CHARACTERISTICS OF TEST ITEMS FOCUSING ON MEANINGFUL LEARNING: A CASE STUDY IN PRE-VOCATIONAL GEOGRAPHY EDUCATION IN THE NETHERLANDS

Erik BIJSTERBOSCH
Windesheim University of Applied Sciences, Zwolle, The Netherlands
H.Bijsterbosch@Windesheim.nl

Tine BÉNEKER
Utrecht University, Utrecht, The Netherlands
T.Beneker@uu.nl

Wilmad KUIPER
Utrecht University, Utrecht, The Netherlands
W.Kuiper@uu.nl

Joop van der SCHEE
VU University, Amsterdam, The Netherlands
J.A.vander.Schee@vu.nl

Abstract

Summative assessments tend to encourage students' rote learning rather than meaningful learning. Yet, summative assessments might contribute to meaningful learning when they meet certain criteria, such as the use of test items and corresponding scoring rubrics that appeal to higher cognitive processes and to divergent assessment. In 2016, a small-scale study was conducted with six geography teachers of pre-vocational education to examine which type of test items and accompanying scoring rubrics are feasible and practical to support meaningful learning and which strategies can scaffold both teachers and students. The results showed that teachers were most positive about pre-structured test items. Both teachers and students were also positive about the application of a flow chart to scaffold students in answering the test items. The results showed that teachers encountered problems in scoring open, more complex test items focusing on evaluating and creating.

Keywords: *geography education, meaningful learning, summative assessment, test items, scoring rubrics.*

1. INTRODUCTION

The effect of assessment on learning has been studied extensively in recent decades. Several studies on this relationship have documented that teachers' classroom practices tend to encourage rote learning instead of meaningful learning (Black & Wiliam, 1998a, 1998b; James & Gipps, 1998; Klenowski & Wyatt-Smith, 2011). This observation seems to hold true in geography education as well. A study of K-12 classroom and large-scale geography assessments in the USA revealed that these assessments mainly test students' recall of geographical facts (Wertheim, Edelson, & The Road Map Project Assessment Committee, 2013).

Meaningful learning refers to an active construction of knowledge based on prior subject-specific knowledge and new information; it includes the cognitive processes of understanding, applying, analysing, evaluating and creating (Anderson, Krathwohl, et al., 2001). Meaningful learning, in this sense, is the opposite of rote learning, which stimulates the recall of knowledge. Furthermore, this approach to meaningful learning implies that students ‘can actively engage in the process of constructing meaning’ (Anderson, Krathwohl, et al., 2001, p. 65) and are able to apply or extend their specific conceptual and procedural knowledge.

The learning process benefits when multiple assessment approaches are used, including a variety of test items (Bell & Cowie, 2001; James & Gipps, 1998). These test items should be accompanied by clearly specified criteria for judging and marking (Harlen, 2005). Clearly specified criteria for judging and marking, or scoring rubrics, should be brought into line with students’ progress in learning. Assessment of students’ progress in learning “starts from the aim to discover what the learner knows, understands or can do” (Pryor & Crossouard, 2008, p. 5). Pryor and Crossouard defined this principle as divergent formative assessment. Divergent assessment can be distinguished from convergent assessment, which aims at identifying “if the learner knows, understands or can do a predetermined thing” (Pryor & Crossouard, 2008, p. 5). Although developed for formative assessment, this principle of divergent assessment could be relevant to summative assessment as well.

To date, little is known about the relationship between summative assessment in geography education in the Netherlands and its potential contribution to meaningful learning. Prior research by the authors has provided some insights into the relationship between summative assessments and meaningful learning in pre-vocational geography education in the Netherlands (Bijsterbosch, Van der Schee, & Kuiper, 2017; Bijsterbosch, Van der Schee, Kuiper, & Béneker, 2016). A content analysis of internal school-based examinations in pre-vocational secondary education showed that a majority of test items (62%) assess a form of remembering as a cognitive process. In the examinations, test items barely appealed to higher-order cognitive processes, such as evaluating and creating. The results of a questionnaire completed by teachers of pre-vocational geography education (n=74) showed that teachers rarely construct test items themselves and that they estimated the percentage of test items assessing meaningful learning to be higher (66%) than the results of the analysed school-based examinations (38%) showed. However, we must interpret these results cautiously because the group of respondents to the questionnaire was not the same as the group of teachers who completed the internal school-based examinations. Yet, these outcomes are relevant because they might indicate that teachers’ perceptions deviate from their practices.

The study in this paper is designed to examine the characteristics of feasible test items (and corresponding scoring rubrics) in school-based summative assessments that stimulate students’ learning in a meaningful way. Additionally, this study examines which strategies can feasibly and practically scaffold teachers to construct and judge these test items and scaffold students to cope with these test items. The research question guiding this study, therefore, is the following:

“What are the characteristics of feasible test items, scoring rubrics, instruments and strategies that contribute to meaningful learning in the context of internal school-based examinations in pre-vocational geography education in the Netherlands?”

To answer this research question, a designed toolkit was tested and evaluated in a small-scale case study with six geography teachers in pre-vocational education.

2. DESIGN OF THE TOOLKIT AND PROVISIONAL DESIGN PRINCIPLES

A toolkit on summative assessment and meaningful learning was designed to identify feasible test items that contribute to meaningful learning, feasible corresponding scoring rubrics, and feasible instruments and strategies to scaffold teachers and students on this issue. The toolkit served as input for an intervention to increase the use of test items – on school-based examinations – that contribute to meaningful learning and to support the professional growth of teachers regarding this aim.

This intervention is part of a design study on meaningful learning and internal school-based examinations in pre-vocational geography education in the Netherlands. The intervention is meant to contribute to the solution of the following problem: most test items used on school-based examinations assess a form of remembering, and teachers do not construct many test items themselves. Evaluation of the intervention must, first, provide insight into which test items and corresponding scoring rubrics are feasible and can be used on internal school-based examinations to increase the percentage of test items contributing to meaningful learning. Second, the intervention must also provide insight into which instruments and strategies for teachers and students are feasible and practical and how the professional growth of teachers – with respect to this identified problem – can be fostered. How, and to what extent, teachers' professional growth can be fostered will be reported in a separate study.

The toolkit for this intervention is based on provisional design principles that reflect the results from the first phase of the design study: the phase of analysis and exploration, which also included a literature review and an analysis of current practices. The provisional design principles for the toolkit are formulated in such a way that they reflect the aim of the toolkit - to provide test items, corresponding scoring rubrics, instruments and strategies that support the construction of test items that contribute to meaningful learning - and specify the characteristics of the elements of the toolkit. The toolkit contains three separate sections, and each section focuses on a part of the identified problem.

The first section of the toolkit contains examples of test items that appeal to distinct cognitive processes related to meaningful learning (for examples, see Bijsterbosch, 2018). Some of the examples come from existing examinations in the Netherlands and England; others were constructed by the researcher. The examples of the test items should give the participating teachers an idea of the characteristics of test items that support meaningful learning.
The characteristics of these test items are as follows:

- Test items contribute to meaningful learning when they appeal to cognitive processes that transcend rote learning; i.e., understanding, applying, analysing, evaluating and creating.
- Test items contribute to meaningful learning when they appeal to the integration of newly provided information and prior subject-specific knowledge.
- Test items contribute to meaningful learning when they stimulate divergent assessment; i.e., test items should aim to discover what the learner knows, understands or can do instead of assessing if the learner knows, understands or can do a predetermined thing.

The examples in the first section of the toolkit were chosen to be consistent with the learning objectives and should reflect the characteristics of the test items. To align the examples with the learning objectives, the examples were classified in a taxonomy table, which, for the purpose of this study, was slightly adjusted to the original taxonomy table of the revised taxonomy of Bloom (Anderson, Kratwohl, et al., 2001).
The first section contains examples such as,

- ‘constructed response tasks’ that appeal to different types of understanding, e.g. In which place is the average temperature in January lower, place A or B? Explain why the average January temperature is lower in this place.
- ‘executing familiar tasks’, that appeal to different ways of applying knowledge, e.g., “Calculate how many children per 1000 inhabitants were born in (year) in (country).” These items assess the ability to apply certain skills as part of procedural knowledge.
- ‘Differential items’. Differential items are characterized by a structure with multiple criterion-referenced tasks reflecting a sequence in the cognitive dimension. The structure of these items is based on Stimpson’s structure of ‘superitems,’ which are based on the SOLO-taxonomy (Stimpson, 1992). First, students need to describe what is displayed by a given figure or table. Second, students need to recall what they already know about this topic. Third, students have to relate the given information in the test item with the knowledge they already possess. Finally, students have to evaluate or generalize. Differential items, as such, are consistent with multiple levels of the cognitive dimension and the scoring rubrics.
- Examples of test items that appeal to higher-order cognitive processes, such as predicting and decision-making. These items combine the ability to solve a problem or to predict with more complex conceptual and procedural knowledge. These items are very suitable for use as ‘cases’ in test items.
- ‘short essays’. These test items are among the most challenging and complex items for students. Students usually have to evaluate, by attributing or criticizing the points of view of others, and provide reasonable arguments for their evaluations.

The second section of the toolkit contains a model with scoring rubrics and prescriptions regarding how to judge and mark these test items. This section of the toolkit is crucial. As Harlen (2005) noted, the extent to which the criteria used for judging and marking are clearly specified is a key variable when implementing test items contributing to meaningful learning. In particular, the more complex and open test items must be accompanied by clearly prescribed scoring rubrics for these items, based on the following characteristics:

- The model with scoring rubrics reflects the characteristics of the test items appealing to meaningful learning; i.e., whether a student is able to use the given information in the test items, whether a student is able to recall subject-specific knowledge, whether a student is able to integrate this existing subject-specific knowledge with the given information and, finally, whether what a student knows, understands or can do is assessed, instead of if the student knows, understands or can do a predetermined thing (principle of divergent assessment).
- The scoring rubrics are linked to the geographical conceptual knowledge in the objectives for the internal school-based examinations.
- The scoring rubrics include multiple levels to judge and mark students’ responses, which gives teachers the opportunity to reward what students know and to what extent they are able to integrate newly provided information with prior subject-specific knowledge.

To design a model to assess, judge and mark students' levels of performance in pre-vocational geography education in the Netherlands, several existing approaches from other researchers to identify levels of performance were compared. During early attempts to develop such a model, levels of performance were related to Piagetian stages of cognitive development. Peel (1972) distinguished three levels of students' responses, which were related to their age but also to other factors, such as students' background or the form of questioning. A more geographical attempt to define levels of performance in relation to the student's age – and an elaboration of Peel's model – was undertaken by Rhys (1972), who identified, in a pilot-study, four levels of understanding: (1) not reality-oriented, (2) single piece of evidence, (3) limited deductive analysis and (4) deduction from a guiding hypothesis. A similar approach to identifying levels of performance was launched by Biggs and Collis (1982). They introduced the SOLO taxonomy (Structure of the Observed Learning Outcomes), which was also based on Piaget's stages of cognitive development. An important diversion from Piaget's approach was their assertion that students' responses did not directly reflect their stage of development but rather a criterion-referenced level of performance. Other, more recent models for judging and marking students' understanding have been introduced by Entwistle and Smith (2002) and Smith (2002). Entwistle and Smith proposed a hierarchy of understanding that distinguished among mentioning, describing, relating, explaining and conceiving. This hierarchy has been reduced by Smith (2002), for modelling purposes, to three levels of understanding: unconnected understanding, descriptive understanding and explanatory understanding. Table 1 presents an overview of this comparison. The approaches are compared with each other in an attempt to distinguish general levels of performance.

Table 1. Comparison of attempts to define levels of performance

Level	Peel (1972)	Rhys (1972)	SOLO taxonomy (Biggs & Collis, 1982)	Entwistle & Smith (2002)	Smith (2002)
0		Not reality-oriented	Students are not able to answer in a structured way (pre-structural)	Mentioning: students are only able to provide incoherent bits of information without a structure	
1	Logically immature individuals tend to answer tautologically	Single piece of evidence, reality-oriented	Student's answer relates to one relevant feature (unistructural) or multiple but unrelated features (multistructural)	Describing: students are able to give brief descriptions of the topic, which they've derived from the provided material (tautological)	Unconnected understanding: students know facts but do not know how to relate them
2	The individual is dominated by the content	Limited deductive analysis, items of evidence combined	Students' answers reflect relational thinking (relational)	Relating: students give a personal explanation but without supportive arguments	Descriptive understanding: students do bring the facts together to form a description
3	Individual is able to think beyond the given content to evoke possible hypotheses from own experience	Deduction from a guiding hypothesis, comprehensive judgement		Explaining: students do use relevant evidence to come up with structured arguments	Explanatory understanding: students bring facts and descriptions together to form explanations
4			The student is able to combine the given information with prior knowledge to deduce more abstract principles and apply them to another situation (extended abstract)	Conceiving: students show individual conceptions, which they've developed through continuing reflection	

Although the compared models did not contain a uniform number of levels, it seems possible to identify five that reflect students' levels of performance. The five different levels - in fact, four levels, when the lowest level is not regarded as a performance level – reflect the characteristics of test items contributing to meaningful learning and have been transformed into a model to assess, judge and mark students' levels of performance (table 2). Students' answers can be marked at level 1, 'Repeating', when the answer of the student is merely tautological. The student describes geographical features that are already given by texts, figures or tables accompanying the test item. When a student is able to recall geographical knowledge related to the test item but does not really integrate this knowledge with the given information, the answer can be marked at the second level, 'elementary understanding'. At the third level, 'relating', the student shows the ability to relate the given information to pre-existing knowledge and thus the ability to describe and explain geographical relationships. Finally, at the highest level, 'Evaluating or Generalizing', the student demonstrates the ability to reason geographically. Geographical reasoning is more demanding for students because, to some extent, they have to evaluate or predict based on reasonable arguments derived from the geographical context and from geographical models or theories. Hooghuis et al. (2014, p. 243) defined geographical reasoning as 'reasonable reflective thinking about the relationship between mankind and environment focused on deciding what to believe or do in situations where location matters'. This highest level is only applicable when test items appeal to the skills of evaluating or creating.

Table 2. General model to judge and mark, including distinct levels of performance

Level	Description for each level
0	**Unstructured**: The student's response contains no substantive correct elements.
1	**Repeating:** The answer of the student is tautological. The student describes geographical features that are already given by texts, figures or tables accompanying the test item. The student does not integrate this information with pre-existing knowledge.
2	**Elementary understanding**: A student is able to recall geographical knowledge related to the test item but does not really integrate this knowledge with the given information. The student is not able to describe or explain geographical relationships.
3	**Relating:** The student shows the capability to relate the given information to pre-existing knowledge and thus the ability to describe and explain geographical relationships.
4	**Evaluating or Generalizing**: The student demonstrates the ability to reason geographically. The student not only demonstrates the ability to describe or explain geographical relationships but also demonstrates the ability to evaluate or predict based on reasonable arguments derived from the geographical context and geographical models or theories.

This designed model is a general model that can be applied to test items appealing to different types of meaningful learning. Yet, for each test item, the model has to be supplemented with specific geographical conceptual and procedural knowledge that the students are expected to demonstrate in their answers. For each test item, a separate marking scheme must be constructed based on the distinct levels of performance supplemented with the required geographical knowledge.

The third section of the toolkit contains instruments and coaching strategies to help teachers and students understand and answer test items appealing to meaningful learning. Students must become aware of teachers' expectations, which are reflected by the scoring rubrics. Awareness of scoring rubrics is quite essential to enhancing students' performance on test items stimulating meaningful learning (Black & Wiliam, 2011). To train and scaffold students, the instruments and learning strategies that are supposed to be effective have the following characteristics:

- The instruments scaffold students to answer the test items appealing to meaningful learning in accordance with the levels of the scoring rubrics.
- The strategies make students aware of the scoring rubrics for the test items appealing to meaningful learning.

One important and supposedly effective instrument for students is a flow chart (table 3) to help them understand these test items. The flow chart contains four steps. These steps are consistent with the scoring rubrics and, therefore, reflect the requirements of answering the test items.

Table 3. Flow chart with steps to answer a test item

Step 1	Which elements does your answer have to contain (a description, relationship, evaluation, prediction)?
Step 2	What do you already know about this topic?
Step 3	What kind of information is given by the texts, figures or tables accompanying the test item?
Step 4	Combine the knowledge you already have with the given information to answer the question. Make sure your answer includes the required elements (a description, relationship, evaluation, prediction).

A strategy that can scaffold students to answer the test items is the analysis of both 'good practices' and the corresponding scoring rubrics of test items that appeal to meaningful learning. Analysis of 'good practices' by students could help them to gain insight into the requirements of answering these test items. Other strategies that are suggested in the toolkit are classroom discussions about the test items and self- or peer assessment by students. These strategies should stimulate the formative use of summative assessment and give both teachers and students handholds for practice and evaluation.

3. METHOD

3.1 Outline of the case study

In the spring of 2016, a first prototype of the toolkit was evaluated by four experts: two experienced geography teacher educators and two educational scientists. An important element in this phase of the design study is formative evaluation by expert appraisal and interviews (McKenney & Reeves, 2012; Nieveen, 2010; Thijs & van den Akker, 2009). The evaluation, therefore, was formative, and it focused on the relevance, consistency and practicality of the toolkit. The outcomes of this evaluation were used to redesign the toolkit.

The redesigned toolkit was tested in a small-scale case study with six geography teachers from September until December 2016. All teachers worked in the third grade of pre-vocational education. In the third grade, the content of geography lessons pertains to three different areas of geography: sources of energy, poverty and wealth, and boundaries and identity. These three areas are part of the examination program for internal school-based geography examinations in pre-vocational secondary education and, as such, they are obligatory.

Participating teachers were recruited by the first author. Recruitment was conducted simply by sending e-mails with an invitation to teachers working in pre-vocational education in the vicinity of the institute of the first author. Approximately 50 teachers were directly invited to participate. Teachers were asked to participate in a teacher professional development program on internal school-based examinations and meaningful learning. Six teachers responded to the invitation and actually participated in this program.

The program consisted of three meetings of four hours each, followed by six weeks of collaborative practice. During these weeks, the teachers worked in pairs of two on constructing test items, and they practiced with their students. The program ended with a meeting to evaluate and discuss the results of what the teachers had done. The meetings were led by the first author of this article.

In the first meeting, the participating teachers discussed their beliefs and values regarding the aim of geography education, the purpose of summative assessment in geography education, and more specifically, the purpose of the internal school-based examinations. The aim was that teachers should become aware of their beliefs and values and the extent to which these beliefs and values influence how they think about the relationship among summative assessment, geography education and meaningful learning. The second step in the first meeting was to activate teachers' pre-existing knowledge regarding summative assessment and meaningful learning. The teachers received a few examples of test items from national exams and discussed what type of knowledge and cognitive processes were required for students to be able to answer these test items. Finally, the teachers received some instruction and materials regarding the relationship among summative assessment, test items and meaningful learning.

In between the first and second meetings, the teachers were asked to practice with the taxonomy table (as part of the instruction materials). They had to classify selected test items in this table, and the outcomes of this exercise were discussed at the beginning of the second meeting, which occurred two weeks later.

During the second meeting, the teachers were provided with some examples of test items appealing to understanding and evaluating. Demonstration of and instruction on these test items were followed by collaborative practice on the construction of test items. Furthermore, teachers practiced using the scoring rubrics on these test items. Practice exercises, in between the second and third meeting, were again part of the materials.

At the third meeting, test items that appeal to evaluating and creating, as well as the differential items, were introduced. The teachers were also instructed on strategies to scaffold students on how to address these test items. An important element of these strategies was the flow chart for students. Finally, the teachers received a flow chart for themselves on how to construct test items.

Over the six following weeks, the participating teachers worked in pairs of two on the construction of test items for the first internal school-based examination. The teachers constructed test items and provided each other with feedback. They also practiced with their students during the lessons. The constructed test items were discussed at the final meeting with the whole group. At the final meeting, the three sections of the toolkit were evaluated with the teachers as well.

3.2 Data collection

During the final stage of the study, the materials and the outline of the toolkit were evaluated with the teachers. The evaluation of the toolkit was formative and provided answers to the research question. First, the teachers completed a survey on the feasibility of test items on internal school-based examinations appealing to meaningful learning and on the feasibility of the scoring rubrics for these items. For each item and criterion, the teachers had to fill in – on a 1-to-5 point Likert-scale – the extent to which this item was feasible in relation to the intended outcomes. The teachers were also asked to elicit their scores.

The qualitative data that came from the elicitations were coded and analysed using a coding scheme that reflected the characteristics of the test items and scoring rubrics. Each guiding characteristic received a different code. When a teacher, for example, mentioned that a test item was highly valued because it enabled an assessment of what students had learned, this item was scored as contributing to divergent assessment (the third characteristic). The elicitations were independently scored by the first author and by another geography teacher educator. An interrater reliability test showed that Cohen's Kappa was 0.74, indicating a good level of agreement. After the coding, the outcomes were discussed with regard to how to interpret the statements of the teachers. Only the statements that had full agreement between the two scorers were used for further analysis.

The outcomes of the analysis were discussed with the whole group in a group interview. The group interview was semi-structured and focused on the question of which type of test items were feasible and to what extent the scoring rubrics were feasible. The main findings of the survey results were used as a guideline for the group interview.

Finally, classroom observations and subsequent mini-interviews with students were used to analyse to what extent and how the participating teachers practiced with their students. Students were observed while practicing with test items and strategies in the classroom. After the lessons, some students were interviewed regarding how they perceived the feasibility and practicality of the test items, the scoring rubrics, and the strategies that were supposed to scaffold them.

4. RESULTS

4.1 The feasibility of test items appealing to meaningful learning

The participating teachers were asked, by means of a survey, to indicate whether the examples of test items used in the instruction materials were feasible to appeal to meaningful learning and to use in summative assessments. Second, the teachers were asked to elicit why they believed that these test items were feasible or not feasible. Teachers' individual remarks were later discussed with the group of participating teachers.

Teachers were positive about the feasibility of the examples of test items, especially the ones that were more 'structured', such as the constructed response tasks appealing to different types of understanding or those pertaining to 'executing familiar tasks.' One of the reasons why teachers were positive about these test items was that these items have a clear structure, which makes it easier for students to know what is expected from them. The differential item was also valued as feasible. One of the teachers mentioned that the differential test item was possibly more directing but that this makes it easier for students to come up with a correct answer.

Test items appealing to higher-order cognitive skills, such as evaluating and creating, were valued positively by the teachers, yet these test items were considered less feasible and practical. One of the reasons why test items focusing on evaluating and creating were regarded as less feasible was the problem some students encountered when answering these test items.

One of the teachers mentioned that several students had difficulties answering these test items because there was confusion regarding what a correct answer would be. These difficulties emerged when the teacher evaluated students' answers at the debriefing.

A second reported reason why test items focusing on evaluating and creating were valued less positively was that these test items are more challenging for students whose literacy is below average. Writing essays is more difficult for these students, as one of the teachers mentioned. Third, some teachers mentioned that these test items required students to follow multiple steps, creating a risk that students would forget or skip steps. The fourth reason why teachers were less positive about the test items focused on evaluating and creating had to do with difficulties in scoring these test items. These difficulties were not always related to the content but sometimes to the perceived difficulty of scoring a test containing these items. As one of the teachers mentioned,

> "It is, of course, very idealistic and nice, but to score it is….eh, well now I am already busy for hours scoring a test."
> (Teacher A, group interview)

Overall, teachers were positive about the feasibility of the example test items in the instruction materials. However, they preferred the items that were more structured, and thus less demanding for students to answer and for teachers to score. One of the teachers also mentioned a positive effect of the summative assessment as a whole:

> "The whole set of test items now is more varied and challenging".
> (One of the teachers eliciting this aspect in the survey)

During the case study, the lessons of four participating teachers were observed when they practised the test items appealing to higher order cognitive skills with their students. After the lessons, mini-interviews with small groups of students (four or five) were held to reveal the extent to which the students thought that the test items were feasible. The students, who participated voluntarily, were asked to share why they thought that these test items were feasible or not feasible.

Most students thought that the test items were different from what they were used to, but not too difficult. As one of the students mentioned,

> "I did not find these test items very difficult, but it is another kind of questioning."
> (Student 2, mini-interview after lesson with teacher M)

Other students agreed on this point, especially with respect to the test items focusing on evaluating or creating. The students realized that these test items were sometimes more demanding in terms of meaningful learning:

> "You have to think deeper about the subject".
> (Student 1, mini-interview after the lesson with teacher H)

> "You have to add your own ideas, not just the information you have learned".
> (Student 3, mini-interview after lesson with teacher M)

Some of the students admitted that they encountered problems in answering the test items focused on evaluating or creating. For these students, answering these test items was more time consuming. Because they were afraid of running out of time during the test, these students were more critical with respect to the feasibility of these test items. According to some students, another reason why they were anxious about these test items was because they were uncertain how extensive their answers should be.

4.2 The feasibility of the scoring rubrics

The instruction materials included a general model for judging and marking answers at distinct levels of performance. The model was based on scoring rubrics and included four levels of performance. The teachers perceived the feasibility of this model as quite low, noting that they were confronted with several problems when trying to apply this model.

One of the problems was that teachers had difficulties scoring students' answers based on this model. It was especially difficult to determine students' levels of performance on test items that were more demanding in terms of evaluating or creating:

> "I have tried to apply the model in which you give marks based on the level of performance, but I stopped doing so at a certain time. It was so arbitrary. I could not explain to myself anymore what I had done."
> (teacher A, group interview)

Other reasons the teachers mentioned as to why the model with scoring rubrics was not feasible referred to the time-consuming process of marking these test items and the problems students would encounter when answering these test items.

Although the feasibility of the model with scoring rubrics was quite low, the teachers were much more positive about the individual principles that constituted the model with scoring rubrics. The two principles that were especially highly valued by the teachers were the students' ability to integrate pre-existing subject knowledge with given information and, second, the students' ability to show what they know, understand or can do instead of showing if they know, understand or can do a predetermined thing (principle of divergent assessment).

In the group interview, the teachers referred multiple times to this principle of divergent assessment. One teacher commented,

> "It really depends on how a student interprets the question….if he or she reasons in a certain way, the reasoning does not have to be wrong".
> (Teacher N, group interview)

Another teacher alluded to the notion of divergent assessment in summative assessment:

> "I like it when the test contains items that assess what a student knows instead of judging what he or she does not know".
> (Teacher A, group interview)

What also emerged from the group interview was that teachers not only apply this principle in their tests but also during their lessons:

> “To find out what students know instead of what they don’t know. I see myself doing this during my lessons, in the way I ask my students questions… I do not ask anymore ‘What is this?’, but ‘what do you know about this?’ ...Students find this more difficult, more difficult than recalling knowledge, but students respond to me that they understand the content better, because they have to explain it to me.“
> (Teacher Ar, group interview)

4.3 The feasibility and practicality of the instruments and strategies for teachers and students

From the survey and the interviews, two important issues emerged. The first issue was the use of the taxonomy table. The taxonomy table was introduced as an instrument to align the objectives for the internal school-based examinations with instruction and assessment. Most teachers were familiar with some type of taxonomy, but not the taxonomy table of the revised taxonomy of Bloom. The taxonomies that were used most by the teachers were Bloom’s original taxonomy and the so-called RTTI taxonomy. The RTTI taxonomy consists of four categories: remembering (R), executing a familiar task (T1), implementing an unfamiliar task (T2) and comprehension (I). This taxonomy is used frequently in Dutch secondary education.

The teachers reported that the taxonomy table was feasible. Yet, at the same time, some teachers reported that the practicality of the taxonomy table was less obvious. One teacher reported that the taxonomy table was quite overwhelming because of the number of options and amount of information it provided.

Other teachers reported that the taxonomy table helped them to become more aware of the objectives. One teacher mentioned that he purposely used the table to bring the constructed test in line with the requested objectives as written in the national guide for teachers regarding internal school-based examinations:

> “I am more aware now of the test items… I tried to use the objectives when I constructed the test items. I had the objectives open in another tab. I purposely worked towards these objectives, you know?”
> (Teacher M, group interview)

The second issue that emerged was that of the flow chart for students as a strategy of scaffolding. Teachers regarded this flow chart as a feasible instrument. In the opinion of one teacher, it helped the students learn how to answer the test items. Another teacher mentioned that the flow chart was very helpful in achieving the goal of divergent assessment. Some teachers also noted that the answers provided by the students were more structured when they used the flow chart. In the opinion of two teachers,

> “A number of students used the flow chart… then I could notice that the level of performance increased, the answers became more structured. I was quite happy with that”.
> (teacher An, group interview)

> “…you see much more structure in their answers.”
> (Teacher Ar, group interview)

Not only were the teachers positive about the flow chart, the students were positive about it as well. In their words, the flow chart was ‘handy’. It helped them to structure their answers and to create overviews. Talking about this issue, one of the students said,

> “The flow chart makes it easier to practice for the test. When you don’t have the flow chart, you will not be able to perform well on the test.”
> (Student 3, mini-interview after the lesson with teacher Ar)

Although most students were positive about the flow chart, some students were also anxious about using the flow chart during the test. In their opinion, it takes more time to answer the test items when they use the flow chart. As one student put it,

> “Probably it will cost you marks (overall) when you use the flow chart because you will run out of time and score less points on other test items.”
> (Student 1, mini-interview after the lesson with teacher A)

Other instruments or strategies did not emerge from the analysis as feasible instruments or strategies. Asked about other instruments or strategies, the students reported that they had not analysed ‘good practices’ of test items that focused on meaningful learning and corresponding scoring rubrics before they practiced with the test items. The students also reported that classroom discussions were not part of their teachers’ repertoire when scaffolding the students to practice the test items.

The observed lessons, in which the teachers practiced with the test items that focused on meaningful learning, confirmed this impression. Classroom discussions about students’ answers on the test items were not held. Although some teachers did some type of debriefing at the end of the lesson, in the observed lessons, little time was spent discussing the answers of the students and the reasons why they came up with these answers. The debriefing merely focused on what the ‘correct’ answer should have been. This way of debriefing seemed to be in line with students’ expectations. Most students reported during the mini-interviews that a recapitulation of the correct answer was the purpose of the debriefing. In their words, they were satisfied with the way the debriefing went because they wanted to know what the ‘correct’ answer was. Only some students reported that they were interested to hear what other students had answered and to learn from it.

In the group interview, the teachers admitted that they had spent less practice time with the students than was initially planned. The teachers also mentioned that they wanted to continue to practice the test items with their students. Some teachers, therefore, had already discussed this with their colleagues at school.

5. CONCLUSIONS AND DISCUSSION

A first outcome of this study suggests that teachers value pre-structured test items as most feasible for students in pre-vocational education. Test items focused on understanding and applying knowledge are considered to fall into this category. Test items that appeal to higher cognitive processes, such as evaluating and creating, are considered to be less feasible. When test items focused on evaluating and creating are desired, the application of differential items that assess a sequence of cognitive tasks seems to be most promising.

Teachers mentioned several reasons why they perceive the more open test items (those that focus on evaluating and creating) as less feasible. The first reason was that the students encountered problems in answering these items because they are more demanding in terms of literacy and structuring. Another reason why these items are considered to be less feasible is that the teachers had problems scoring these items. The feasibility of test items and scoring rubrics seems to depend, therefore, on students' literacy and ability to structure their answers, and on teachers' ability to score these items.

The teachers' valuation of the model with scoring rubrics was consistent with these outcomes. This model was perceived to be not very feasible due to problems the teachers encountered when scoring students' levels of performance, which might have been induced by a lack of understanding of the geography curriculum by the teachers (Brysch & Boehm, 2014). A second reason why this model was perceived as less feasible was that it seemed to give teachers the impression that scoring test items with this model was more time consuming. Teachers also indicated that they were not convinced of the feasibility of the highest levels of the model when scoring students' answers.

A second – and perhaps somewhat contradictory – outcome of this study, compared to teachers' valuation of the model with scoring rubrics, is that teachers appear to be positive about the constitutive principles of the model as a way to score test items. Especially the principle of divergent assessment - i.e., assessing what the student knows, understands or can do instead of assessing if the student knows, understands or can do a predetermined thing - was highly valued. The other constitutive principles of the model appeared to be feasible as well. Most teachers mentioned that they became more aware of how to score students' ability to recall pre-existing subject knowledge, to use new information in answering the test items and to integrate both types of knowledge in their reasoning. Teachers' valuation of these principles was quite strongly related, however, to pre-structured test items.

A third important outcome of this study indicates that scaffolding students with strategies such as the flow chart is very helpful. Both the teachers and students mentioned that the flow chart helped the students to structure their answers. The quality of students' answers was perceived to increase when students used the flow chart to answer test items focused on meaningful learning.

What is unknown is whether the flow chart helps students to enhance their geographical understanding. Although both teachers and students mentioned that the flow chart helped the students to structure their answers - and even that the quality of the given answers seemed to improve - this study has not determined whether this also means that students were better able to demonstrate a grasp of cause and effect (Peel, 1972), to make a systematic analysis of cases not directly related to their own experience (Rhys, 1972), or to make sense or give meaning to something (Bennetts, 2005). Future research should provide more insight into the potential of the flow chart to enhance students' performance with respect to geographical understanding.

From the survey and interviews, it emerged that teachers hardly used the other suggested instruments and strategies from the toolkit. The observed lessons confirmed the impression that teachers did not really practice with strategies such as analysing examples of answers to test items or classroom discussions. Although there was some debriefing at the end of the lessons, the observed lessons did not really include these strategies. One of the reasons could be that the teachers, as they reported, had spent less time on practice with students than expected.

These findings raise intriguing questions regarding what characterises feasible test items and corresponding scoring rubrics focused on meaningful learning. Should summative assessments in pre-vocational geography education intended to stimulate meaningful learning focus on pre-structured test items due to the problems the students and teachers encountered with the more open test items? Or could these problems be overcome when both students and teachers are scaffolded more and over a longer period of time?

To realize the full potential of test items in summative assessment that contribute to meaningful learning, mutual understanding between students and teachers regarding the intended outcomes is important (Entwistle & Smith, 2002). Mutual understanding becomes even more important when the test items are different from what the students are used to. To enhance mutual awareness between students and teachers, the outcomes of this study suggest that instruments such as the flow chart could be helpful.

This flow chart seems to have the potential to structure students' answers. Students are forced to construct their answers based on recalling what they have learned and to integrate this with the new information in the test items. In this sense, the flow chart could help students to actively construct knowledge and give meaning to it, which is one definition of what meaningful learning should be (Anderson, Kratwohl, et al., 2001). The flow chart also seems to have the potential to make students more aware of teachers' expectations concerning the intended outcomes, as the flow chart was consistent with the constitutive principles of the model used to score test items. As mentioned above, it is still uncertain whether the flow chart also has the potential to enhance students' geographical understanding.

What is unknown from this study is how teachers can become more confident when applying the model to the scoring of more open test items, namely those that focus on evaluating or creating. If they practiced more often with the model when scoring students' performance, could teachers become more confident when applying these principles and the model to test items focusing on evaluating or creating? Or, can teachers' self-efficacy in applying this model be enhanced if they recognize that their scoring of these test items is in line with the scoring of their colleagues? This is also an important issue for future research.

Another interesting finding from this study suggests that teachers integrate their summative assessment practices with the more formative purposes of assessment when they apply the constitutive principles for scoring test items. Several teachers mentioned, for instance, that they not only tried to apply the principle of divergent assessment in their summative assessments but in their classroom practices as well. Students' responses during classroom practice seem to have enforced teachers' valuation of this principle.

Consequently, application of this principle seems to have brought summative assessment more in line with formative assessment. Formative assessment is often considered to be more effective at stimulating students' learning (Sluijsmans, Joosten-ten Brinke & Van der Vleuten, 2013). The results of this study seem to enforce the idea, however, that the application of principles for summative assessment has the potential to bridge the gap with formative assessment and, as such, contribute to and stimulate students' learning as well. To ensure that summative assessment contributes to meaningful learning, the results of this study also suggest that more time is needed for teachers to practice and to apply other instruments and strategies.

Some final remarks should be made about this study. The current study is limited in several ways. First, only six teachers in pre-vocational education participated. With this small sample size, caution must be applied to the results. Additionally, the selection of participating teachers and students was not fully at random. Second, the teachers participated for a period of three months. It would be interesting to see what the results would be if teachers were to practice and were scaffolded over a longer period of time.

A remark must also be made regarding this type of research. The qualitative method used in this study relies heavily on what teachers and students reported in the survey and the interviews. Although this method is suitable to explore the reasons for teachers' and students' remarks, more research is needed to verify the results from these two groups.

There are still many unanswered questions about the way teachers and students can be scaffolded to construct, score and answer test items in pre-vocational geography education in ways that contribute to meaningful learning. An important issue is to what extent teachers will become able to score these test items reliably, particularly the more open and complex items. A second issue is to what extent teachers' practices with respect to summative assessment will change, particularly over a longer period of time. Finally, future research is needed to determine how and to what extent teachers' knowledge, beliefs and values interfere with the previous two issues.

REFERENCES

Anderson, L. W. (Ed), Krathwohl, D. R. (Ed), Airasian, P.W., Cruikshank, K. A., Mayer, R. E., Pintrich, P.R., Raths, J. & Wittrock, M. C. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* (*Complete Edition*). New York: Longman.

Bell, B., & Cowie, B. 2001. The characteristics of formative assessment in science education. *Science education:* 85 (5): 536-553.

Bennetts, T. 2005. Progression in Geographical Understanding. *International Research in Geographical and Environmental Education:* 14 (2): 112-132.

Biggs, J. B., & Collis, K. F. 1982. *Evaluating the quality of learning.* New York: Academic Press.

Bijsterbosch, H. 2018. *Professional development of geography teachers with regard to summative assessment practices.* Enschede: Ipskamp publishing.

Bijsterbosch, H., Van der Schee, J. A., & Kuiper, W. 2017. Meaningful learning and summative assessment in geography education: An analysis in secondary education in the Netherlands. *International Research in Geographical and Environmental Education:* 26 (1), 17-35.

Bijsterbosch, H., Van der Schee, J. A., Kuiper, W., & Béneker, T. 2016. Geography teachers' practices towards summative assessments: a study in pre-vocational education in the Netherlands. *Review of International Geographical Education Online:* 6 (2), 118-134.

Black, P., & Wiliam, D. 1998a. Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice:* 5 (1): 7-74.

Black, P., & Wiliam, D. 1998b. *Inside the black box: raising standards through classroom assessment*. London: NferNelson.

Black, P., & Wiliam, D. 2011. Assessment for Learning in the Classroom. In *Assessment and Learning (2nd ed.)*, ed. J. Gardner, 206-231. London: Sage Publications Ltd.

Brysch, C. P., & Boehm, R. G. 2014. Online professional development in geography: The learning cluster method and teacher leadership. *European Journal of Geography:* 5 (1): 62-69.

Entwistle, N., & Smith, C. 2002. Personal understanding and target understanding: mapping influences on the outcomes of learning. *The British journal of educational psychology:* 72 (3): 321-342.

Harlen, W. 2005. Teachers' summative practices and assessment for learning – tensions and synergies. *Curriculum Journal:* 16 (2): 207-223.

Hooghuis, F., van der Schee, J., van der Velde, M., Imants, J., & Volman, M. 2014. The adoption of Thinking Through Geography strategies and their impact on teaching geographical reasoning in Dutch secondary schools. *International Research in Geographical and Environmental Education:* 23 (3): 242-258.

James, M., & Gipps, C. 1998. Broadening the basis of assessment to prevent the narrowing of learning. *Curriculum Journal:* 9 (3): 285-297.

Klenowski, V., & Wyatt-Smith, C. 2011. The impact of high stakes testing: the Australian story. *Assessment in Education: Principles, Policy & Practice:* 19 (1): 65-79.

McKenney, S. E., & Reeves, T. C. 2012. *Conducting educational design research*. New York: Routledge.

Nieveen, N. M. 2010. Formative Evaluation in Educational Design Research. *In An Introduction to Educational Design Research,* eds. T. J. Plomp & N. M. Nieveen, 89-103. Enschede: SLO.

Peel, E. A. 1972. The quality of understanding in secondary school subjects. *Educational Review*: 24 (3): 163-173.

Pryor, J., & Crossouard, B. 2008. A socio-cultural theorisation of formative assessment. *Oxford Review of Education:* 34 (1): 1-20.

Rhys, W. T. 1972. Geography and the adolescent. *Educational Review:* 24 (3): 183-196.

Sluijsmans, D., Joosten-ten Brinke, D., & Van der Vleuten, C. 2013. *Toetsen met leerwaarde*. *Een reviewstudie naar de effectieve kenmerken van formatief toetsen* [Formative assessment. A review study on characteristics of formative assessment]. Den Haag: NWO-PROO.

Smith, C. A. 2002. Supporting teacher and school development: learning and teaching policies, shared living theories and teacher-researcher partnerships. *Teacher Development:* 6 (2), 157-179.

Stimpson, P. (1992). Assessment in Geography: an evaluation of the SOLO Taxonomy. *In Empirical Research and Geography Teachin*g, eds. H. Schrettenbrunner & J. van Westrhenen, 157-177. Utrecht; Amsterdam: Koninklijk Nederlands Aardrijkskundig Genootschap; Centrum voor Educatieve Geografie Vrije Universiteit.

Thijs, A., & van den Akker, J. (Eds.). 2009. *Leerplan in ontwikkeling* [Curriculum Development]. Enschede: SLO.

Wertheim, J. A., Edelson, D. C., & The Road Map Project Assessment Committee 2013. A Road Map for Improving Geography Assessment. *The Geography Teacher:* 10 (1): 15-21.