European Journal of Geography Volume 9, Number 1: 105-120, March 2018
©Association of European Geographers

A TEST TO MEASURE STUDENTS' SYSTEMS THINKING ABILITIES IN GEOGRAPHY

Marjolein COX
KU Leuven, Department of Earth and Environmental Sciences, Leuven, Belgium
marjolein.cox@kuleuven.be

Jan ELEN
KU Leuven , Faculty of Psychology and Educational Sciences, *Leuven, Belgium*
jan.elen@kuleuven.be

An STEEGEN
KU Leuven, Department of Earth and Environmental Sciences, Leuven, Belgium
an.steegen@kuleuven.be

Abstract

In high school geography students are taught about complex geospatial relations that often focus on the interaction between humans and their environment. Systems thinking as an approach concentrates in general on understanding interconnections between variables in a system to see the bigger picture. Therefore systems thinking is an appropriate approach in geography education. Recently it is also seen as an important cognitive skill in some European countries. However, an appropriate tool to assess systems thinking in geography is currently missing. In this article, a test and its development are described. The test was administered with 617 students of 16 to 18 years old in secondary education in Flanders. Based on the validation process, the reliability measures and the distribution of the students' scores, it is concluded that the proposed test is a valid and reliable assessment instrument for future research.

Keywords*: systems thinking, measuring tool, geography education*

1. INTRODUCTION

Within science education, including geography education, and education for sustainable development, the importance of systems thinking is recognized. Different studies are examining how to foster systems thinking in education (Assaraf & Orion, 2005; Hmelo-Silver, Liu, Gray, & Jordan, 2015; Kali, Orion, & Eylon, 2003; Sweeney & Sterman, 2000). Students need to acquire insights in complex global and local interconnections in order to be able to make more sustainable decisions (Rempfler & Uphues, 2012; Riess & Mischo, 2010; Yoon & Hmelo-Silver, 2017). Geography as a course has an important role to fulfill as many themes taught in class are about sustainability issues and the core of the course is about the interaction between humans and the environment (International Geographical Union, 2016).

However, if the goal is to foster systems thinking in geography education, it should be possible to assess students on this cognitive skill. While efforts have been engaged in (Rempfler & Uphues, 2012), no validated and widely accepted framework is currently available. In a pilot study a test was developed to measure high school students' systems thinking skills in geography in Flanders (Cox et al, submitted). Although important insights were revealed such as large differences in students' level of systems thinking depending on their study background, it was noticed that adjustments of the test were needed in order to turn it into a valid and reliable research instrument. Therefore, a second test is developed as a measuring tool on systems thinking in geography. This article describes the theoretical background of this new test, the development of the test, the validation process it went through and the results on reliability measurements.

2. THEORETICAL BACKGROUND

Understanding spatial relations is crucial in geography. To understand geospatial issues it is necessary to have insight in the interactions that are prevalent in and between regions. This includes also interconnections between different spatial scales, between humans and their environment and between different time scales. These interactions are often not linear one-dimensional but multilateral, and including reinforcing and balancing loops (Rempfler & Uphues, 2012). More in general, geospatial relational thinking is an important goal in geography education (Favier & van der Schee, 2014). As systems thinking is about understanding relations and seeing the larger picture as well, it is evident that systems thinking is an indispensable cognitive skill that needs to be trained in geography education. In Flanders and Germany for example, system thinking is recently seen as a basic concept for geography education (Katholiek Onderwijs Vlaanderen, 2017; Rempfler & Uphues, 2012). In order to properly design test items, knowledge on earlier definitions and tests of systems thinking was acquired in a literature study, summarized in the following paragraphs.

2.1 Understanding systems thinking as a broad concept

Systems thinking is a broad concept or construct and defined in several ways. The concept is linked to and developed on system dynamics, but has a broader interpretation and does not necessarily refer to a quantitative and dynamic simulation analysis used to understand systems behavior as is the case in system dynamics (Forrester, 2007). Many authors specified and distinguished elements that are part of systems thinking. Arnold and Wade (2015) compared eight definitions of systems thinking and identified several elements reoccurring in one or more of the definitions. In half of the compared definitions reoccurring elements are 'interconnections/interrelationships', 'wholes rather than parts', 'feedback loops' and 'dynamic behavior'. Several authors seem to agree that these elements should be understood in order to be a good systems thinker. Based on their comparison Arnold and Wade (2015) suggest a new definition of systems thinking and specify eight important elements in systems thinking: mainly based on Hopper and Stave (2008), Plate and Monroe (2014), and Sweeney and Sterman (2000). (1) recognizing interconnections, (2) identifying and understanding feedback, (3) understanding system structure, (4) differentiating types of stocks, flows, variables, (5) identifying and understanding non-linear relationships, (6) understanding dynamic behavior, (7) reducing complexity by modeling systems conceptually, and (8) understanding systems at different scales.

These eight cognitive skills are also quite comparable to the list of eight emergent characteristics of systems thinking of Assaraf and Orion (2010). The difference is that the latter organize these characteristics into a hierarchic model, the Systems Thinking Hierarchical model, representing the development of systems thinking in the context of earth systems education into three levels. The model has a pyramid structure where each level is necessary to acquire for the development of the skills in the next level. The first and lowest level is called 'analysis of system components' and contains only the first characteristic, namely (1) the ability to identify the components of a system and processes within the system. The second level, called synthesis of system components, contains four different skills: (2) the ability to identify relationships among the system's components, (3) the ability to identify dynamic relationships within the system, (4) the ability to organize the systems' components and processes within a framework of relationships, and (5) the ability to understand the cyclic nature of systems. The third and highest level, called implementation, consists out of three characteristics: (6) the ability to make generalizations, (7) understanding the hidden dimensions of the system and, (8) thinking temporally: retrospection and prediction. The first level is comparable to what Brandstädter, Harms, and Großschedl (2012, p.2148) call 'structural systems thinking', and describe as 'the ability to identify a system's relevant elements and their interrelationships'. The second and third level are included in the term procedural systems thinking, namely 'the ability to understand the dynamic and time-related processes that emerge from the systems' structure' (Brandstädter et al., 2012, p.2148).

2.2 Measuring systems thinking: external representations to visualize mental models and thinking

Several intervention studies in different domains and age groups have measured the impact of their intervention with different methods, such as observations by the teacher, interviews, multiple choice questions, open ended questions, drawings, word associations and concept maps (Assaraf & Orion, 2005; Hmelo-Silver, Jordan, Eberbach, & Sinha, 2017; Hopper & Stave, 2008; Kali et al., 2003; Riess & Mischo, 2010; Schuler, Fanta, Rosenkraenzer, & Riess, 2017). These tools are often very specifically developed to measure the effectiveness of the intervention and are used on a small scale. No standardized measuring tool to test the level of systems thinking in general, nor in geography, is available today. Authors such as Hopper and Stave (2008) and Plate (2010) explicitly express the need for an appropriate instrument to assess systems thinking in general for educational researchers, as well as for educators to test the effectiveness of methods for teaching systems thinking.

In order to know how well someone understands systems, we should be able to understand what this person is thinking. In other words, the mental model of the person should be made external. Concept maps, developed by Novak and Cañas (2008), are suggested by several authors as an appropriate tool to approximate the invisible cognitive structures (Assaraf & Orion, 2005; Lücken & Sommer, 2010; Mehren, Rempfler, & Ullrich-Riedhammer, 2015; Novak & Cañas, 2008). Indeed, a concept map that consists of concepts connected to each other, is helpful to evaluate the conceptual understanding of the internal system structure (Brandstädter et al., 2012; Yin, Vanides, Ruiz-Primo, Ayala, & Shavelson, 2005). In some studies, the computer as a tool to construct concept maps, is suggested because its use results in a considerable higher complexity compared to paper-pencil mapping (Brandstädter et al., 2012). Computer mapping is also more convenient for the processing afterwards if measurements are conducted on a large scale.

Causal maps, in which only causal relations are visualized, are a variation on concept maps, where a combination of several kind of relations can be used. In a causal map, relations between variables are visualized by arrows as in concept maps, but a plus or minus sign is added to the arrow to clarify whether it is a positive relation (an increase of variable A leads to an increase of variable B, or a decrease of variable A leads to a decrease of variable B) or a negative relation (an increase of variable A leads to a decrease of variable B, or a decrease of variable A leads to an increase of variable B) (Öllinger, Hammon, von Grundherr, & Funke, 2015). Despite the fact that Öllinger et al. (2015) mention the lack of theoretical underpinning and empirical research, it is assumed that elaborating such causal maps fosters the understanding of interconnectedness. Authors as Plate (2010) recognize the development of cause-effect maps as an important step in the process of understanding the structure and dynamics of a system. These maps can contribute to the understanding that one cause can have multiple effects and therefore serve as a tool to go beyond linear thinking.

Within the field of system dynamics the term causal maps is not often used, but causal loop diagrams are. The main idea is the same, although in causal loop diagrams the focus is much more on feedback loops. It is used as 'a communication tool of the feedback structure representing the principal feedback loops of the systems which generate the reference dynamic of the systems (Bala, Arshad, & Noh, 2017: p.37)'. According to Lane (2008) drawing causal loop diagrams is not sufficient and only a simulation is effective to fully understand the system. But, he also recognizes the advantages of these diagrams e.g. to start discussions, to communicate about certain issues, the limited equipment necessary to draw them and the quick overview. As causal maps are closer related to the systems structure, whereas concept maps are often hierarchical and linear, it might serve as an appropriate tool to test the systems thinking abilities of students. To avoid confusion with geographical maps, the term causal diagram is used further on.

2.3 Operationalization of the construct

In order to develop an appropriate measuring tool, the notion was made operational as follows.

Systems thinking is a cognitive skill that enables to (Arnold & Wade, 2015; Assaraf & Orion, 2010):

(1) construct a causal diagram based on the information of a given source. This means
   (1a) identifying the relevant variables in the information
   (1b) recognizing the relations between the different variables
   (1c) assigning the nature of the relationship (+ or -)
(2) describe relations between variables in words
(3) explain the influence within a system if there is an interference

3. METHODS

As the aim of this study is to develop a measuring tool for systems thinking in geography, the developing and validation process is described in detail before elucidating the format of the test. A full translation of the test is added in the appendix.

3.1 Development and validation process of the test

During the development process of the test several actions were undertaken regarding the validity of the test. In a first phase, the items, including the text that students have to read to answer the items, were constructed by the first author and subsequently reviewed by both co-authors. Apart from a general review they were asked to focus on the agreement with the operational definition of systems thinking, the clarity of the questions and texts, the feasibility of the available time for students and the accuracy of the ideal answers on the questions. Thorough revisions were made e.g. the subject of the text in the first item was changed. The format of the other items was also revised from drawing a diagram towards providing a diagram and asking questions on it. It measures if the students can read a diagram and describe relations in their own words (second part of the definition) and if they are able to explain influences on the system in case of an interference (third part of the definition). This also provided more variation in item format and was expected to save time for the students when completing the test.

In a second phase, the revised test was completed by a student of the 11th grade. The goal was to examine whether the provided time was sufficient and whether the questions could elicit the desired responses. Based on the feedback of the student some items were slightly rephrased such as the word ‘slum’ into an easier Dutch synonym.

In the third phase, the revised version of the test was discussed by two expert groups and another expert individually. During this expert panel validation participants were requested to complete the test before a discussion was held on the accuracy of the geographical content, the phrasing of the instructions used in the test and the correctness of the model answer. The first group consisted of one retired professor in Geography and two teacher trainers in geography. Some suggestions were made regarding specific variables used in the diagram provided in the test, the rephrasing of the second item and the use of another graph in the fourth item in order to avoid confusion. These remarks were considered before another meeting with an expert professor. This geography professor made some suggestions to improve the model answer. In the last group 3 PhD-students and 3 post-doctoral researchers from a geography division at the same university suggested to improve the general instructions about the constructions of a causal diagram. Furthermore, the construct validity of the test was also discussed during these meetings by comparing the test to the operational definition of systems thinking. It was then discussed whether the test measures systems thinking as defined. All the experts concluded that all the elements of the definition were clearly tested with this measuring tool.

3.2 Format of the test

The paper and pencil test itself consists of six items (see appendix). Each of these items is constructed to measure a part of the operational definition. The first part of the definition which states that a systems thinker is able to construct a causal diagram based on information in a given source, is measured in the first and second item. In these items, the students have to read a text and construct a causal diagram that serves as an answer on a given research question. This means that the students have to identify the relevant variables in item 1, recognize the relations between the different variables in the text and assign the nature of the relation in item 2.

The second part of the definition which states that a systems thinker is able to describe relations between variables in words, is tested in item 3 and 4. In these items a causal diagram is provided in which several global challenges are linked to each other. In the third item students are asked to describe in their own words the relation between two of these given variables. Students therefore have to be able to correctly read the diagram. Also for the fourth item they have to describe in their own words the relations between two variables, but in addition to the third item they have to take into account a change in the status of the variables. So, in other words they have to be able to explain what effect a change of one of the variables will have on another variable. The latter (item 4) is therefore already testing the third part of the definition which states that a systems thinker is able to explain the influence in a system in case of an interference within the system. In addition this is also measured in item 5 and 6. In the fifth item a map and a graph are provided in addition to the causal diagram with global challenges used in item 3 and 4. Based on the information in both sources, students have to identify two relevant variables and add those to the provided diagram. Therefore students should be able to identify how these additional variables are connected to the existing variables in the diagram. The sixth item is similar to the fifth one, but a short article is provided instead of a map or graph, and the students have to add one variable and some connections to the diagram.

As established in the validation process a maximum of 45 minutes is sufficient to complete the test.

3.3 Creating a scoring guide

A scoring guide was developed and reviewed by both co-authors. The main idea behind this scoring guide is the comparison with the model answer that was approved by the experts. Generally the responses are inventoried before actual scores are given. This was all done in MS Excel. The scoring is explained in detail for the different items, and the scoring procedure for item 1 and 2 is given as an example in Figure 1 .

In the first item, in which students have to identify variables, the variables present in the model answer are looked for. If a variable is present in the model answer, but not in the student's response, a code zero is given. For each variable that is present in both the model answer and the student's response, a code is given in agreement with the amount of times the variable is present. This means for example that if a variable is used once, a code one is given, if a variable is used twice a code two is given etc. A list of acceptable synonyms of the variables in the model answer was created. Afterwards the given codes were translated into an actual score. Codes zero and one stayed zero and one, but the other codes were changed into a gradual score between 0 and 1 as a causal diagram is better if a variable is only used once. A code two becomes 0.5 and a code three becomes 0.33. To count for a total score on 1 on the first item, all the scores were summed up and divided by 14 as the number of variables in the model answer is 14.

In the second item, in which the students have to draw the relations between the variables in a causal diagram, all possible relations with the variables present in the model answer are searched for in the student's answer. If the relation is not found, a code zero is given. If a relation is found, a code is given according to how the student assigns the nature of the relation. E.g. if a relation is present visualized by an arrow and a minus sign, a code 1 is given; if a relation is present visualized by an arrow and a plus sign, a code 2 is given; if a relation is present visualized by an arrow accompanied with a word like 'more or increase', than a code 3 is given etc. In total 15 different codes could be given. Afterwards only the relations that were present in the model answer were taken into account. For these relations

the given codes were translated into a score between 0 and 1. Only if an arrow was drawn with the correct sign, the code was translated into a 1. But e.g. code 3 was translated into 0.8 if the model answer was an arrow with a plus sign and into a 0 if the model answer of that relation was an arrow with a minus sign. A table to translate the codes into scores was part of the scoring guide. The entire detailed scoring guide is available on request with the authors. The total score on this item was received by a summation of the score for each relation present in the model answer, divided by 19 as there were 19 relations that had to be found to have a complete diagram.

In the third and fourth item the students had to explain the relation between several variables in words. The variables that were present in the explanation in the model answer were searched for in the student's response. For each variable or synonym present and interpreted in a correct way, a code 1 was given. For each variable or synonym present, but interpreted in a wrong way, a code 2 was given. Afterwards the codes 2 were translated into 0.5. To count the total score, the scores per variable were summed up and divided by 6 in the third item and by 4 in the fourth item, as this is the number of variables that were present in the model answer for these items.

For the fifth and sixth item in which the students had to add variables and relations to the provided diagram, the score was determined by checking for the presence of the variables of the model answer on the one hand and the presence of the relations on the other hand. According to the procedures for the first and second item a code was given and translated into a gradual score between zero and one. The total score on this items was a summation of the scores for the variables present and for the present relations, divided by 6 for the fifth item and by 4 in the sixth item.

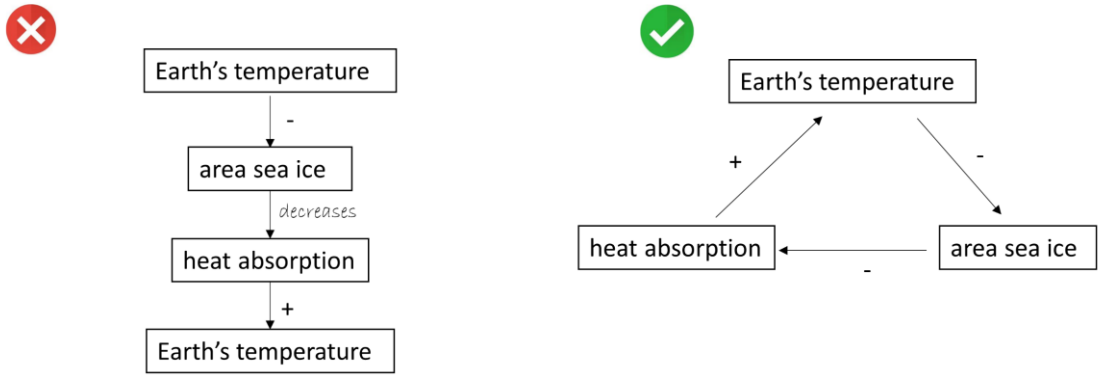
The total score on the whole test was the un-weighted sum of the six items, thus considering the different items as equally important.

A simple example to clarify the scoring guide:

Imagine that a student has to construct a causal diagram based on this text:

If the earth's temperature is rising, the area of sea ice will decrease faster. Due to a lower area of sea ice, the earth will absorb more heat because there is less ice that will reflect sunlight. Therefore Earth's temperature will increase further.

The students answer is shown on the left and marked as incorrect. The model answer is shown on the right and marked as correct.



Based on the scoring guide a comparison is made between the student's answer and the model answer. For the first item, each of the three variables would receive a code:

- Earth's temperature: code 2 as the word is used twice.
- Area sea ice: code 1
- Heat absorption: code 1

Code 2 is translated into a score of 0.5 and code 1 stays a score of 1. The total score for the first item is the summation of the scores for each variable and would be 2.5/3.

For the second item, the drawn relations are taken into account. For each possible relation a code is given. Code 0 means that the relation is not present. Code 1 means that the relation is present with a minus sign. Code 2 means that the relation is present with a plus sign. Code 4 means that the relation is present with a word meaning decreasing (e.g. decreases, decline, minus,…)

- Relation from 'Earth's temperature' to 'area sea ice': code 1
- Relation from 'Area sea ice' to 'heat absorption': code 4
- Relation from 'heat absorption' to 'Earth's temperature': code 2
- All other relations e.g. relation 'area sea ice' tot 'Earth's temperature': code 0

These codes are translated into a score according to the correct answer. Therefore, in this example both codes 1 and 2 receive a score of 1. Code 4 receives a score of 0.8. The total score on the item is the summation of the scores on the relations present in the model answer. In this case the student would earn 2.8 on 3. The relations are all present in the student's answer, but in one of the three relation the arrow is accompanied with a word instead of the correct symbol.

Figure 1. Example of the scoring guide for item 1 and 2.

3.4 Measuring the reliability

To check for the inter-rater reliability twenty randomly selected tests were scored by two independent raters. To allow for coincidental agreement, a Cohens Kappa was calculated for each item in IBM SPSS Statistics 24.

The internal consistency of the test was measured by calculating the Cronbach's alpha coefficient on a sample of 617 students in the last or penultimate year of high school in Flanders (age 16-18). The group consisted of 310 female students and 307 male students in different study programs from 15 different schools spread over Flanders.

4. RESULTS

4.1 Reliability measures

The Cohen's kappa coefficient was calculated to evaluate inter-rater reliability. As the inter-rater agreement was calculated for each of the variables and connections, an average score was taken to represent the inter-rater agreement for each item. These average scores range from 0.75 to 0.97 (item 1= 0.81; item 2= 0.83; item 3=0.97; item 4= 0.79; item 5= 0.78; item 6= 0.75) and can be considered substantial to almost perfect (Landis & Koch, 1977).

The internal consistency between the different items in the test, measured by the Cronbach's alpha coefficient is 0.611. According to Hinton et al. (2014) this can be classified as moderate reliability. The small amount of items and the heterogeneous character of the construct might be reasons for this rather low value.

4.2 Distribution of test scores

In Figure 2 the distribution of the total test scores is shown for the 617 participants. This total test score is obtained by the summation of the scores on the 6 items and rescaled to a score between 0 and 1. The histogram shows a frequency distribution that approximates a normal distribution. The mean score is 0.46 with a standard deviation of 0.12. The distribution is slanted to the left with a slightly negative skewness of -0.31. Overall, the test is successful in differentiating between levels in systems thinking, but with a 50th percentile of 0.47 and a 95th percentile is 0.65, it is clear that even the students with the highest score do not really perform excellent on the test. This might indicate that the test is rather difficult.

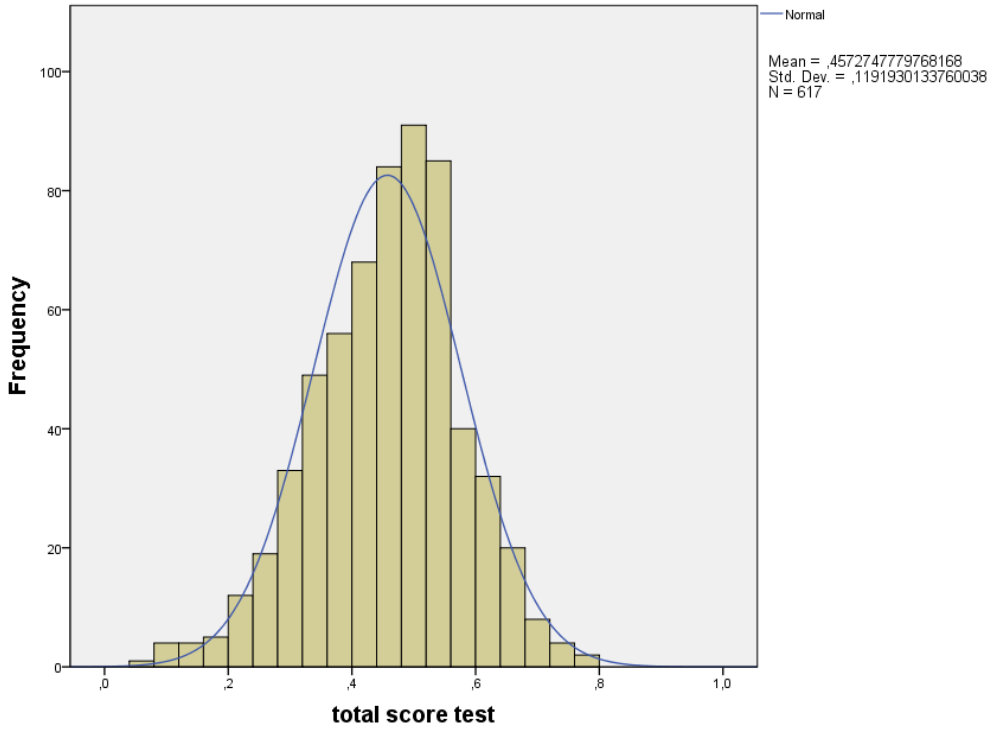
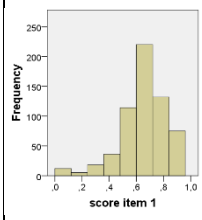
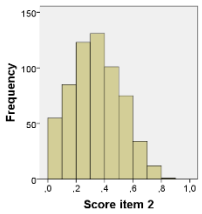
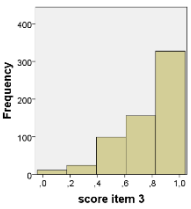
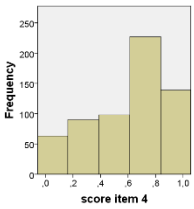
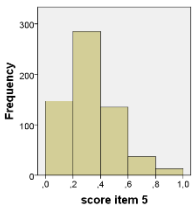
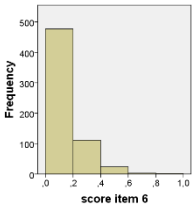


Figure 2. Histogram with the distribution of the total score on the test.

To have a better insight in the features of the test, the distribution for each item is given in Table 1. The mean scores on the different items vary from 0.06 to 0.74, which indicates that very difficult, intermediate, but also rather easy items are included in the test. Item 3, in which students have to read a given causal diagram and explain a relation in words, has the highest mean score of 0.74. The distribution is skewed to the left with a negative skewedness of -0.81. Therefore it can be considered as a rather easy item. Item 1 and 4 have a mean score of respectively 0.66 and 0.62. The distribution of the responses are both left skewed with a negative skewness of -1.22 for the first item and -0.57 for the fourth item. These items seem to be from an easy to intermediate level for the students to answer.

Table 1. Some characteristics of the distribution of the scores for each item separately.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Mean	0.66	0.33	0.74	0.62	0.33	0.06
Median	0.70	0.34	0.83	0.75	0.33	0
St. Dev.	0.17	0.18	0.21	0.31	0.20	0.12
Minimum	0.00	0.00	0.00	0.00	0.00	0.00
Maximum	1.00	0.82	1.00	1.00	0.83	0.80
Skewness	-1.22	0.15	-0.81	-0.57	-0.04	2.43
Histogram						

Item 2, which is dependent on item 1 as the relations are drawn between the variables found in item 1, and item 5 both have a mean score of 0.33. With respectively 0.82 and 0.83 as highest score obtained by the students for these items, the maximum score of 1 was not obtained by the participants. In addition the distribution of the scores is slightly skewed to the right for item 2 with a skewness of 0.15 and quite symmetric for item 5 with a skewness of -0.04. Therefore these items might be considered as intermediate to rather difficult. The last item has a very low mean of only 0.06 and a clear positive skew is visible in the distribution of the scores. This item is perceived very difficult.

5. DISCUSSION

Regarding the importance of systems thinking mentioned by several authors (Rempfler & Uphues, 2012; Riess & Mischo, 2010; Yoon & Hmelo-Silver, 2017) it is clear that a tool is necessary to measure students systems thinking abilities in general but also in geography in particular (Hopper & Stave, 2008; Plate, 2010). Based on former studies (Arnold & Wade, 2015; Assaraf & Orion, 2010; Brandstädter et al., 2012) an operational definition of systems thinking and test is proposed in this article. To discuss whether this test is actually measuring

systems thinking in geography, it is important not only to explain the validation process and the reliability measurements of the test, but also to give an overview of the different elements that make the test domain specific (Table 2). The four elements mentioned by Arnold and Wade (2015) that are reoccurring in several definitions of systems thinking in general are 'interconnections', 'wholes rather than parts', 'feedback loops', and 'dynamic behavior'. Our measuring tool contains all of these elements spread over the different items. Furthermore, our definition and test also represents each of the three levels of the systems thinking hierarchical model of Assaraf and Orion (2010). The first level, called 'analysis of system components' is clearly present in part 1a of our definition, and in items 1, 5 and 6 of the test. The second level, 'synthesis of system components', is present in part 1 and 2 of the definition and items 2, 3, 4, 5 and 6. Finally, the third level, called 'implementation', can be found in part 3 of the definition and in items 4, 5 and 6 of the test.

Table 2. The different cognitive skills regarding systems thinking in general, systems thinking in geography, and types of task present. It shows which item in the test includes which skill.

skills measured in the test	item 1	item 2	item 3	item 4	item 5	item 6
systems thinking in general						
understanding interconnections	-	x	x	x	x	x
seeing wholes rather than parts	x	x	x	x	x	x
understanding feedback loops	-	x	-	-	-	-
understanding dynamic behaviour	-	-	-	x	x	x
systems thinking in geography						
identify variables on different spatial scales	x	-	-	-	x	-
drawing/describing interconnections between different spatial scales	-	x	x	-	x	-
drawing/describing interconnections between human and environment	-	-	x	-	x	x
drawing/describing interconnections between different time scales	-	-	x	-	x	-
type of task						
drawing interconnections in diagram	-	x	-	-	x	x
identifying variables	x	-	-	-	x	x
describing interconnections in words	-	-	x	x	-	-
extract information from a combination of sources	-	-	-	-	x	-

Regarding the cognitive skills that are additionally important in geography a distribution over the different items is visible as well (Table 2). Working on several spatial scales is present in four items in which the students have to draw or describe interconnections between variables on different spatial scales. However, it is only in item 1 and 5 that the students also have to identify the variables on different spatial scales themselves. This is due to the causal diagram that is provided and should be interpreted in a correct way in items 3 and 4. Besides the focus on spatial scales itself the interaction between variables from natural sciences and from human sciences is present in item 3, 5 and 6. In item 3 and 5 also the interconnections on different time scales are present. Apart from the presence of these domain-specific cognitive skills, the geographical content itself is mostly provided in a text or in figures such as a map and graph. This means that it is not required to possess this content knowledge in order to succeed on the test.

The qualitative analysis of the test by multiple experts in geography and the quantitative analysis of the reliability show that this test is valid and reliable. Consequently, it can be used in this early stage of explicitly working and evaluating on systems thinking in geography courses and research in geography education. The rather low value of the Cronbach's alpha coefficient can probably be explained by the limited number of items in the test as this has a large impact due to the way in which this coefficient is calculated. But also the character of the construct itself can have an influence. The items in the test are supposed to partly measure different aspects of systems thinking, which implies that a very high value of Cronbach's alpha would not be feasible.

Furthermore, the distribution of the scores for each item shows that both rather difficult and easy items are present in the test. Based on the description of the test it is also clear that the test contains a variety of different tasks. An overview can be found in the bottom section of Table 2. The frequency diagram of the total score on the test indicates that the test can detect different levels of systems thinking in geography. However, no high scores above 0.8 are obtained by the students. Together with the very low score on item 6, it might be a reason to adjust this item into a slightly easier one when used in future studies. This would allow the best students to obtain a better total score on the test as well and would therefore create an instrument which allows to differentiate an even wider range of students' systems thinking abilities.

For the inter-rater reliability substantial to almost perfect scores of agreement were found. This means that the developed scoring guide is able to score the students' responses in a rather objective way. Some reflections that popped up while scoring the tests are worth mentioning here in the discussion. First, the way of scoring might induce a slightly overestimation of the real systems thinking abilities. During the scoring variables and relations are looked for that are also present in the model answer. For those that are present one can earn points, but for the connections that are present in the students answer but not in the model answer no points are subtracted. It is difficult to develop a rationale to add this 'problem' into a scoring guide as some of these connections that are added by the student might be correct as well. This is due to the fact that the model answer, as validated by the experts, only takes into account the information in the provided text. The student can use extra prior knowledge to add connections. Second, if one would calculate a total score on this test, the different items have an equal weight, so with six items each item would contribute for 16.7% of the total score. One can discuss on this distribution as it does not necessarily mean that all parts of the definition are equally important.

Even though this test is valid and reliable, more research is necessary in the future to examine the influence of language proficiency. After all, one should be able to read and understand texts as well as express yourself quite good in a language to be able to draw the diagrams and explain the relations in their own words. Furthermore, it is to be explored whether the spatial aspect of geographical systems can be integrated in a more explicit way in the test. It was observed during the study that students are not always aware of the different spatial scales or regions on which certain variables can cause an effect (Cox, Steegen, & Elen, accepted for publication). Despite improvements made in comparison with the initial measuring tool, it is observed that more attention for this spatial aspect is required.

6. CONCLUSIONS

The proposed paper-and pencil test is shown to be a valid and reliable tool to measure the level of systems thinking in geography. It is suggested that there is a difference between domain-general and domain-specific systems thinking, for which the important elements are distinguished. Observations during the scoring of the test revealed possibilities for further improvement of the measuring tool concerning the focus on feedback loops, the spatial aspect of geographical systems and the scoring efficiency. Future research could increase insight in the different possibilities, but meanwhile we hope that this tool can be used in intervention studies or other studies concerning the improvement of systems thinking in geography. Furthermore, it might also serve to evaluate systems thinking in geography by teachers.

REFERENCES

Arnold, R. D., and Wade, J. P. 2015. A definition of systems thinking: A systems approach. *Procedia Computer Science*, *44*, 669–678.

Assaraf, O. B.-Z., and Orion, N. 2005. Development of system thinking skills in the context of earth system education. *Journal of Research in Science Teaching*, *42*(5), 518–560.

Assaraf, O. B.-Z., and Orion, N. 2010. Four case studies, six years later: Developing system thinking skills in junior high school and sustaining them over time. *Journal of Research in Science Teaching*, *47*(10), 1253–1280.

Bala, B. K., Arshad, F. M., and Noh, K. M. 2017. *System dynamics, Modelling and Simulation*. Signapore: Springer.

Brandstädter, K., Harms, U., and Großschedl, J. 2012. Assessing System Thinking Through Different Concept-Mapping Practices. *International Journal of Science Education*, *34*(14), 2147–2170.

Cox, M., Steegen, A., and Elen, J. (accepted for publication). Using causal diagrams to foster systems thinking in geography education. *International Journal of Designs for Learning*.

Favier, T. T., and van der Schee, J. A. (2014). The effects of geography lessons with geospatial technologies on the development of high school students' relational thinking. *Computers and Education*, *76*, 225–236.

Forrester, J. W. 2007. System dynamics- a personal view of the first fifty years. *System Dynamics Review*, *23*(2/3), 345–358.

Hinton, R. P., McMurray, I., and Brownlow, C. 2014. *SPSS Explained*. New York: Routledge.

Hmelo-Silver, C. E., Jordan, R., Eberbach, C., and Sinha, S. 2017. Systems learning with a conceptual representation: a quasi-experimental study. *Instructional Science*, *45*(1), 53–72.

Hmelo-Silver, C. E., Liu, L., Gray, S., and Jordan, R. (2015). Using representational tools to learn about complex systems: A tale of two classrooms. *Journal of Research in Science Teaching*, *52*(1), 6–35.

Hopper, M., and Stave, K. 2008. Assessing the effectiveness of systems thinking interventions in the classroom. *Proceedings of the 26th International Conference of the System Dynamics Society*, 1–26.

International Geographical Union. 2016. International Charter on Geographical Education. Retrieved from http://www.igu-cge.org/Charters-pdf/2016/IGU_2016_def.pdf

Kali, Y., Orion, N., and Eylon, B. S. 2003. Effect of knowledge integration activities on students' perception of the earth's crust as a cyclic system. *Journal of Research in Science Teaching*, *40*(6), 545–565.

Katholiek Onderwijs Vlaanderen. 2017. *Leerplan secundair onderwijs: aardrijkskunde derde graad tso/kso*. Brussel.

Landis, J. R., and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data, *Biometrics*, *33*(1), 159–174.

Lane, D. C. 2008. The emergence and use of diagramming in system dynamics: A critical account. *Systems Research and Behavioral Science*, *25*(1), 3–23.

Lücken, M., and Sommer, C. 2010. System competence – Are elementary students able to deal with a biological system? *Nordic Studies in Science Education*, *6*(2), 125–143.

Mehren, R., Rempfler, A., and Ullrich-Riedhammer, E. M. 2015. Diagnostik von Systemkompetenz mittels Concept Maps. Malariabekämpfung im Kongo als Beispiel. *Praxis Geographie*, (7–8), 29–33.

Novak, J. D., and Cañas, a J. 2008. The Theory Underlying Concept Maps and How to Construct and Use Them. *IHMC CmapTools*, 1–36.

Öllinger, M., Hammon, S., von Grundherr, M., and Funke, J. 2015. Does visualization enhance complex problem solving? The effect of causal mapping on performance in the computer-based microworld Tailorshop. *Educational Technology Research and Development*, *63*(4), 621–637.

Plate, R. 2010. Assessing individuals' understanding of nonlinear causal structures in complex systems. *System Dynamics Review*, *26*(1), 19–33.

Plate, R., and Monroe, M. 2014. A Structure for Asssessing Systems Thinking. The Creative Learning Exchange. *23*(1), 1–12.

Rempfler, A., and Uphues, R. 2012. System competence in geography education: Development of competence models, diagnosing pupils' achievement. *European Journal of Geography*, *3*(1), 6–22.

Riess, W., and Mischo, C. 2010. Promoting Systems Thinking through Biology Lessons. *International Journal of Science Education*, *32*(January 2015), 705–725.

Schuler, S., Fanta, D., Rosenkraenzer, F., and Riess, W. 2017. Systems thinking within the scope of education for sustainable development (ESD) – a heuristic competence model as a basis for (science) teacher education. *Journal of Geography in Higher Education*, *8265*(June), 1–13.

Sweeney, L. B., and Sterman, J. 2000. Bathtub Dynamics : Initial Results of a Systems Thinking Inventory Bathtub Dynamics : Initial Results of a Systems Thinking Inventory. *System Dynamics Review*, *16*(4), 249–286.

Yin, Y., Vanides, J., Ruiz-Primo, M. A., Ayala, C. C., and Shavelson, R. J. 2005. Comparison of two concept-mapping techniques: Implications for scoring, interpretation, and use. *Journal of Research in Science Teaching*, *42*(2), 166–184.

Yoon, S. A., and Hmelo-Silver, C. 2017. Introduction to special issue: models and tools for systems learning and instruction. *Instructional Science*, 1–4.