The Impact of National Examinations on Geography Teachers' Assessment practices in the Netherlands

Erik Bijsterbosch
Windesheim University of Applied Sciences, Zwolle, the Netherlands

Abstract

Geography teachers' school-based (internal) examinations in pre-vocational geography education in the Netherlands appear to be in line with the findings in the literature, namely that teachers' assessment practices tend to focus on the recall of knowledge. These practices are strongly influenced by national (external) examinations. This paper provides empirical evidence about the impact of the national examinations on internal assessment practices. An analysis of test items in national examinations from 2015, 2016 and 2017 shows that the majority of these items focus on remembering and are in a format that can be marked reliably. Teachers' tendency to copy these formats in their school-based assessment raises questions regarding validity. This paper explores these concerns and contributes to the discussion on effective assessment in secondary school contexts.

Introduction

Emphasis is often placed on the negative impact of teachers' summative assessment practices on students' learning (Harlen, 2004). In geographical education, the tendency to focus on the recall of knowledge has been identified. For example, The Road Map Project (2013) revealed that the majority of large-scale assessments in the United States tested students' recall of geographical facts (Wertheim, Edelson, & The Road Map Project Assessment Committee, 2013). This tendency could be a result of the demand to produce reliable test items "that are relatively closed in nature and require minimal or no subjective judgement. In short, they are safe" (Stimpson, 2006, p. 79).

The pressure to produce reliable results is stronger when systems are based on high-stakes tests. The results of these tests are often used for purposes of accountability which can lead to a teaching to the test strategy. Klenowski and Wyatt-Smith (2011) in their analysis of the impact of high-stakes testing in Australia found that many schools used the teaching to the test strategy to improve literacy and numeracy. Equally, Kuiper, Van Silfhout, and Trimbos (2017) in the Netherlands, in their reflections on the relationship between curriculum and assessment, emphasised the *pre-shadowing* effect of national examinations on teachers' classroom assessment practices and the enacted curriculum.

The Dutch national examinations impact on the enacted curriculum and internal assessment practices and this can cause problems with regard to validity. As Kuiper et al. (2017, p. 86) stated, "what is tested makes beloved and what stays untested makes unbeloved". The enacted curriculum often reflects the content and structure of the national examination rather than the aims and objectives of the intended curriculum. This situation is not typically Dutch, as it can also be found in other countries. Spielman (2017), Ofsted's chief inspector in England, recently reported on this issue:

> There need be no tension between success on these exams and tests and a good curriculum. Quite the opposite. A good curriculum should lead to good results. However, good examination results in and of themselves don't always mean that the pupil received rich and full knowledge from the curriculum. In the worst cases, teaching to the test, rather than teaching the full curriculum, leaves a pupil with a hollowed out and flimsy understanding.

To prevent summative classroom assessment practices being too focused on national examinations, teachers should consider the validity of assessment items. Assessment programs should reflect the full content and objectives of the intended curriculum. A dependability approach, as suggested by Harlen (2005), could contribute to this outcome. Harlen (2005, p. 213) defined dependability as the sum of reliability and validity:

> The interdependence between the concepts of reliability and validity means that increasing one tends to decrease the other. Dependability is a combination of the two, defined in this instance as the extent to which reliability is optimized while ensuring validity. This definition prioritizes validity, since a main reason for

using teachers' assessment rather than depending entirely on tests for external summative assessment is to increase the construct validity of the assessment.

Up to now, far too little attention has been paid to geography teachers' summative assessment practices in the Netherlands. Although data on students' results in national examinations are collected each year, there have only been a few investigations into geography teachers' assessment practices and how such practices are influenced by the national examinations. This paucity of evidence is in line with the lack of published research pertaining to geographical education and assessment in general (Lane & Bourke, 2017).

This paper contributes to the discussion of the impact of high-stakes tests on geography teachers' classroom assessment practices. An analysis of test items in Dutch national examinations will be compared with results from prior studies of teachers' assessment practices. The consequences of these results and their implications for assessment in schools will then be discussed.

Background

Structure of geography and examinations in the Netherlands

Pre-vocational education in the Netherlands is a four-year course and is one of the possible tracks in secondary education. The other two tracks are a five-year general education track and a six-year pre-university education track. Students enter secondary education at the age of twelve. Geography is only compulsory in the first two years of secondary education. In the final two years, students choose six or seven subjects as part of their examination program. For those who choose geography the program consists of two parts: a national examination and an internal school-based examination. Both contribute 50% to the overall result at the end of secondary education.

Since 2013, the examination program in pre-vocational geographical education has contained six content domains: (1) Sources of energy, (2) Poverty and wealth, and (3) Boundaries and identity (internal school-based examinations) (4) Weather and climate, (5) Water, and (6) Population and place (domains of the national examination). A separate domain with specifications for geographical skills and methods is also included.

Prior research has highlighted two problems with regards to the alignment of internal (school-based) and external examinations. First, school-based examinations are dominated by the content of the national examination program. Results of a questionnaire conducted by Noordink, Oorschot, and Folmer (2017) showed that three-quarters of teachers in pre-vocational geographical education assessed the content domains of the national examination program in their school-based examinations (Noordink et al., 2017). These results were confirmed by Bijsterbosch, Van de Schee, Kuiper, and Béneker (2016) who found an even higher proportion of teachers structuring their assessment in this way. The second problem relates to the format of the internal school-based examinations. In a questionnaire by Bijsterbosch et al. (2016) geography teachers (n=74) responded that the purpose of internal examinations was preparation for the external assessment. The majority of these teachers believed that using a similar test format (multiple-choice questions or short, constructed responses) benefited students in this preparation. They also believed that these formats supported greater reliability in marking. Open test items demanding longer answers from students were less common. This suggests that teachers were more concerned with reliable test results than they were with the validity of their school-based examinations. These results are consistent with the findings of Harlen (2005), Black, Harrison, Hodgen, Marshall, & Serret, (2010) and Bijsterbosch, Van der Schee, and Kuiper (2017) regarding the reliability and validity of internal assessment practices. One of the consequences of these practices is that geography teachers' summative assessments in pre-vocational geographical education in the Netherlands do not always initiate meaningful ways of learning. More than 60% of these test items focus on recall of knowledge only (Bijsterbosch et al., 2017). Test items focusing on higher-order cognitive skills, such as evaluating or creating, are rarely included in these examinations.

These findings deviate from teachers' stated goals for geographical education. During the panel interviews, teachers confirmed that their goals went beyond the recall of knowledge (Bijsterbosch et al., 2017). Most teachers felt that geographical education should aim to support deep understanding and should scaffold students to become citizens who can make informed decisions about their world in the future. This raises serious questions about the impact of the national examinations on the design of school-based assessment and the accuracy of teachers' perceptions of the content domains of the national assessment.

Content analysis of national examinations

To identify the extent to which the national examinations reflect teachers' perceptions, the national examinations from 2015, 2016 and 2017 were analysed with regard to geographical knowledge and cognitive dimensions. Both dimensions were scored using the revised taxonomy by Bloom (Anderson, Krathwohl et al., 2001). In this table, the knowledge dimension consists of four categories. The first category is factual knowledge containing knowledge of specific details and elements, and knowledge of simple concepts. The second category is conceptual knowledge, which comprises knowledge of geographical principles or relationships between concepts. The third category, procedural knowledge, focuses on geographical skills and methods. The final category, metacognitive knowledge, includes knowledge of learning strategies.

The second dimension of the taxonomy consists of five cognitive processes: remembering, understanding, applying, evaluating and creating. Remembering refers to students' abilities to recall knowledge. Understanding is a more comprehensive category, containing cognitive processes such as explaining or inferring. The third cognitive process, applying, refers to students' abilities to choose and apply geographical skills. Evaluation requires students to attribute or critique the opinions of others, or give an opinion themselves. Finally, creating refers to the processes of developing a new idea or solution. In the analysis of test items, an important distinction was made between remembering and the other cognitive processes. Test items focusing on understanding, applying, evaluating and creating must contain new information. Otherwise, it is assumed that students will be able to answer the task correctly solely based on what they have already learned.

All test items in the national examinations (N=133) were scored by the author. A random selection of twenty-six test items were scored by another geography teacher educator in order to achieve inter-coder agreement. An interrater reliability test showed that Cohen's Kappa was 0.77 ($p<0.001$) for the scores of the test items in the distinct cells of the taxonomy table, which indicates a substantial agreement. The results of the analysis (Table 1) show that the majority of test items focus on remembering. Sixty per cent of items analysed assessed the recall of conceptual knowledge (see Appendix A, Examples 1, 2 and 4). The second most important category is 'understanding conceptual knowledge' (Appendix A, Example 3). Only seven per cent of the test items focused on applying (Appendix A, Example 5) and there were no examples of items assessing evaluation or creation of knowledge.

The examinations were also analysed by the assessment developers, the National Institute for Educational Measurement (Cito). These analyses are published on-line (Cito, 2015, 2016, 2017) and include psychometric indicators, such as the P-value or Rit/Rir-value of the test items. In their analysis Cito assigned each item to a category. An overview of the number and percentages of test items assigned to each category is provided in Table 2. The first three categories refer to the types of test item – open, multiple choice and pre-structured – while the remaining categories denote the targeted cognitive process: items with *statements*, *mention/cite* items, and *explanation* items. The definitions of these distinct categories come from Cito, but have been translated by the author. Note that the categories can overlap – a test item can be pre-structured and also include statements. Appendix A contains examples of test items in each category.

Table 1: Cumulative percentages of test items (2015, 2016 and 2017) in the taxonomy table.

Knowledge Dimension	Cognitive Process Dimension					
	Remember	Understand	Apply	Evaluate	Create	Total
Factual Knowledge	11					11
Conceptual Knowledge	49	33				82
Procedural Knowledge			7			7
Metacognitive Knowledge						
Total	60	33	7			100

Table 2: Numbers and percentages of test items in national examinations according to Cito in 2015, 2016 and 2017.

	2015 (n=43)	2016 (n=45)	2017 (n=45)
Open tasks	21/49	30/67	24/53
Multiple choice	14/33	15/33	15/33
Pre-structured	8/19		6/13
Statements	6/14	6/13	4/9
Mention/cite	12/28	16/36	11/24
Explanation	11/26	11/24	16/36

Conclusions/discussion

The outcomes of the content analysis of national examinations by the author is in line with the outcomes of the previous analysis of school-based examinations (Bijsterbosch et al., 2017). The majority of test items focus on the recall of knowledge. Higher-order cognitive processes, such as evaluating and creating, are completely absent.

A comparison with the analysis by the test developer was more complicated because Cito used distinct categories for the cognitive dimension. These categories do not match the categories of the revised taxonomy (Anderson et al., 2002), nor do they reflect the categories that are prescribed in the examination program (describe, explain, evaluate, problem solve, predict). The key question is whether the national examinations reflect the requirement of the syllabus for students to demonstrate higher-order cognitive processes. The content analysis of the external examination outlined above suggests that this is not the case. This raises questions regarding the construct validity of the test items in the examination. In this regard, some critical comments can be made about the format of test items. Cito distinguished three categories of test items – open, multiple choice and pre-structured. Most of the items in the external examinations adopted an open format. These tasks could best be defined as constructed response tasks that require a short answer (see Appendix A Example 2). Multiple-choice and pre-structured tasks were also common. These item formats are preferred because they provide reliable results. While there is nothing wrong with striving for reliability, this focus should not be at the expense of content and construct validity.

Greater attention to the validity of the examinations, both national and school-based, is needed. This problem has been previously highlighted in the literature. Kuiper et al. (2017) identified the need to ensure a balance between assessment reliability, validity and transparency. Kuiper et al. (2017) also drew a distinction between broad curriculum goals and specific achievement standards. The curriculum goals are expressed in generic terms and provide schools and teachers with choice regarding the selection of topics and learning objectives. The achievement standards are a set of attainment targets that students are supposed to demonstrate and, as such, are fundamental for both the internal and external examination program.

Ideally, the exam content and structure should align with the curriculum goals and the achievement standards. This does not mean that the exams fully reflect the content and objectives of the entire curriculum; rather, the knowledge and cognitive processes students are supposed to demonstrate in the exams are in line with the broader educational goals of the subject – in the context of this paper, the educational goals for geographical education. The exams are supposed to follow the content and objectives of the curriculum, not vice versa (Kuiper, 2017). This sequence becomes problematic when teachers *teach to the test* and exams dictate the enacted curriculum. This has the effect of widening the gap between the intended and enacted curricula.

To bridge this gap, greater focus on constructive alignment is necessary. Constructive alignment focuses on the relationship between educational goals, instruction, pedagogy, assessment and achievement standards. These five aspects should be in line with each other. An approach based on powerful knowledge, as suggested by Lambert (2011) and others, might be helpful in achieving this. According to Lambert, three domains are essential for powerful knowledge:

1. deep descriptive and explanatory world knowledge;
2. development of relational thinking in geography; and
3. an enhanced propensity to think about how places, societies and environments are made.

Powerful knowledge, in this sense, is strongly connected to a capabilities approach. A

capabilities approach invites teachers and curriculum leaders to reflect on how education contributes to human autonomy and potential (GeoCapabilities, 2016). This approach also allows teachers to connect the subject-specific knowledge to the goals for which to strive. As Lambert (p. 258) states:

> A 'capabilities' geography expresses geography in terms of educational goals. The curriculum content, beyond the statutory knowledge requirements (including possibly a core knowledge sequence), still has to be selected. But the goals articulate what we are trying to achieve with young people: an improved knowledge and understanding of the world and their relationship with it.

In geographical education, this approach is helpful in determining what the content and objectives should be and which pedagogies should be applied to meet these goals. This approach would also be helpful in order to align the ultimate goals in geographical education with geography examinations, both national and internal.

A dependability approach, with a strong focus on constructive alignment that rebalances the focus on reliability with validity in the construction of examinations, is required to create meaningful examinations in the Netherlands. A rethinking of the examination program, and the relationship between content and purpose of school-based and national examinations, is also necessary. The distinction between a school-based examination programme and a national examination programme has led to undesirable effects, as described above. Reconsidering this distinction therefore seems to be necessary.

Another issue worth reconsidering is whether the current content of the examination program is appropriate. Noordink et al. (2017, p. 13) note that many teachers of pre-vocational geography believe that the examination program is "overloaded" and that the range of topics and regions should be reduced. Progression in geographical understanding is often considered to reflect increasing breadth, increasing depth, a move from the concrete to the abstract, and the use of a wider range of techniques (Taylor, 2013). The current focus on breadth in the examination program might be at the expense of increasing depth. This may also promote a focus on the recall of knowledge. A less overloaded examination program, therefore, might be required. There is an urgency to rethink the design of the examination program, to ensure that both school-based and national examinations contribute in meaningful ways of learning in geography. The promotion of more meaningful ways of teaching, learning and assessing geography is a responsibility of the entire geography community in the Netherlands.

References

Anderson, L. W. (Ed.), Krathwohl, D. R. (Ed.), Airasian, P., 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.

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*, 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*, 118–134.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*, 215–232.

Cito. (2015). *Test and item analysis pre-vocational geography education 2015*. Retrieved from https://www.cito.nl/

Cito. (2016). *Test and item analysis pre-vocational geography education 2016*. Retrieved from https://www.cito.nl/

Cito. (2017). *Test and item analysis pre-vocational geography education 2017*. Retrieved from https://www.cito.nl/

GeoCapabilities. (2016). Retrieved from http://www.geocapabilities.org/training-materials/

Harlen, W. (2004). A systematic review of the evidence of the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes. In *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

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

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

Kuiper, W. (2017). Ruimte, richting en ruggesteun [Space, direction and scaffolding]. In E. Folmer, A. Koopmans-Van Noorel, & W. Kuiper (Eds.), *Curriculumspiegel 2017* [Curriculum mirror 2017] (pp. 11–27). Enschede: SLO.

Kuiper, W., Van Silfhout, G., & Trimbos, B. (2017). Curriculum en toetsing [Curriculum and assessment]. In E. Folmer, A.Koopmans-Van Noorel, & W. Kuiper (Eds.), *Curriculumspiegel 2017* [Curriculum mirror 2017] (pp. 83–109). Enschede: SLO.

Lambert, D. (2011). Reviewing the case for geography, and the 'knowledge turn' in the English national curriculum. *Curriculum Journal*, *22*, 243–264.

Lane, R., & Bourke, T. (2017). Assessment in geography education: A systematic review. *International Research in Geographical and Environmental Education*, 1–15. https://www.tandfonline.com/doi/full/10.1080/10382046.2017.1385348.

Noordink, H., Oorschot, F., & Folmer, E. (2017). *Monitoring vernieuwde examenprogramma aardrijkskunde vmbo. Samenvattend eindrapport* [Monitoring new geography examination programme in pre-vocational education. Summarising report]. Enschede: SLO.

Spielman, A. (2017). *HMCI's commentary: recent primary and secondary curriculum research October 2017*. Retrieved from https://www.gov.uk/government/speeches/hmcis-commentary-october-2017

Stimpson, P. (2006). Changing assessments. In J. Lidstone & M. Williams (Eds.), *Geographical education in a changing world: Past experience, current trends and future challenges* (pp. 73–84). Dordrecht: Springer.

Taylor, L. (2013). What do we know about concept formation and making progress in learning geography? In D. Lambert & M. Jones (Eds.), *Debates in geography education* (pp. 302–313). London: Routledge.

Wertheim, J. A., Edelson, D. C., & The Road Map Project Assessment Committee (2013). A road map for improving geography assessment. *The Geography Teacher*, *10*, 15–21.

Appendix A

1. Example of a test item that has been classified as testing 'remembering conceptual knowledge' by the author and as a 'mention/cite' item by Cito:

   The construction of the Aswan Dam in South Egypt has advantages and disadvantages for the people living in the area of the lower reaches of the Nile.

   Describe an advantage for the people living in the area of the lower reaches of the Nile.

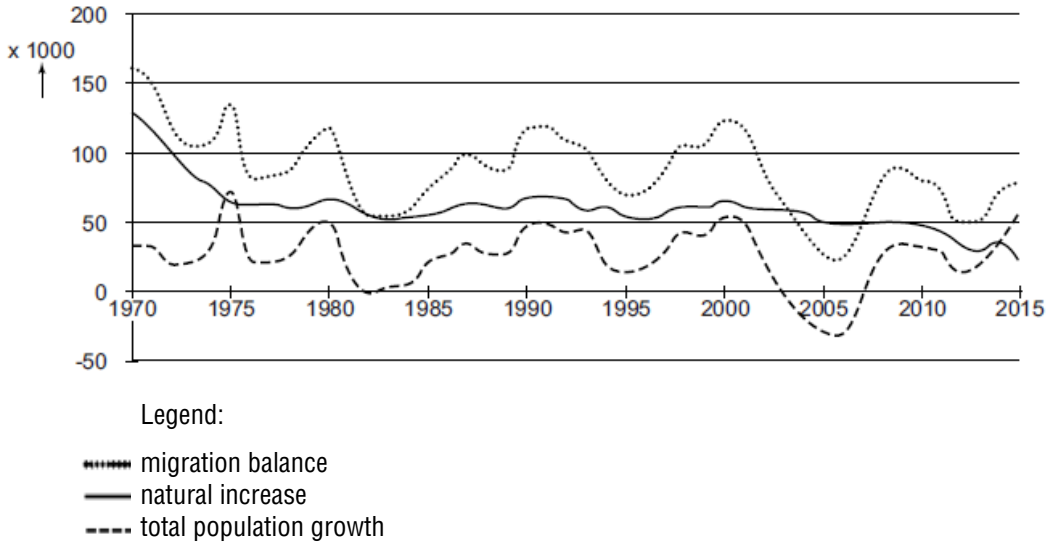
2. Example of a test item that has been classified as testing 'remembering conceptual knowledge' by the author and as an 'open task and explanation' item by Cito:

   Tornados and hurricanes are both manifestations of extreme weather conditions. In general, hurricanes lead to more victims than do tornados. Despite this, hazard management for tornados is more difficult than it is for hurricanes.

   Mention a reason why this is the case.

3. Example of a test item that has been classified as testing 'understanding conceptual knowledge' by the author and as a 'pre-structured and statement' item by Cito:

   Figure 32 Population development in the Netherlands, 1970–2015



Study Figure 32.

*Below are thr*ee statements, based on Figure 32.

Statement 1: In 2015, more people died than were born.

Statement 2: In 2006, the total population growth was less than was the natural increase.

Statement 3: Between 1970 and 2015, the Dutch population mainly grew because of natural increase.

Write the number*s 1, 2 and 3 on your paper and write whether the statement is correct or incorrect*

4. Example of a test item that has been classified as testing 'remembering conceptual knowledge' by the author and as a 'multiple-choice and statement' item by Cito:

   Two students make a statement about air pressure.

   Statement 1: The tighter the packing of the isobars, the weaker the wind blows.

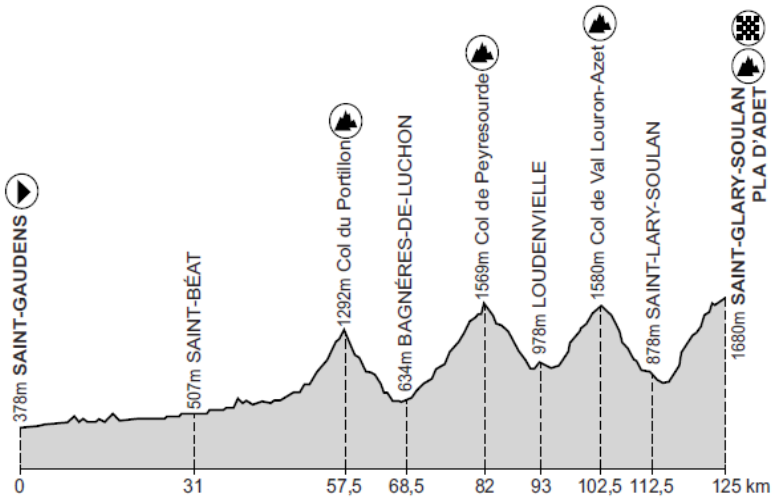
   Statement 2: In high-pressure areas, the air rises, thus creating a greater chance of precipitation.

   Which is correct?

a. *Only statement 1 is correct*
b. *Only statement 2 is correct*
c. *Both statements are correct*
d. *Both statements are incorrect.*

5. Example of a test item that has been classified as testing 'applying procedural knowledge' by the author and as a 'multiple-choice' item by Cito:

Figure 4. Stage 17 in the Tour de France 2014



In stage 17, the cyclists had to climb. That day, the weather was calm. The temperature in Saint-Béat was 24 degrees Celsius. How many degrees Celsius lower was the temperature at the top of the Col de Peyresourde?

a. *Approximately 0,6 degrees Celsius*
b. *Approximately 1,0 degrees Celsius*
c. *Approximately 6,0 degrees Celsius*
d. *Approximately 10,0 degrees Celsius.*