The translation presented above was made for illustration purposes only.

References

1. Dörner, D. *Die Logik des Mißlingens: Strategisches Denken in komplexen Situationen*; Rororo: Hamburg, Germany, 2000.
2. Booth-Sweeney, L.; Sterman, J.D. Thinking about systems: Student and teacher conceptions of natural and social systems. *Syst. Dyn. Rev.* **2007**, *23*, 285–311. [CrossRef]
3. Hmelo-Silver, C.E.; Marathe, S.; Liu, L. Fish Swim, Rocks Sit, and Lungs Breathe: Expert-Novice Understanding of Complex Systems. *J. Learn. Sci.* **2007**, *16*, 307–331. [CrossRef]
4. Yoon, S.A. An Evolutionary Approach to Harnessing Complex Systems Thinking in the Science and Technology Classroom. *Int. J. Sci. Educ.* **2008**, *30*, 1–32. [CrossRef]
5. Meadows, D.H.; Meadows, D.L.; Randers, J.; Behrens, W.W. *The Limits to Growth*; Universe Books: New York, NY, USA, 1972.
6. Scheunpflug, A. *Biologische Grundlagen des Lernens*; Cornelsen Scriptor: Berlin, Germany, 2001.
7. Mehren, R.; Rempfler, A.; Buchholz, J.; Hartig, J.; Ulrich-Riedhammer, E.M. System competence modelling: Theoretical foundation and empirical validation of a model involving natural, social and human-environment systems. *J. Res. Sci. Teach.* **2018**, *55*, 685–711. [CrossRef]
8. Wiek, A.; Withycombe, L.; Redman, C.L. Key competencies in sustainability: A reference framework for academic program development. *Sustain. Sci.* **2011**, *6*, 203–218. [CrossRef]
9. UNESCO. *Education for Sustainable Development Goals: Learning Objectives*; UNESCO: Paris, France, 2017; ISBN 9789231002090.
10. Rieckmann, M. Schlüsselkompetenzen für eine nachhaltige Entwicklung der Weltgesellschaft. Ergebnisse einer europäisch-lateinamerikanischen Delphi-Studie. *GAIA Ecol. Perspect. Sci. Soc.* **2011**, *20*, 2011.
11. Buckler, C.; Creech, H. *Shaping the Future We Want: UN Decade of Education for Sustainable Development (2005–2014): Final Report*; UNESCO: Paris, France, 2014; ISBN 9789231000539.
12. UNESCO. *UNESCO Roadmap for Implementing the Global Action Programme on Education for Sustainable Development*; UNESCO: Paris, France, 2014.
13. UN General Assembly. *Transforming Our World: The 2030 Agenda for Sustainable Development*; UN General Assembly: New York, NY, USA, 2015.
14. European Comission. *Integrating the Environment and Climate Change into EU International Cooperation and Development*; European Commission: Brussels, Belgium, 2016.
15. Tilbury, D.; Janousek, S.; Denby, L.; Elias, D.; Bacha, J.; Haddad, C. *Monitoring and Assessing Progress during the UNDESD in the Asia-Pacific Region: A Quick Guide to Developing National ESD Indicators*; UNESCO Asia and Pacific Region Bureau for Education: Bangkok, Thailand, 2007; ISBN 92-9223-115-4.
16. Köck, H.; Rempfler, A. *Erkenntnisleitende Ansätze—Schlüssel zur Profilierung des Geographieunterrichts: Mit erprobten Unterrichtsvorschlägen*; Aulis-Verl. Deubner: Köln, Germany, 2004; ISBN 978-3-7614-2483-4.
17. Fischer-Kowalski, M.; Weisz, H. Society as hybrid between material and symbolic realms: Toward a theoretical framework of society-nature interaction. *Adv. Hum. Ecol.* **1999**, *8*, 215–252.
18. Bertalanffy, L.V. *General System Theory*; Braziller: New York, NY, USA, 1968.
19. Rempfler, A.; Uphues, R. Sozialökologisches Systemverständnis: Grundlage für die Modellierung von geographischer Systemkompetenz. *Geogr. Didakt.* **2010**, *38*, 205–217.
20. Ben-Zvi Assaraf, O.; Orion, N. Development of system thinking skills in the context of earth system education. *J. Res. Sci. Teach.* **2005**, *42*, 518–560. [CrossRef]
21. Lecher, T.; Hoff, E.H. *Die Umweltkrise im Alltagsdenken*; Beltz: Weinheim, Germany, 1997; ISBN 3621273727.
22. Jacobson, M.J. Problem solving, cognition, and complex systems: Differences between experts and novices. *J. Complex* **2001**, *6*, 41–49. [CrossRef]
23. Sommer, C. Untersuchung der Systemkompetenz von Grundschülern im Bereich Biologie. Ph.D. Thesis, Christian-Albrechts-Universität zu Kiel, Kiel, Germany, 2006.
24. Mehren, R.; Rempfler, A.; Ullrich-Riedhammer, E.-M.; Buchholz, J.; Hartig, J. Systemkompetenz im Geographieunterricht. *ZfDN* **2016**, *22*, 147–163. [CrossRef]
25. Rieß, W.; Stahl, E.; Hörsch, C.; Schuler, S.; Schwab, S.; Fanta, D.; Bräutigam, J.; Rosenkränzer, F.; Kramer, T. Förderung Systemischen Denkens bei Lehramtsstudierenden: Theoretische Grundlagen und Eingesetzte Messinstrumente. In Proceedings of the Conference "Theorie, Empirie & Praxis" of the "Fachsektion Didaktik der Biologie", Kassel, Germany, 17 September 2013; Mayer, J., Hammann, M., Wellnitz, N., Arnold, J., Werner, M., Eds.; Kassel University Press: Kassel, Germany, 2013.
26. Sweeney, L.B. *Thinking about Everyday Systems*; Harvard: Cambridge, MA, USA, 2004.
27. Tripto, J.; Ben-Zvi Assaraf, O.; Snapir, Z.; Amit, M. The 'What is a system' reflection interview as a knowledge integration activity for high school students' understanding of complex systems in human biology. *Int. J. Sci. Educ.* **2016**, *38*, 564–595. [CrossRef]
28. Ben-Zvi Assaraf, O.; Orion, N. Four case studies, six years later: Developing system thinking skills in junior high school and sustaining them over time. *J. Res. Sci. Teach.* **2010**, *47*, 1253–1280. [CrossRef]
29. Rosenkränzer, F.; Hörsch, C.; Schuler, S.; Riess, W. Student teachers' pedagogical content knowledge for teaching systems thinking: Effects of different interventions. *Int. J. Sci. Educ.* **2017**, *39*, 1932–1951. [CrossRef]
30. Cox, M. *A Systems Thinking Approach in Secondary Geography Education*; KU Leuven: Leuven, Belgium, 2018.

31. Nesbit, J.C.; Larios, H.; Adesope, O.O. How students read concept maps: A study of eye movements. In Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Vancouver, BC, Canada, 25–29 June 2007; Montgomerie, J., Seale, C., Eds.; Association for the Advancement of Computing in Education (AACE): Vancouver, BC, Canada, 2007.
32. Schockemöhle, J. *Außerschulisches Regionales Lernen als Bildungsstrategie für Eine Nachhaltige Entwicklung: Entwicklung und Evaluierung des Konzeptes, Regionales Lernen 21+*; Selbstverlag des Hochschulverbandes für Geographie und ihre Didaktik e.V. (HGD): Weingarten, Germany, 2009; ISBN 978-3-925319-31-0.
33. Schreiber, J.-R.; Siege, H. *Orientierungsrahmen für den Lernbereich Globale Entwicklung im Rahmen Einer Bildung für Nachhaltige Entwicklung: Ein Beitrag zum Weltaktionsprogramm 'Bildung für Nachhaltige Entwicklung: Ergebnis des Gemeinsamen Projekts der Kultusministerkonferenz (KMK) und des Bundesministeriums für Wirtschaftliche Zusammenarbeit und Entwicklung (BMZ)*, 2nd ed.; Cornelsen: Berlin, Germany, 2016.
34. Di Giulio, A.; Ruesch Schweizer, C.; Adomflent, M.; Blaser, M.; Bormann, I.; Burandt, S.; Fischbach, R.; Kaufmann-Hayoz, R.; Krikser, T.; Künzli David, C.; et al. *Bildung auf dem Weg zur Nachhaltigkeit. Vorschlag Eines Indikatoren-Sets zur Beurteilung von Bildung für Nachhaltige Entwicklung*; Universität Bern: Bern, Switzerland, 2011; ISBN 978-3-906456-66-9.
35. Schroeder, N.L.; Nesbit, J.C.; Anguiano, C.J.; Adesope, O.O. Studying and Constructing Concept Maps: A Meta-Analysis. *Educ. Psychol. Rev.* **2018**, *30*, 431–455. [CrossRef]
36. Benninghaus, J.; Mühling, A.; Kremer, K.; Sprenger, S. Complexity in Education for Sustainable Consumption—An Educational Data Mining Approach using Mysteries. *Sustainability* **2019**, *11*, 722. [CrossRef]
37. de Haan, G. The development of ESD-related competencies in supportive institutional frameworks. *Int. Rev. Educ.* **2010**, *56*, 315–328. [CrossRef]
38. Hemmer, I. Bildung für nachhaltige Entwicklung: Der Beitrag der Fachdidaktiken. In *Befähigung zu Gesellschaftlicher Teilhabe: Beiträge der Fachdidaktischen Forschung*; Menthe, J., Höttecke, D., Zabka, T., Hammann, M., Rothgangel, M., Eds.; Waxmann: Münster, Germany; New York, NY, USA, 2016; ISBN 9783830985600.
39. Bagoly-Simó, P. Implementierung von BNE am Ende der UN-Dekade. Eine internationale Vergleichsstudie am Beispiel des Fachunterrichts. *ZGD* **2014**, *42*, 221–256.
40. Alavi, S.M. On the adequacy of verbal protocols in examining an underlying construct of a test. *Stud. Educ. Eval.* **2005**, *31*, 1–26. [CrossRef]
41. Rexroth, M.; Prüfer, P. *Zwei-Phasen-Pretesting: ZUMA-Arbeitsbericht 2000/08*; Zentrum für Umfragen, Methoden und Analysen: Mannheim, Germany, 2000.
42. Embretson, S.E.; Reise, S.P. *Item Response Theory for Psychologists*; Reprinted 2009 by Psychology Press; Psychology Press: New York, NY, USA, 2009; ISBN 0805828192.
43. Robitzsch, A.; Kiefer, T.; Wu, M. TAM: Test Analysis Modules. R Package Version 3.5-19. 2020. Available online: http://finzi.psych.upenn.edu/R/library/TAM/html/TAM-package.html (accessed on 10 March 2021).
44. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.
45. Yen, W.M. Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Appl. Psychol. Meas.* **1984**, *8*, 125–145. [CrossRef]
46. Chen, W.-H.; Thissen, D. Local Dependence Indexes for Item Pairs Using Item Response Theory. *J. Educ. Behav. Stat.* **1997**, *22*, 265–289. [CrossRef]
47. Little, T.D.; Cunningham, W.A.; Shahar, G.; Widaman, K.F. To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits. *Sci. Ed.* **2002**, *9*, 151–173. [CrossRef]
48. Rosseel, Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *J. Stat. Softw.* **2012**, *48*, 1–36. [CrossRef]
49. Schecker, H.; Klieme, E.; Niedderer, H.; Ebach, J.; Gerdes, J. *Abschlussbericht zum DFG-Projekt Physiklernen mit Modellbildungssystemen: Förderung Physikalischer Kompetenz und Systemischen Denkens durch Computergestützte Modellbildungssysteme*; Institut für Didaktik der Physik an der Universität Bremen und Max-Planck-Institut für Bildungsforschung: Berlin, Germany, 1999.
50. Mambrey, S.; Timm, J.; Landskron, J.J.; Schmiemann, P. The impact of system specifics on systems thinking. *J. Res. Sci. Teach.* **2020**, *57*, 1632–1651. [CrossRef]
51. Brockmüller, S. Erfassung und Entwicklung von Systemkompetenz—Empirische Befunde zu Kompetenzstruktur und Förderbarkeit Durch den Einsatz Analoger und Digitaler Modelle im Kontext Raumwirksamer Mensch-Umwelt-Beziehungen. Ph.D. Thesis, Pädagogische Hochschule Heidelberg, Heidelberg, Germany, 2019.
52. Cox, M.; Elen, J.; Steegen, A. Fostering students geographic systems thinking by enriching causal diagrams with scale. Results of an intervention study. *Int. Res. Geogr. Environ. Educ.* **2020**, *29*, 112–128. [CrossRef]