Professional development of geography teachers with regard to summative assessment practices

Erik Bijsterbosch

Professional development of geography teachers with regard to summative assessment practices

Title: Professional development of geography teachers with regard to summative assessment practices

ISBN: 978-94-028-1020-2

Cover design: Yvonne van Staalen

Publisher: Ipskamp Printing

Professional development of geography teachers with regard to summative assessment practices

Professionele ontwikkeling van aardrijkskundedocenten op het gebied van summatieve toetsing

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen
op woensdag 6 juni 2018 des middags te 12.45 uur

door

Henricus Bijsterbosch

geboren op 19 oktober 1966
te Heerde

Promotoren: Prof.dr. J.A. van der Schee
Prof.dr. W.A.J.M. Kuiper
Prof.dr. T. Béneker

Table of contents

Chapter 1

Introduction

1.1 Teachers' assessment practices

Teachers' competence plays a vital role in students' learning. Teachers can affect students' learning and can be the greatest contributors to it. What is of the greatest importance in this regard is the teachers' "mind frame within which they ask themselves about the effect that they are having on student learning" (Hattie, 2012, p. 15).

Part of teachers' competence is their assessment literacy. Teacher assessment literacy consists of the essential knowledge and skills to construct, score and administer assessments in order to use students' results to make sound decisions about students' learning (Brookhart, 2011; DeLuca, LaPointe-McEwan, & Luhanga, 2016; Xu & Brown, 2016). Teachers' assessment literacy is, therefore, strongly connected to teachers' assessment practices and students' learning.

An important finding from research on the relationship between assessment and learning, is that teachers' classroom assessment practices do not always contribute to types of learning that have been identified as deep or meaningful (Black & Wiliam, 1998b; Harlen, 2005; James & Gibbs, 1998). Instead, teachers' classroom assessment practices seem to enforce rote learning, a type of learning that is considered to be less beneficial in terms of learning and progress. Consequently, "classroom assessments, which have the potential to enhance instruction and learning, are not being used to their fullest potential" (National Research Council, 2001, p. 1).

The tendency to focus on the assessment of rote learning has been identified in teachers' formative and summative assessment practices. However, more emphasis is generally placed on the negative impact of teachers' summative assessment practices on students' learning (Harlen, 2004b). Factors found to influence the tendency to focus on rote learning in summative assessments are, firstly, a propensity to produce reliable results and therefore use more test items that can be marked readily and reliably (Harlen, 2005). In general, these test items are more focused on the recall of knowledge. Other factors that might influence the tendency to focus on rote learning are external factors determined by national examinations and the examination programme. In their reflections on the relationship between the curriculum and assessment in the Netherlands, Kuiper, Van Silfhout and Trimbos (2017) emphasised that examination programmes and national examinations have a

'pre-shadowing' effect on classroom assessment practices and the enacted curriculum. Hence, what is assessed in national examinations influences what teachers teach in their classrooms, as well as their assessment practices. When these examinations contain more test items that can be marked readily and reliably with a focus on the recall of knowledge, this can enforce a strategy of 'teaching to the test' and might strengthen the tendency to focus on rote learning.

To prevent teachers' summative assessment practices being too focused on national examinations, reliable results and test items recalling knowledge, these assessments should also meet the criterion of validity. The criterion of validity can be met when summative assessments reflect the full content and objectives of the intended curriculum. In this sense, when summative assessments include test items that have a focus on meaningful learning, the pre-shadowing effect can also function the other way around. Summative assessments as such, can function as a lever for the enacted curriculum and bring the enacted curriculum more in line with the requirements of the intended curriculum, also in terms of meaningful learning (Kuiper et al., 2017).

In geography education, the tendency to focus on the recall of knowledge in summative assessments has also been identified. A study in the USA showed that the majority of large-scale assessments tested students' recall of geographic facts (Wertheim, Edelson, & The Road Map Project Assessment Committee, 2013). Although not always underpinned by empirical evidence, others have identified this tendency in geography education as well. The tendency to focus on the recall of knowledge in summative assessments in geography education is enforced by a demand to produce reliable results and the use of test items "that are relatively closed in nature and require minimal or no judgement. In short, they are safe" (Stimpson, 2006, p. 79).

To date, there is hardly any empirical evidence pertaining to geography teachers' summative assessment practices in the Netherlands. The most important sources of information about geography assessments in the Netherlands are the external exit examinations (CE) and students' results in these examinations. The external examinations contribute 50% to the overall result for geography at the end of secondary education. The other 50% is determined by internal school-based examinations (SE) that are constructed by the teachers. Figures for the exit examinations are only partly satisfactory

because they provide no insight into the quality of internal school-based examinations and how teachers construct these summative assessments. In 2008, the National Institute for Educational Measurement (Cito), at the request of the Inspectorate of Education, investigated the quality of internal school-based examinations in secondary education for four subjects; biology, mathematics, English and the Dutch language (Cito, 2008). Unfortunately, geography as a subject was not included in this investigation.

This paucity of evidence on geography teachers' summative assessment practices in the Netherlands is in line with a recent international review study, which revealed that there is little published data pertaining to geography education and assessment. In their review study, Lane and Bourke (2017) showed that only 30 empirical peer-reviewed studies on geography education and assessment were published internationally between 2000 and 2016. These 30 studies covered a variety of themes, including international assessment, spatial reasoning and formative assessment, besides alignment with curriculum standards, performance standards and methods to assess geographical understanding. Therefore, in their conclusion, Lane and Bourke stated that "this systematic review confirmed the dearth of research on assessment in geography education" (2017, p. 12).

To date, little is also known about the extent to which geography teachers in secondary education in the Netherlands work on their professional development, their competencies in general or, specifically, on their assessment literacy. In the Netherlands, the Inspectorate of Education has a supervisory role on the competence of teachers and their professional development. Because this role of the Inspectorate was implemented by law in 2012, the Inspectorate wanted to determine the extent to which teachers in the Netherlands enhanced their professional knowledge and skills. An important conclusion in this report was that teachers' overall competency should be enhanced (Inspectie van het Onderwijs, 2013). The Inspectorate acknowledged that teachers in secondary education were willing to work on their professional development, but that their workloads were an important constraint. Although this report provided some valuable information about teachers' professional development, the report contains no specific information about teachers' professional development in terms of their assessment literacy. The report also did not contain information about teachers' professional development in specific subjects, such as geography.

This thesis examines how professional growth can be fostered with regard to teachers' assessment practices in pre-vocational geography education in the Netherlands with reference to internal school-based examinations. These examinations are part of the final school examinations and, therefore, have a supposed pre-shadowing effect on teachers' classroom practices and their assessment practices in prior years. Furthermore, these examinations are an important exemplar of teachers' summative assessment practices.

The focus in this thesis is on teachers' development in terms of the knowledge, skills, beliefs and practices with regard to the content of internal school-based examinations, specifically the test items in the examinations. The investigation will consider the type of test items that predominate in the internal school-based examinations, how these test items are related to students' learning, and how teachers can be scaffolded to alter their knowledge, skills, beliefs and practices. Before proceeding to describe the research question and research strategies employed in this investigation, it is important to define the concepts of learning, meaningful learning, teacher professional development, the relationship between assessment and learning, assessment literacy and teachers' conceptions.

1.2 Learning and meaningful learning

The learning sciences, such as cognitive psychology and educational psychology, have identified some fundamental characteristics of learning (Bransford, Brown, & Cocking, 1999). Firstly, many scholars consider learning to be a process in which knowledge is constructed or adjusted based on past experiences, prior knowledge and new information. People either add information to their existing knowledge, which is known as conceptual growth, or adjust their knowledge, which is called conceptual change. Both depend not only on the newly provided information, but also on the existing conceptual structures in the mind. The more developed and powerful these conceptual structures are, the more effectively the learning process is facilitated.

A second fundamental characteristic of learning is that learning is considered to be the outcome of active knowledge construction on the part of the learner. In this process, learners play an active role and learn to take control of their own learning processes. Control of the learning process implies the

self-regulation of learning, which can be stimulated when learners reflect on their goals and are subsequently able to apply and adjust their learning strategies (Cleary & Zimmerman, 2004; Zimmerman, 2002).

Whereas these first two characteristics of learning reflect the individual's role in the active construction of knowledge, other perspectives on learning have emphasised the social context thereof. A third fundamental characteristic of learning, according to many scholars, is that learning is not an individual process; instead, it is influenced by – and is situated in – multiple cultural settings. Learning is not exclusively individual, as it takes place in the presence of other learners (Borko, 2004; Bransford et al., 1999; Putnam & Borko, 2000).

Therefore, in educating people, acknowledgement of learners' socio-cultural contexts, previous experiences and existing knowledge is essential. This existing knowledge, however, might be influenced by prior experiences and socio-cultural contexts in such a way that learners may have developed misconceptions or false beliefs. Educators need to take these misconceptions into account during the learning process, as they may need to be corrected to enable the transfer of knowledge to other contexts. The transfer of knowledge to other contexts is essential if learning is to be considered meaningful (Bransford et al., 1999).

Several attempts have been made to define meaningful learning. In this regard, meaningful learning is often differentiated from rote learning, which is considered to be the opposite. One definition of meaningful learning encompasses all cognitive processes that transcend rote learning, more specifically: understanding, applying, analysing, evaluating and creating (Anderson, Krathwohl, et al., 2001). Harlen and James (1997) referred to this process as deep learning, and distinguished it from surface learning. A deep learning approach, in their view, requires that students are actively involved in developing their personal understandings by relating new ideas to existing knowledge, whereas the surface learning approach is focused on reproducing and accepting ideas passively. James and Gipps (1998) made a comparable distinction between shallow and deep learning. Shallow learning resembles rote learning, but has a focus on memorisation. Deep learning, on the other hand, involves the intention to understand and requires an active approach to learning. Fullan and Langworthy (2014) transformed the concept of deep learning into a process of learning in which students not only create new

knowledge but are also able to connect it to real-life problems and are able to work together with their teachers on real-life problem solving.

In each of the above conceptualisations, it is essential that learners actively develop higher levels of understanding; in other words, they are active in the sense that learners create their own knowledge. It is also essential that learners are able to use subject-specific conceptual and procedural knowledge. In this constructivist approach to learning, it is recognised that learners create their knowledge based on both new information and prior knowledge, and then give meaning to it. To accomplish this type of learning, not only must instruction go beyond the recall of merely factual knowledge, so must assessment (Anderson, Krathwohl, et al., 2001).

Although all of the above conceptualisations of meaningful learning address the notion of knowledge construction in some way, none of them are related explicitly to subject-specific content. In 1972, Peel proposed a way to describe cognitive processes that demand more than memorisation and retention in relation to subject-specific content. In an attempt to discern different levels of understanding over the course of adolescence, Peel related understanding to the application of concepts within school subjects. For geography education, understanding was defined as “a grasp of cause and effect, a capacity to follow a sustained argument and a power to evaluate” (Peel, 1972, p. 164) using substantive geographical concepts.

This concept of understanding in geography education has been explored more intensively by Bennetts (2005a, 2005b, p. 113), who defined understanding as “making sense of something or giving meaning to something”. In this respect, Bennetts stressed the importance of the active engagement of learners in developing understanding through several mental processes in relation to their experiences and ideas. Ideas are the subject-specific constructs and include concepts, generalisations, models and theories. Understanding, in this sense, is strongly related to perspectives on geography and subject-specific key concepts.

1.3 Teacher professional development

Over the past few decades, several studies have provided important information regarding how teachers’ professional development can be

fostered. Whitcomb, Borko and Liston (2009, p. 208) identified five important features of professional development programmes, “namely that professional development programs be situated in practice, focused on student learning, embedded in professional communities, sustainable and scalable, and both supported and accompanied by carefully designed research”. Other researchers have added some important characteristics, such as the duration of the activity; professional development programmes are more effective when they are stretched over time and when teachers have the opportunity to implement the intended changes (Bransford et al., 1999; Penuel, Fishman, Yamaguchi, & Gallagher, 2007). An additional characteristic is to focus on the subject matter or content (Garet, Porter, Desimone, Birman, & Yoon, 2001).

A review of effective interventions on teachers’ professional development (Van Veen, Zwart, Meirink, & Verloop, 2010) summarised the characteristics of successful interventions. Successful interventions

1) focus on teachers’ daily practices and the pedagogical content knowledge of teachers,
2) stimulate teachers as active participants and learners,
3) stimulate collaborative learning,
4) are stretched out across time, and
5) are incorporated into the policy of the school or national innovations.

According to the authors of this study, teacher learning in real-world settings was expected to be another characteristic of successful interventions. However, hardly any empirical evidence was found to support the expectation that interventions in real-world settings are more effective than are more traditional, out-of-school interventions.

A more recent study of teacher professional development (Maandag, Helms-Lorenz, Lugthart, Verkade, & Van Veen, 2017) confirmed these characteristics of successful interventions. In addition, in this review, which was based on recent empirical studies, other characteristics of successful interventions were suggested, such as coaching by trained programme leaders and reflection on teaching skills in line with the teachers’ development stages. Both review studies also revealed that most reported teacher professional development programmes seemed to lack an idea how teacher learning can be fostered by the intervention, namely a ‘theory of improvement’ or a ‘theory of teacher learning’.

Although a theory of learning appeared to be missing, the importance of having a theory of learning supporting the teacher professional development programme was emphasised in both review studies. With regard to the relationship between professional development and learning, professional development is usually considered to be the result of teacher learning. In this respect, many scholars have emphasised that teacher learning through professional development programmes is 'situated' (Borko, 2004; Putnam & Borko, 2000). In this perspective, situated means that teacher learning is considered to be the outcome of participation in social systems and the individual use of knowledge within these systems. Teacher learning, therefore, is regarded as the result of active individual construction of knowledge and skills within multiple participative contexts of teachers' daily practice.

1.4 Assessment and learning

In research on the relationship between assessment and students' learning, the importance of formative assessment – or the formative use of assessments – has been stressed (Black & Wiliam, 2009, 2012; Hattie & Timperley, 2007; Sluijsmans, Joosten-ten Brinke, & Van der Vleuten, 2013). Although there is no single definition of formative assessment, most scholars agree that essential elements of formative assessment are the provision of feedback on learning tasks, the learning process and self-regulation (Hattie & Timperley, 2007). Formative assessment also includes the application of learning strategies to scaffold students' learning, such as questioning, test dialogues, self-assessment and peer-assessment. Formative assessment, in this sense, is mainly referred to as Assessment for Learning (AfL), and is often considered to be the opposite of summative assessment, or Assessment of Learning (AoL). Summative assessment is considered merely to determine students' academic achievement and is often regarded as contributing less to progress in learning (Dunn & Mulvenon, 2009).

Evidence from research, however, shows that summative assessment can also contribute to progress in learning. Brookhart (2001) showed that good students used assessment information for formative purposes to regulate their own learning in both summative and formative situations. In 50 interviews with students, successful students seemed to integrate the formative and summative purposes of assessment by thinking about "how well they did and summing up their accomplishment to date, on the one hand,

while realising that they had information with which to approach future learning, on the other" (Brookhart, 2001, p. 167). These results showed that, while formative and summative assessments might have different functions, both can be used for formative and summative purposes and can thus stimulate students' learning, provided that summative assessments are accompanied by feedback.

The sometimes strict distinction between formative and summative assessment with regard to the potential benefits for learning has also been criticised by others (Bennett, 2011; Taras, 2007, 2009). When summative test results are used to identify potential gaps in students' knowledge, these results could also be used for formative purposes, thus helping to stimulate students' learning. Feedback on the results and on students' performances is, in this respect, extremely important (Harlen & James, 1997). Consequently, the purposes of formative and summative assessments are brought more in line with each other (Bennett, 2011).

Another potentially positive effect of summative assessment on learning has been defined as the testing effect (Roediger & Karpicke, 2006). The testing effect demonstrates that testing produces greater retention over time than does (re)studying the study material, and is therefore a "powerful means of improving learning, not just assessing it" (Roediger & Karpicke, 2006, p. 249). This testing effect has been demonstrated not only with regard to the retention of factual knowledge, but also for higher-order cognitive processes. In their study, Dirkx, Kester and Kirschner (2014, p. 361) showed that "testing benefitted not only the retention of facts from a mathematics text but also the application of the principles and procedures contained in that text".

An important prerequisite if summative assessment is to contribute to learning is that the assessments must be in line with subject-specific objectives (National Research Council, 2001; Stimpson, 2006). The constructive alignment of assessments with objectives and instruction is crucial for supporting learning (Biggs, 1996). Bennett (2011) also stressed that summative assessments can support learning when the tests include rich domain-relevant processes and strategies.

In conclusion, despite the potential negative effects, summative assessments can contribute to progress in learning, particularly when summative

assessments are used for formative purposes, are accompanied by feedback and are in line with subject-specific objectives.

1.5 Teachers' assessment literacy

As stated previously, teachers play an essential role in enhancing students' thinking and learning. In order to be able to play this role fully, teachers should possess the knowledge and skills to align their assessment practices with the goals of education and instruction. To be able to do so, teachers should be literate with regard to assessments. Teacher assessment literacy has been defined in several ways. According to DeLuca et al. (2016, pp. 251-252) assessment literacy "involves the ability to construct reliable assessments and then administer and score these assessments to facilitate valid instructional decisions anchored to … educational standards". Xu and Brown (2016, p. 149) referred to assessment literacy "as a basic understanding of educational assessment and related skills to apply such knowledge to various measures of student achievement". Throughout this thesis, the term 'assessment literacy' will refer to the essential knowledge and skills to construct, score and administer assessments in order to use students' results to make sound decisions.

To identify the extent to which teachers experience assessment literacy, several standards to measure teachers' assessment literacy have been proposed. One of the earliest, and perhaps most influential attempts, was released as the *Standards for Teacher Competence in Educational Assessment of Students* by the American Federation of Teachers, National Council on Measurement in Education, and National Education Association in 1990 (AFT, NCME, & NEA, 1990; Brookhart, 2011). Teachers' assessment literacy in this document consists of seven standards, which reflect the necessary knowledge and skills for teachers in terms of "understanding principles of sound assessment" (Levy-Vered & Nasser-Abu Alhija, 2015, p. 378). The seven standards are:

1. Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.
2. Teachers should be skilled in developing assessment methods appropriate for instructional decisions.

3. Teachers should be skilled in administering, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods.
4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
5. Teachers should be skilled in developing valid pupil grading procedures that use pupil assessments.
6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.
7. Teachers should be skilled in recognising unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information (AFT et al., 1990).

Since 1990, the definition of teachers' assessment literacy and its constituting standards have evolved further. An important criticism of the *Standards for Teacher Competence in Educational Assessment of Students* was that these did not incorporate current conceptions of formative assessment (Brookhart, 2011; DeLuca et al., 2016) and teacher education (DeLuca et al., 2016). The original *Standards* were merely based on summative assessment practices and, therefore, were somewhat outdated. Brookhart (2011) then proposed a new set of standards that incorporated these new insights.

Other scholars emphasised the influence of teachers' conceptions of assessment on their assessment literacy and on their willingness to evaluate their assessment practices and the potential contribution of these practices to students' thinking and learning. Teachers' conceptions regarding assessment practices are strongly influenced by teachers' beliefs and values (Levy-Vered & Nasser-Abu Alhija, 2015). In addition, these beliefs and values are affected by teachers' experiences with assessment practices and by regulations from outside the classroom (Xu & Brown, 2016).

Recent evidence from research has pointed out that teachers' assessment literacy can be enhanced by education (Koloi-Keaikitse, 2016; Levy-Vered & Nasser-Abu Alhija, 2015). These studies not only showed a positive effect of teacher training on teachers' knowledge, skills and assessment practices, but also a strong relation with teachers' conceptions of assessment. Teacher education on assessment literacy, therefore, "needs to encompass both technical knowledge of assessment and more consciousness-arousing

components that prompt teachers to re-examine their conceptions" (Xu & Brown, 2016, p. 154).

In educating teachers about assessment literacy, it is important to make a distinction between pre-service and in-service teacher training. As pointed out by DeLuca et al. (2016, p. 269), "these two teacher populations may have differing learning needs in assessment and value different learning structures". Teacher training in assessment literacy, therefore, should be "situated within the requirements of different educational contexts" (Xu & Brown, 2016, p. 155), serving different needs of teachers at different stages of their development.

1.6 Teachers' conceptions of assessment

It is necessary at this point to clarify exactly what is meant by teachers' conceptions of assessment. According to Xu and Brown (2016, p. 156), teachers' "conceptions of assessment denote the belief systems that teachers have about the nature and purposes of assessment, and that encompass their cognitive and affective responses". The cognitive dimension of this system refers to teachers' beliefs regarding what is true and false about assessment. Teachers' acceptance of new information or knowledge about assessment depends on whether this new knowledge is congruent with their existing knowledge. In other words, teachers tend to accept new knowledge about assessment when this new knowledge fits in their existing knowledge.

The affective dimension refers to teachers' emotions about assessment. Teachers' emotional inclinations are related to prior experiences with assessment practices and perceptions of students' learning. Prior experiences can be either positive or negative. Both can affect teachers' conceptions into deeply-held belief systems that are more open or resistant to change.

Other terms that are related to teachers' belief systems are also used in research on teaching. One of these terms is teachers' dispositions (Jo & Bednarz, 2014; Schussler, 2006; Schussler, Stooksberry, & Bercaw, 2010). Teachers' dispositions refer to teachers' tendencies to act in addition to their attitudes. Teachers' dispositions, as such, connect teachers' will to enact certain knowledge, skills or strategies with their intentions to achieve educational goals.

Teachers' dispositions have an intellectual dimension, a cultural dimension and a moral dimension. The intellectual dimension is related to teachers' "inclination to process knowledge of content and pedagogy, their awareness of what the educational context requires for desired learning outcomes to be reached, and their inclination to put their knowledge and awareness to use accordingly in the classroom" (Schussler et al., 2010, p. 352). The intellectual dimension is, therefore, strongly related to teachers' pedagogical content knowledge. The cultural dimension is related to teachers' will to meet the educational needs of all the students in the classroom. This depends on teachers' awareness of their own culture, of the students' culture and how to connect instruction with learning intentions to meet the students' needs most effectively while taking these backgrounds into account. The moral dimension refers to teachers' deepest held (moral) values.

Teachers' beliefs systems are often referred to as teachers' beliefs and attitudes in research on teacher professional development (Clarke & Hollingsworth, 2002; Guskey, 1986, 2002). These beliefs and attitudes are distinguished from knowledge and skills, yet they are all seen as part of teachers' personal domain and are thus interconnected. This composition of teachers' personal domain is therefore similar to the classification used to identify the components of teachers' dispositions and conceptions.

There is a cognitive and affective dimension in each of the above conceptualisations of teachers' belief systems. The cognitive dimension refers to teachers' (pedagogical content) knowledge and beliefs about assessment, whereas the affective dimension refers to teachers' emotions. In this thesis, the terms 'conceptions', 'dispositions' and 'beliefs' are used interchangeably, and refer to teachers' deepest held beliefs about assessment, the relationship with their educational goals and the way these are connected to their perceptions of students' meaningful learning.

1.7 Research questions and research strategy

The importance of teachers' role in students' learning and the fact that teachers' classroom assessment practices tend to encourage rote learning instead of meaningful ways of learning raises the question of how teachers' professional growth in this area can be fostered. The focus in this thesis is on teachers' knowledge, skills, beliefs and practices pertaining to internal school-

based examinations in secondary pre-vocational geography education in the Netherlands. These school-based examinations are an important case of summative assessments that have a supposed effect on teachers' classroom assessment practices in prior years.

The research was conducted in a secondary pre-vocational education setting. In the Netherlands, about 53% of all students experience this type of education (see Appendix A). This type of education is comparable to lower secondary education and, for most students, is the final stage of their basic education. Consequently, most students will no longer attend geography classes after graduation. Although pre-vocational education includes a large amount of secondary geography education, and although it has many unresolved issues regarding what and how to teach and assess, research on this part of secondary geography education is scarce. This is an important reason that the context of secondary pre-vocational education was chosen in this research.

The main research question for this thesis is:

How can geography teachers' professional growth in secondary pre-vocational geography education in the Netherlands be fostered with regard to their practices, knowledge, skills and beliefs in relation to school-based examinations and meaningful learning?

The main research question has been divided into the following sub-questions:

1a. What kind of geographical knowledge and which cognitive processes are prevalent in test items in school-based geography examinations in pre-vocational secondary education in the Netherlands?

1b. What kind of beliefs, attitudes and conceptions do geography teachers in pre-vocational secondary education in the Netherlands have upon the school-based geography examinations?

2a. What are the current practices, beliefs and values of geography teachers in pre-vocational secondary education in the Netherlands regarding internal school-based examinations?

2b. What is the relationship between geography teachers' practices in pre-vocational secondary education in the Netherlands and their perceptions of test items that appeal to distinct cognitive processes in their internal school-based examinations?

2c. What is the relationship between the background characteristics of geography teachers in pre-vocational secondary education in the Netherlands and their practices regarding the construction of school-based examinations?

3. 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?

4. How practical and feasible is a teacher professional development programme on internal school-based examinations and meaningful learning in pre-vocational geography education to foster teacher professional growth?

5. How can a designed teacher professional development programme on summative assessment and meaningful learning in pre-vocational geography education in the Netherlands contribute to the professional growth of teachers in terms of changes in teachers' practices, knowledge, skills and beliefs through reflection and enactment?

The research was conducted as an educational design research study (EDR) based on the model by McKenney and Reeves (2012). EDR can be defined as

> the systematic study of designing, developing and evaluating educational interventions (such as programs, teaching-learning strategies and materials, products and systems) as solutions for complex problems in educational practice, which also aims at advancing our knowledge about the characteristics of these interventions and the processes of designing and developing them. (Plomp, 2010, p. 9).

The yield of EDR, therefore, is twofold: a designed intervention that will contribute to a solution to a problem and a contribution to the knowledge regarding how and why the interventions functions. In addition to these yields, EDR contributes to the professional development of the participants via

the collaborative learning of researchers and practitioners (McKenney & Reeves, 2012; Plomp, 2010). EDR usually consists of three consecutive phases: (1) the phase of analysis and exploration, (2) the phase of design and construction, and (3) the phase of evaluation and reflection (McKenney & Reeves, 2012).

In the first phase of this research, a content analysis of school-based internal examinations, a questionnaire among 74 geography teachers, panel interviews with both teachers, and experts and a literature study provided insight in the practices, knowledge and beliefs of geography teachers in pre-vocational education. For the second and third phases, a teacher professional development programme (TPDP) was designed and tested with groups of geography teachers in pre-vocational education in the Netherlands. The designed intervention was first evaluated for consistency and practicality with educational experts. The redesigned intervention was then tested in a pilot study with six teachers. The outcomes of this pilot study were used to evaluate the practicality of the intervention. After the evaluation, the intervention was redesigned and tested with another group of eight teachers. The results from this group were used to identify potential professional growth among the teachers.

The intervention was tested and evaluated in multiple progressive cycles. Thus, the research process was not linear, but cyclical and iterative in nature. The first prototype for the intervention was based on tentative design principles that were derived from the phase of analysis and exploration. Finally, based on an overall evaluation, more conclusive design principles were formulated to provide insight into the function and characteristics of the design.

The research in this thesis consisted of a combination of quantitative and qualitative methods. In the first phase of the research, the content analysis, panel interviews and questionnaire provided qualitative and quantitative results. In the second and third phases, the initial testing and evaluation of the intervention produced qualitative results. The evaluation of the design at this stage was mainly formative, and was intended to improve the designed intervention. As stated by Nieveen (2010, p. 93), formative evaluation is “a systematically performed activity (including research design, data collection, data analysis, reporting) aiming at quality improvement of a prototypical

intervention and its accompanying design principles". In the final phase of testing and evaluating the intervention, a mixed-method study was used to reveal the extent to which professional growth could be identified. Quantitative data in this phase were derived from a content analysis, while qualitative data were obtained from the questionnaire and from the interviews.

1.8 Outline of the thesis

In line with the EDR approach, the chapters in this thesis reflect the different stages of the research design (Figure 1.1). Chapters 2 and 3 form part of the phase of analysis and exploration. Chapters 4 and 5 describe the design of the intervention, and the first formative evaluation of the intervention. In Chapter 6, an evaluation of the tested design will be presented. In Chapter 7, the findings of the research are summarised. This chapter also provides the main conclusions, design principles, a reflection on the research, and recommendations for teacher education concerning assessment literacy and further research.

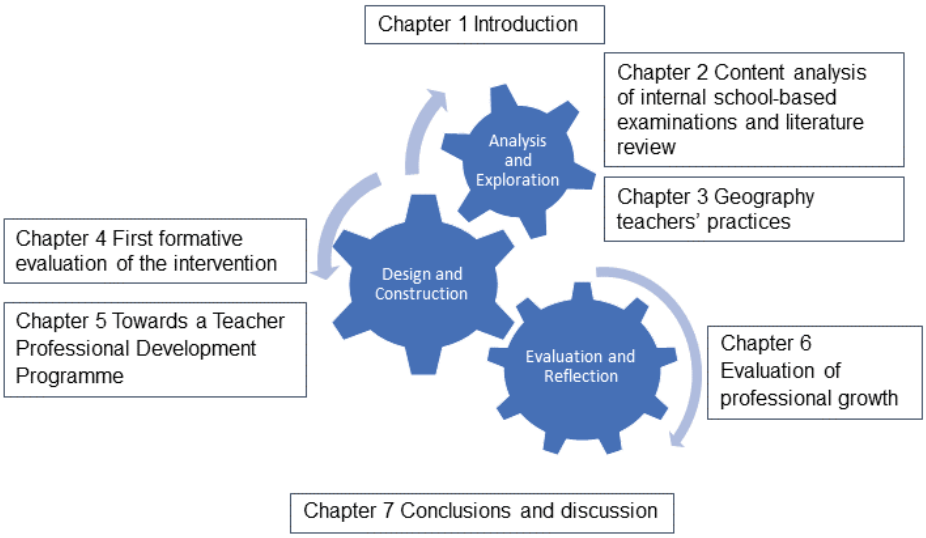


Figure 1.1 Outline of the thesis

Chapter 2. The content of internal school-based examinations in pre-vocational geography education

The study in this chapter reveals the extent to which internal school-based examinations in pre-vocational geography education in the Netherlands contained test items focusing on meaningful learning. Using an adapted version of a taxonomy table based on the revised taxonomy of Bloom, 1108 test items in 49 school-based examinations were scored along a cognitive dimension axis and a knowledge dimension axis. The outcomes of this content analysis were discussed with two panels of geography teachers and experts in (geography) teacher education. The purpose of this study was to investigate the types of geographical knowledge and cognitive processes that prevailed in internal school-based examinations. Panel interviews were used for purposes of validation, and to reveal conceptions regarding the purpose of the internal school-based examinations and alignment with the objectives of these examinations.

Chapter 3. Geography teachers' practices in pre-vocational education regarding internal school-based examinations

This chapter describes an investigation of teachers' practices regarding internal school-based examinations. The aim was to investigate how teachers constructed these examinations and the extent to which they used instruments such as a taxonomy. Furthermore, this study intended to provide insights into teachers' beliefs and values regarding these examinations and summative assessments in general, and the relationship of teachers' dispositions to their practices. In this study, the outcomes of a questionnaire among 74 geography teachers in pre-vocational education were used to reveal teachers' practices and dispositions.

Chapter 4. Characteristics of test items and scoring rubrics contributing to meaningful learning

The outcomes of the first two studies and the literature review were used to design an intervention with two main goals: The first was to identify the characteristics of test items and scoring rubrics in pre-vocational education that have the potential to contribute to meaningful learning in pre-vocational geography education, and the second was to foster teacher professional growth towards summative assessments and meaningful learning. Therefore,

the designed intervention consisted of two separate yet integrated parts. Firstly, a toolkit containing instruction materials, examples of test items, scoring rubrics and strategies for both teachers and students was designed to stimulate the application of test items in school-based examinations focusing on meaningful learning. Secondly, a teacher professional development programme (TPDP) was designed to foster teacher professional growth towards summative assessment and meaningful learning. The toolkit served as a set of materials in the external domain of the TPDP.

An evaluation of the toolkit is provided in this chapter. The toolkit was first presented to four teacher educators for expert appraisal and, as a redesigned toolkit, was tested and evaluated with a group of six geography teachers. These teachers followed a training programme for three months, including instruction, demonstration, collaborative practice and (peer) feedback. During the programme, the teachers and some of their students were observed and interviewed to identify test items, scoring rubrics and strategies that were valued as practical.

Chapter 5. Towards a teacher professional development programme for summative assessment and meaningful learning

The first evaluation of the TPDP, is described in this chapter. The evaluation aimed to determine the extent to which the programme was practical. Therefore, the teachers were required to complete a questionnaire and were interviewed about the practicality of the programme and the components thereof.

Chapter 6. Professional growth towards summative assessment and meaningful learning

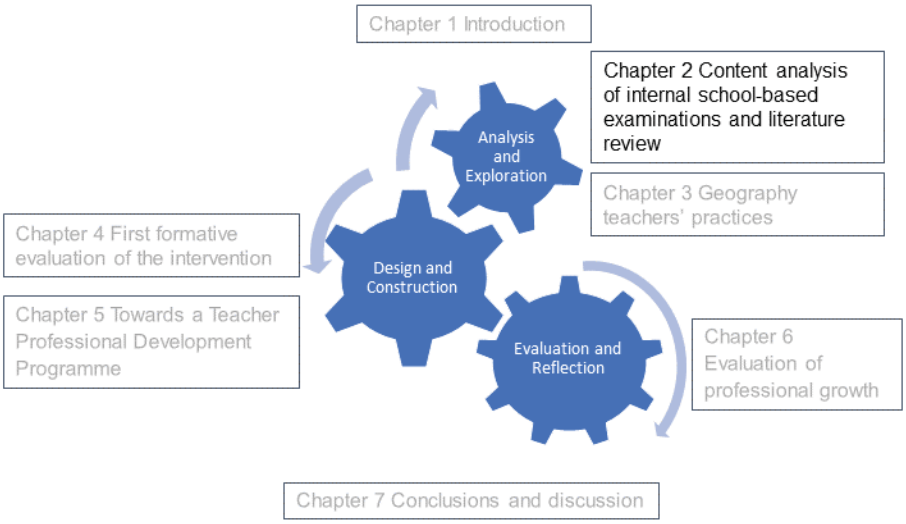
Following the evaluation of the prototype, the intervention was redesigned, tested and evaluated with another group of eight geography teachers in pre-vocational education. The aim was to identify the extent to which professional growth regarding summative assessment and meaningful learning could be identified. In order to identify potential growth, teachers' practices, beliefs and values were measured on multiple occasions during the programme. The outcomes of this evaluation of the programme are described in this chapter.

Chapter 7. Conclusions and discussion

In this final chapter, the main outcomes of the research will be presented. Furthermore, conclusions will be drawn from these outcomes, and the design principles will be formulated. Based on these conclusions, some recommendations for further research and implications for teacher education and geography education will be presented. Reflection on the outcomes and the research will also form part of this chapter.

Chapter 2

Meaningful learning and summative assessment in geography education: An analysis in secondary education in the Netherlands*



* 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.

Abstract

Enhancing meaningful learning is an important aim in geography education. Also, assessment should reflect this aim. Both formative and summative assessments contribute to meaningful learning when more complex knowledge and cognitive processes are assessed. The internal school-based geography examinations of the final exam in pre-vocational secondary education in the Netherlands are an important test case to reveal the extent to which geography teachers construct examinations containing complex knowledge and cognitive processes. In this study internal school-based examinations were analyzed based on a taxonomy table derived from a revision of Bloom's taxonomy (Anderson, Krathwohl, et al., 2001) and discussed with teachers and experts. The results of the content analysis showed that more than half of the test items in the internal school-based examinations are based on remembering knowledge, especially factual and conceptual geographical knowledge.

2.1 Introduction

2.1.1 Meaningful learning and geographical knowledge

An important aim in education is to enhance meaningful learning (Anderson, Krathwohl, et al., 2001; James & Gipps, 1998). Meaningful learning can be defined as constructing knowledge based on new information and prior knowledge (Anderson, Krathwohl, et al., 2001). Meaningful learning, sometimes defined as deep learning, can be distinguished from rote learning. Rote learning refers to remembering or recalling factual knowledge and can be defined as surface or shallow learning (James & Gipps, 1998).

In the past decades emphasis has been on enhancing meaningful learning in geography education. In this respect, the work of David Leat, Margaret Roberts and others have made a significant contribution to the application of teaching and learning strategies (Leat, 1998; Leat, Van der Schee, & Vankan, 2005; Roberts, 2013). Less emphasis has been placed, however, on the contribution of different types of assessments on meaningful learning, in particular the contribution of summative assessments.

Most authors refer to meaningful learning as a combination of several cognitive processes: understanding, applying, evaluating or creating on the one hand and different types of knowledge on the other (Anderson, Krathwohl, et al., 2001; James & Gipps, 1998; Leat & McGrane, 2000; Mayer, 2002; Weeden, 2013). This combination requires an active approach of pupils to learning. Active, in this sense, means pupils have to integrate their knowledge of facts, concepts and procedures with new facts, concepts or procedures in such a way that they construct their own new meaningful knowledge. By constructing this new meaningful knowledge pupils make sense of the new information, whether this new information is provided to them by instruction or assessment.

The construction of new knowledge can offer an important contribution to meaningful learning when pupils are challenged to perform complex tasks. The complexity of the tasks increases when more complex knowledge and cognitive processes beyond remembering are demanded. Although there is no strict hierarchy in the cognitive dimension, evaluating and creating are generally seen as more complex cognitive processes than applying or understanding (Krathwohl, 2002). However, despite this sequence in the

cognitive dimension tasks based on lower order processes can be more demanding for pupils than higher cognitive processes. It depends on the complexity of the knowledge as well.

Understanding is the most comprehensive cognitive process and is sometimes referred to as an overall category for intellectual activities that go beyond recalling knowledge (Bennetts, 2005a), but it is more common seen as a synonym for comprehending (Krathwohl, 2002), one of the former dimensions in the original taxonomy of Bloom. In this sense understanding comprises multiple subcategories as explaining, interpreting, classifying, summarizing, comparing, exemplifying and inferring (Anderson, Krathwohl, et al., 2001). All of these subcategories are important cognitive processes with a huge potential to enhance meaningful learning in education and in particular for geography education.

Meaningful learning, however, becomes less valuable when the higher order cognitive processes are not accompanied in the curriculum by core knowledge. Lambert (2011) stresses the need for defining core knowledge and a knowledge framework for geography education as an important and integral part of the curriculum. Others also write about the importance of defining what kind of geographical knowledge and which concepts besides cognitive processes should prevail in geography education (Brooks, 2008, 2013; Firth, 2013; Haubrich, 1992; Taylor, 2013).

Although Lambert's appeal must be read in the context of the revised National Curriculum in England, the importance of defining geographical knowledge has been an important issue in the Netherlands as well. In the beginning of this century Van der Vaart (2001) already emphasized the need for a geographical framework. This framework consists of (1) core knowledge, (2) knowledge of important geographical issues on different scales, and (3) geographical skills, techniques and methods.

More recently, research has been conducted in the Netherlands on thinking geographically and teaching strategies enhancing geographical reasoning as an important contribution to meaningful learning (Favier & Van der Schee, 2014b; Hooghuis, Van der Schee, Van der Velde, Imants, & Volman, 2014; Karkdijk, Van der Schee, & Admiraal, 2013). These studies contribute to research on the integration of cognitive processes and geographical knowledge. In the past decade this theme has been emphasized by the work

of David Leat but others contributed to this theme as well with publications in journals for geography teachers and books with strategy exemplars (Jackson, 2006; Leat, 1998; Leat & Nichols, 2000; Van der Schee & Vankan, 2006; Van der Schee, Vankan, & Leat, 2003; Vankan & Van der Schee, 2004).

2.1.2 Meaningful learning and assessments

The question, how geographical knowledge and geographical reasoning can be enhanced in such a way that meaningful learning is achieved is not only a question of developing successful teaching strategies, but also a question of constructing powerful tools for assessment. As Bennetts (2005a) pointed out, assessments can be very important in developing understanding amongst pupils. To enhance meaningful learning, both formative and summative assessments are useful. Although formative assessments, also defined as assessments for learning (AFL), have the highest capability of contributing to meaningful learning, summative assessments can contribute to meaningful learning as well. It is important to focus not on just one type of assessment but to use a wide range of types of assessment to support meaningful learning (Harlen, 2005; James & Gipps, 1998).

Caution is needed, however, when emphasis is placed on summative assessments. Several authors have drawn attention to the fact that assessments, mainly summative assessments, can have some negative effects on learning and motivation when the results of the tests are used for purposes other than stimulating learning, such as for purposes of accountability (Bennetts, 2005a; Butt, Weeden, Chubb, & Srokosz, 2006; Harlen, 2005). Accountability purposes can distract the goals of assessment from meaningful learning. The types of questions in the assessments and the methods and procedures that were used to construct the assessments can have a negative impact on learning as well. The test items in the assessments can stimulate rote learning instead of meaningful learning (Davies, 2002; Leat & McGrane, 2000) and teachers can adopt a tendency to ‘teach to the test’ which can have a serious negative impact on learning when the tests mainly assess rote learning (Anderson, Krathwohl, et al., 2001; Harlen, 2005).

Despite these possible negative implications of summative assessments on meaningful learning, summative assessments can contribute to meaningful learning when the negative threats can be overcome by instruments that support meaningful learning. Some authors have put emphasis in this

perspective on developing test items assessing higher order skills (Ediger, 2001; James & Gipps, 1998), others on the role of the teachers' judgement in summative assessments (Harlen, 2005). Airasian and Miranda (2002) emphasized the potential of the taxonomy table of the revised taxonomy of Bloom for developing and stimulating meaningful learning. The taxonomy table, a two-dimensional tool that combines the knowledge and cognitive process dimension, is suitable not only to align assessments with curricular objectives and teacher instruction but with more complex aspects of learning and thinking as well.

A promising assessment instrument in supporting meaningful learning in geography education is the so called SOLO taxonomy. The original SOLO taxonomy, in which SOLO means Structure of the Observed Learning Outcomes, was developed by Biggs and Collis (1982) and meant to evaluate the levels of performance by pupils in five stages; pre-structural, uni-structural, multi-structural, relational and extended abstract. The first stage reflects a level whereby the pupil does not know how to fulfill the task. The second level involves describing one relevant element, the third level multiple elements and on the fourth level the pupil is able to relate these elements. On the fifth and final level the response of the pupil goes beyond induction on the basis of data that were offered in the task and the response includes an abstract principle based on deduction as well.

The SOLO taxonomy has been used by others for further development. Stimpson (1992) combined single test items on the different levels of the SOLO taxonomy to one superitem and tested the validity of this instrument. The results of this study supported the idea that the SOLO taxonomy in combination with these superitems is useful in constructing assessments. The SOLO taxonomy has also been used to test the quality of essays by students (Munowenyu, 2007).

In summary, summative assessments can be powerful in enhancing meaningful learning when the test items demand meaningful learning and evaluation instruments will be used that support these test items. Negative influences as 'teaching to the test' and accountability purposes have to be avoided.

2.1.3 Geography education in the Netherlands

This research is conducted in the theoretical programme of pre-vocational education in the Netherlands (see appendix A for an explanation of the Dutch educational system). In pre-vocational education geography as a subject is compulsory in the first two years, as a separate subject or as a part of social studies. After two years pupils can choose geography as one out of six or sometimes seven subjects for their final exam. In 2013 a renewed examination programme for geography was implemented in pre-vocational education (Examenblad.nl, 2015).

The content of the examination programme of the final exam consists of two parts; the first part is assessed with internal school-based examinations and the second part is assessed with an external end-of-school (exit) examination. The internal school-based examinations programme contains three main areas of geography from the syllabus; (1) Sources of Energy, (2) Poverty and Wealth and (3) Boundaries and Identity. The examination programme for the external end-of-school (exit) examination contains three additional areas of geography; (4) Weather and Climate, (5) Water and (6) Population and Place. The external examination pertains to about one-third of the objectives of the examination programme and the school-based examinations to about two-thirds. Both parts, however, contribute 50% to the overall result for geography.

The objectives of the examination programme for internal school-based examinations are elucidated and exemplified in a syllabus for teachers (SLO, 2012). The syllabus contains the specifications for the three main areas of the internal school-based examinations. The specifications are prescriptive for the content of the programme, yet they do not serve as detailed assessment objectives. Teachers can decide which objectives will be assessed and how. The school is responsible for the choices being made by the teachers.

The syllabus emphasizes the importance of learning the pupils to think and reason geographically. Not only does the syllabus contain a separate area with specifications for geographical skills and methods, the objectives in the three main geography areas also refer to these geographical skills and methods. Pupils are, for instance, expected to compare features and regions within different spatial contexts and draw across physical and human characteristics

to compare geographical features. Furthermore, pupils should conduct a small research in their own neighborhood.

Since half of the result of the final exam for geography in pre-vocational education is based on the internal school-based examinations, it is very important to understand how the objectives for this part of the geography examination programme are aligned with the assessments. Alignment of the objectives with assessments (and other features of a curriculum) is of great importance to achieve the curricular goals (Anderson, 2002; Van den Akker, Kuiper, & Hameyer, 2004).

However, there have been no studies in the Netherlands for geography as a subject in secondary education that examined how geographical knowledge and cognitive processes are assessed in internal school-based examinations. In 2008 the National Institute for Educational Measurement (Cito) conducted research into two pre-vocational education subjects, namely mathematics and Dutch language, and two subjects in general education, biology and English language, to assess the validity and quality of internal school-based examinations (Cito, 2008). As yet no research has been conducted for geography in the Netherlands.

There's a need to know how the geography objectives are aligned with the internal school-based examinations and what kind of geographical knowledge in combination with cognitive processes is assessed in internal school-based examinations. Both are needed to gain more insight in the contribution of these summative assessments in geography education in the Netherlands to meaningful learning. It is also important to know how teachers perceive their school-based examinations with respect to the objectives of the examination programme.

This study explores the content of internal school-based examinations in pre-vocational secondary geography education. The results of this study are meant to give more insight in *what* kind of knowledge and cognitive processes are assessed and *how* teachers perceive their internal school-based examinations in relation to the objectives. These insights help to define to what extend the internal school-based examinations contribute to meaningful learning.

The research questions are as follows:

- *What kind of geographical knowledge and which cognitive processes are prevalent in test items in school-based geography examinations in pre-vocational secondary education in the Netherlands?*
- *What kind of beliefs, attitudes and conceptions do geography teachers in pre-vocational secondary education in the Netherlands have upon the school-based geography examinations?*

2.2 Methodology

For this study two instruments were used to gather data. The first instrument was a taxonomy table derived from the original Revised Taxonomy Table (Anderson, Krathwohl, et al., 2001). This instrument was used for a content analysis of internal school-based examinations to answer the first research question. The results of the analysis were discussed in two panel interviews. This second instrument was meant to give more insight in the beliefs, attitudes and conceptions teachers have upon internal school-based examinations.

The first instrument, a taxonomy table, is based on the original revised taxonomy developed by Anderson, Krathwohl et al. (2001) and the objectives for internal school-based examinations in the examination programme of the final geography exam for pre-vocational education (SLO, 2012). Both the revised taxonomy and the objectives for internal school-based examinations in the examination programme are based on two dimensions: a knowledge dimension and a cognitive process dimension. These two dimensions were brought in line with each other in a taxonomy table (see appendix B).

The first dimension of the taxonomy table, the knowledge dimension, consists of four categories and nine subcategories. The first category is factual knowledge, which can be subdivided into (a) *knowledge of specific details and elements* and (b) *knowledge of simple concepts and terminology*. The second category is conceptual knowledge, which can be subdivided into (c) *knowledge of classifications and categories*, (d) *knowledge of geographical principles or relationships between concepts*, and (e) *knowledge of geographical models and theories*. The third category, procedural knowledge, is subdivided into (f) *geographical skills*, (g) *geographical methods* and (h) *knowledge of criteria concerning geographical skills and methods*. Finally, the fourth category

consists of metacognitive knowledge, i.e. (i) *knowledge of (learning) strategies*.

The second dimension of the taxonomy table consists of five cognitive processes: *remember*, *understand*, *apply*, *evaluat*e and *create*. Unlike the original taxonomy table analysing is not a separate category. The choice to reduce the cognitive processes in the geography taxonomy table to five instead of six processes is defendable, as Anderson, Krathwohl et al. already suggested, because analyzing can be divided into three subcategories that can be allocated to other categories. As they have put it: "Although learning to analyze may be viewed as an end in itself, it is probably more defensible educationally to consider analysis as an extension of *understanding* or as a prelude to *evaluating* or *creating"* (Anderson, Krathwohl, et al., 2001, p. 79).

In March 2014 the taxonomy table was validated in two workshops with geography teachers using the theoretical programme of pre-vocational secondary education (vmbo-tl). In these workshops teachers were asked to score a number of test items in the table. In both workshops there was consensus, about the way the items could be scored in the table.

In view of the content analysis a request for internal school-based examinations in the theoretical programme of study for pre-vocational secondary education (vmbo-tl) was sent to teachers by the different networks of teacher training institutions for secondary education in the Netherlands. The internal school-based examinations were collected during the spring and summer of 2014. A total of 49 internal school-based examinations were sent in by geography teachers from 13 schools across different parts of the Netherlands. The internal school-based examinations were all conducted in the school year 2013-2014 in grade Secondary 3 and part of the renewed examination programme for geography in the theoretical programme of prevocational secondary education (vmbo-tl).

Next, each test was checked in Ephorus on duplications. After removal of the duplications a total number of 1108 unique test items remained to be analyzed and were classified in the taxonomy table. For the purpose of this content analysis, the objectives for the internal school-based examinations have been scored in the taxonomy table as well. This gave the opportunity to compare the outcomes of the analysis of the internal school-based examinations with the intended objectives in the examination programme,

and provided more insight in the alignment of objectives and summative assessments in the internal school-based examination programme.

The results of the content analysis were discussed in two separate panel interviews. The participants of the two panel interviews were selected and invited based on their experience as secondary teachers or their expertise in pre-service teacher education or curriculum development and assessments. Nine participants were secondary teachers in pre-vocational education and eight of them had constructed internal school-based examinations in 2013-2014, four participants were Geography educators, one participant was from the Netherlands Institute for Curriculum Development (SLO) and one from Cito. Six participants attended the first interview and nine participants the second.

In both interviews the participants were asked to respond to the most important outcomes of the content analysis. Both interviews were fully open interviews based on three introductory questions: (1) “What do you think of the outcome of the content analysis”? (2) “What could be an ideal distribution of test items in the taxonomy table”? and (3) “Is it possible to achieve this ideal distribution of test items in internal school-based examinations”?

2.3 Results

This section provides the main findings of the content analysis of internal school-based examinations as well as the main outcomes of the panel interviews. The content analysis gives an answer to the first research question of this study and the panel interviews contribute to answer the second research question.

2.3.1 Content analysis of internal school-based examinations

Table 2.1 shows that a majority of the test items were classified as assessing conceptual knowledge, mainly knowledge of geographical principles or relationships between concepts. About 60% focused on this subcategory of geographical knowledge (see appendix C for examples of test items from the analyzed internal school-based examinations).

The second most important subcategory is *knowledge of simple concepts and terminology*. Almost 23% of the test items dealt with this type of knowledge.

Table 2.1 Percentage (*number*) of test items of analyzed internal school-based examinations and number of objectives for internal school-based examinations, scored for each cell in the taxonomy table.

Knowledge Dimension		Cognitive Process Dimension					
		Remember	*Understand*	*Apply*	*Evaluate*	*Create*	Total
Factual Knowledge	*(a) Knowledge of specific details and elements*	5 (*60*) 34	0 (*2*) 11				6 (*62*) 45
	(b) Knowledge of simple concepts and terminology	16 (*180*) 12	6 (*71*) 9				23 (*251*) 21
Conceptual Knowledge	*(c) Knowledge of classifications and categories*	2 (*21*) 17	1 (*8*) 10				3 (*29*) 27
	(d) Knowledge of geographical principles or relationships between concepts	39 (*430*) 25	20 (*227*) 24		1 (*9*) 10	0 (*2*) 2	60 (*668*) 61
	(e) Knowledge of geographical models and theories				1		1
Procedural Knowledge	*(f) Geographical skills*		2	9 (*97*) 8	4	6	9 (*97*) 20
	(g) Geographical methods		8	0 (*1*) 1	3	3	0 (*1*) 15
	(h) Knowledge of criteria concerning geographical skills and methods				1	1	2
Metacognitive Knowledge	*(i) Knowledge of (learning) strategies*			1	1	1	3
Total		62 (*691*) 88	28 (*308*) 64	9 (*98*) 10	1 (*9*) 20	0 (*2*) 13	100 (*1108*) 195

The other subcategories were less prevalent in the internal school-based examinations. Procedural knowledge, especially *geographical skills*, accounted for 9% and factual *knowledge of specific details and elements* for about 6%. The remaining subcategories, *knowledge of geographical models and theories*, *knowledge of criteria concerning geographical skills and methods* and *knowledge of (learning) strategies* were hardly assessed at all.

In terms of cognitive processes the emphasis is on *remembering.* About 62% of the test items were based on this cognitive process. The second category of this dimension that prevailed in the tests was *understanding*, which accounted for 28%. *Applying* accounted for another 9%, with only 1% left that appealed to *evaluating* or *creating*.

The combination of the two dimensions shows that test items classified as *remembering knowledge of geographical principles or relationships between concepts* accounted for almost 39% of the test items. Two other prevailing cells in the taxonomy table are *understanding knowledge of geographical principles or relationships between concepts* and *remembering knowledge of simple concepts and terminology*, containing 20% and 16% of the test items. The other cells in the taxonomy table are less prevalent. Only *applying geographical skills (9%), understanding knowledge of simple concepts and terminology* (6%) and *remembering knowledge of specific details and elements (5%)* could to some extent be classified in the tests. The other combinations of geographical knowledge and cognitive processes were merely absent in the tests.

The objectives for internal school-based examinations were also scored in the taxonomy table (Table 2.1). Some of the objectives contain different categories of knowledge and different categories in the cognitive dimension and were scored in more than one cell. Pupils are, for example, supposed to describe and explain certain features and the associated objective was scored in more than one cell. The total number of objectives in the taxonomy table, therefore, outlines the total number of objectives in the examination programme.

A comparison of the pattern of objectives in the taxonomy table with the pattern of the analyzed test items showed to some extend the misalignment of objectives and test items. The dominance of *remembering* as cognitive process in the test items compared to the objectives is obvious. Secondly,

higher order cognitive processes like *evaluate* and *create* are more prevalent in the objectives than in the test items.

The classification of test items in the taxonomy table compared for the three main areas of geography in the internal school-based examinations programme displayed no significant difference (Figure 2.1). For each subject the pattern was more or less the same. Most test items could be classified as *remembering knowledge of geographical principles and relationships between concepts* and *understanding knowledge of geographical principles and relationships between concepts* or *remembering knowledge of simple concepts and terminology*, the latter especially in tests on Boundaries and Identity.

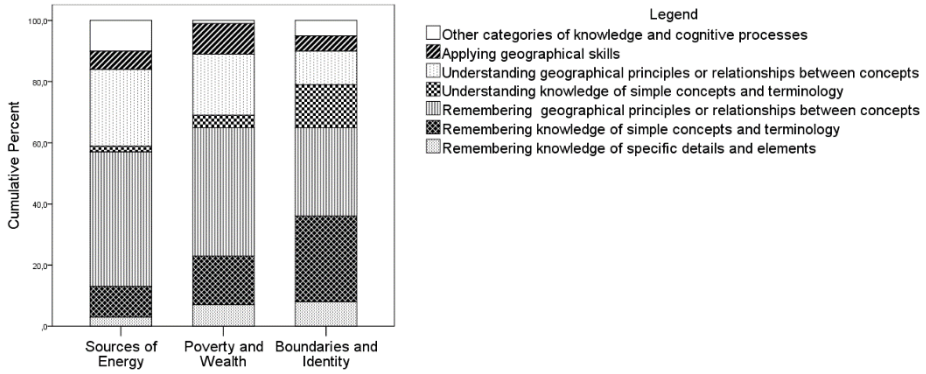


Figure 2.1 Score of test items for three main areas of geography in internal school-based examinations (percentages).

2.3.2 Panel interviews

Most participants on the panel interviews recognized the overall pattern of scored test items in the taxonomy table on internal school-based examinations. As one of the participants mentioned: “Emphasis is on recalling knowledge, but I’m not surprised.” The other participants confirmed that *remembering* is an important cognitive process in assessments in pre-vocational secondary education and particularly *factual* and *conceptual knowledge* is being assessed.

Some of the participants included a kind of judgement in their first reaction. In the first panel interview one of the teachers started with the comment “We

prepare our pupils for the future but obviously this is not a purpose of the internal school-based examinations." This reaction immediately provoked an interpretation and evaluation from the others on the pattern in the taxonomy table. In both panel interviews participants interpreted the scores in the taxonomy table as distinct from a more ideal pattern with more test items on complex knowledge and especially on higher order cognitive processes. Although all the participants agreed on the desirability to assess more complex cognitive processes not all of them were convinced that these higher order processes should be examined in summative assessments, like the internal school-based examinations. Some of the teachers raised the question whether it is desirable and possible to examine higher order cognitive processes in summative assessments in pre-vocational education. Others suggested that these processes could be better examined in formative assessments even when the objectives for the internal school-based examinations request the assessment of more complex knowledge and cognitive processes in these internal school-based examinations.

In both panel interviews there was consensus about the idea that the formats used in the external end-of-school (exit) examination are more than just a guideline for teachers to use the same formats in their internal school-based examinations. By using the same formats teachers feel they do a much better job in preparing their pupils for the external end-of-school (exit) examination. As one of the participants said: "The internal school-based examinations are not meant to prepare pupils for the end-of-school (exit) examination, but when you don't you might have a problem." It is almost a must for teachers to use the same formats, although most of them agreed on the importance to assess higher order cognitive processes in order to achieve the "real" goals with geography education. As one of the teachers admitted, "Preparation for the end-of-school (exit) examination is leading, that's my frustration." All of the participants agreed that a change in formats in the external end-of-school (exit) examination would contribute to the application of other formats in the internal school-based examinations.

The formats in the external end-of-school (exit) examination were not the only felt restriction on assessing more complex knowledge in combination with higher order cognitive processes in internal school-based examinations. Other restrictions mentioned by the participants were a lack of time to practice these other assessment formats with pupils and a lack of confidence

in scoring these other assessment formats. The participants admitted that "good practices of new assessment formats" as well as "instruments to score the performance of the pupils in these formats" would be of great help, especially to overcome problems of reliability. Again, harmonization with formats in the external end-of-school (exit) examination is required according to the participants, as long as accountability remains an important issue in secondary education.

2.4 Conclusions and discussion

The purpose of this study was to investigate what categories of geographical knowledge and cognitive processes prevail in internal school-based examinations in the theoretical programme of pre-vocational education in the Netherlands. Secondly, this study was conducted to examine what kind of beliefs, attitudes and conceptions geography teachers have upon the school-based examinations.

This study has uncovered that a majority of test items deal with the lower categories in the cognitive process dimension, mainly *remembering* and to a somewhat less extent *understanding*. About two-thirds of all the test items are based on rote learning. The more complex cognitive processes like *evaluating* and *creating* are hardly assessed at all. From this point of view, the contribution of the internal school-based examinations to meaningful learning is problematic.

In the knowledge dimension emphasis is being laid on *facts*, *concepts* and *geographical principles and relations between concepts*. *Procedural knowledge of geographical skills and methods* is less prevalent. Remarkably, in none of the test items *knowledge of geographical models or theories* was assessed.

Both dimensions combined reveal that Dutch geography teachers in pre-vocational education tend to focus on testing geographical concepts, geographical principles and geographical relations between concepts in such a way that emphasis is being laid on rote learning and not on different kinds of meaningful learning. In the panel discussions teachers confirmed that remembering is an important dimension in their internal school-based examinations.

The way teachers implement these dimensions of knowledge and cognitive processes in the internal school-based examinations tends to fit in with a broader discussion about geographical knowledge and generic skills. Like in other countries (Lambert, 2011), there seems to be a tendency in the Netherlands in recent years to focus on assessing basic geographical knowledge in the final examinations instead of generic skills (Hooghuis et al., 2014). An important outcome of this tendency could be that teachers put more emphasis on test items in internal school-based examinations that appeal for remembering geographical knowledge instead of test items appealing for more complex knowledge and cognitive processes, although we have to be careful with these statements because we do not really know how internal school-based examinations were constructed in earlier years.

The tendency to put more focus on remembering geographical knowledge does not only raise the question to what extent the internal school-based examinations contribute to the aim of a school geography enhancing meaningful learning but also to what extent these examinations match with the purpose of the examination programme. The syllabus for the internal school-based examinations prescribes that pupils should be taught to think geographically and being able to apply several geographical skills and methods. Based on these prescriptions and the analysis of the objectives for the internal school-based examinations in the taxonomy table you might expect that more different types of knowledge and cognitive processes would be assessed. Almost none of the analyzed internal school-based examinations, however, contained more complex test items appealing to higher cognitive processes. Most analyzed test items were constructed in formats to assess recalling, like matching, true-false or multiple choice, or in assessment formats testing understanding, like constructed response (i.e., supply an answer) or selected response (i.e., choose an answer). To achieve the aim of assessing higher cognitive processes other kinds of test items than those in the analyzed internal school-based examinations seem to be necessary (Anderson, Krathwohl, et al., 2001; James & Gipps, 1998; Lee & Shemilt, 2003; Weeden, 2013; Wood, 2013).

The examination programme is more demanding towards assessing higher cognitive processes than the analyzed internal school-based examinations reflect. One of the main objectives of the examination programme for internal school-based examinations is that pupils have to carry out a simple enquiry-

based exercise in their own neighbourhood. None of the internal school-based examinations that were analyzed contained a kind of assessment as meant in the objectives. This does not justify the conclusion that these assessments are not presented to pupils at all, rather that the internal school-based examinations obviously have another purpose for geography teachers. An important argument for teachers why more complex test items seem to be less prevalent in their internal school-based examinations is that complex knowledge and skills can be just as well, or perhaps even better, assessed formative instead of summative. Assessment for Learning can fill the gap (Wood, 2013) that appears to be in internal school-based examinations concerning meaningful or deep learning. Some teachers confirmed in the panel interviews that these kinds of assessments are part of their programme, not in summative assessments *of* learning but as part of assessments *for* learning in their classrooms. As they put it: "Enquiry-based assessments are part of the curriculum, but not a part of the internal school-based examinations*"*.

These outcomes reveal that the perception of the geography teachers towards the purpose of the internal school-based examinations is aberrant from the standards of the examination programme. Teachers confirmed in the panel interviews that an important purpose of the internal school-based examinations is to prepare the pupils for the external final examinations by using the same assessment formats for test items in the internal school-based examinations as in the external final examinations. In their words: "*We have to prepare our pupils in the same way as they will be assessed in the external final examinations and therefore construct our internal school-based examinations likewise*". In this sense there is a strong tendency of 'teaching to the test' (Anderson, Krathwohl, et al., 2001; Weeden, 2013). Perhaps this tendency can even be better described as 'testing to the test'.

Finally, the results raise a question about the competence and confidence of teachers towards assessing complex knowledge and cognitive processes in internal school-based examinations. A reason why teachers might hesitate to use more complex test items in the internal school-based examinations could be the lack of appropriate instruments to construct more complex test items and instruments to score the performance of the pupils.

Another reason why teachers hesitate to use more complex test items in the internal school-based examinations might be accountability. The results of the internal school-based examinations have to be in line with the results of the external final examinations. Schools have to justify the results towards the Dutch Inspectorate of Education. Weeden (2013) already raises the question whether the tendency to put more emphasis on accountability purposes has led to a loss in teachers' confidence to judge the performance of pupils. Reliable instruments that have been designed and tested can possibly help teachers to overcome this lack of confidence assuming that accountability will continue to play an important role in secondary education in the Netherlands.

If enhancing meaningful learning is an important aim in school geography in secondary education, the assessments should reflect this aim. From this study, it seems that other kinds of assessment formats are needed to contribute to the aim of enhancing meaningful learning by summative assessments, not only in internal school-based examinations but also in the external final exam. Teachers nowadays tend to focus on assessing rote learning and they seem to have a tendency not only of 'teaching to the test' but also of 'testing to the test'.

Caution is demanded, however, drawing firm conclusions from both the content analysis and the panel interviews. First of all, 2013-2014 was the first year of the new geography examination programme for the theoretical programme of prevocational secondary education in the Netherlands. The three main areas of geography belonging to the examination programme that were assessed in the internal school-based examinations were assessed for the first time. Teachers could have avoided risks by conducting test items that mainly assessed *remembering* and *understanding facts*, *concepts* and *geographical principles and relations between concepts* in assessment formats as described above. In the forthcoming years teachers might include test items assessing more complex geographical knowledge and cognitive processes.

Secondly, as stated above, teachers might have assessed the objectives in the examination programme containing more complex geographical knowledge and cognitive processes but not as a part of the summative assessments. Assessments like enquiry-based exercises in their own neighbourhood could have been part of formative assessments in the classroom. In fact, according

to some authors meaningful learning can be achieved just as well or perhaps even better by these kinds of assessment, because these assessments for learning are more effective than summative assessments (Weeden, 2013).

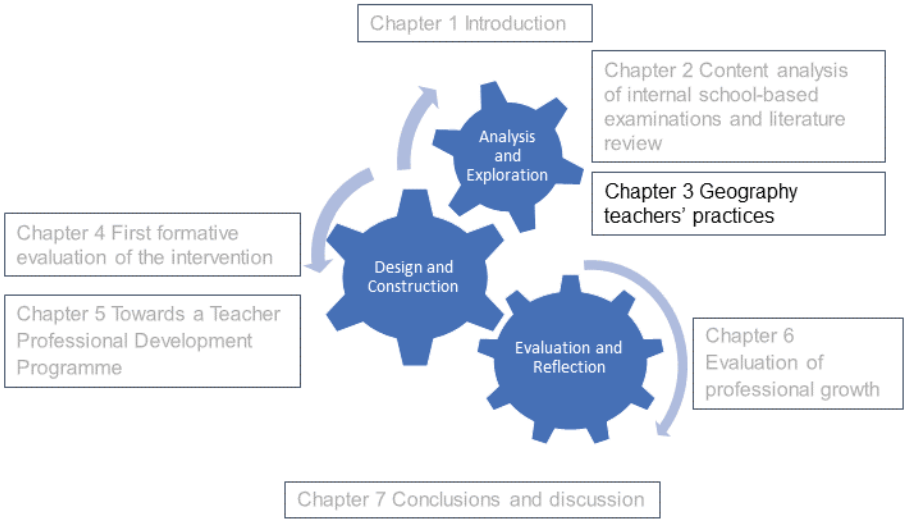
Some comments have to be made on the number of analyzed internal school-based examinations. Although a substantial number of internal school-based examinations and test items were analyzed, still only 13 schoolteachers sent in their internal school-based examinations. To draw more firm conclusions on the assessment of geographical knowledge and cognitive processes an analysis of tests items from more different internal school-based examinations and schools is needed.

Caution is also needed on drawing conclusions concerning the beliefs, attitudes and conceptions of the geography teachers towards the internal school-based examinations. The panel interviews cannot be seen as representative for the geography teachers in pre-vocational education in the Netherlands due do the small numbers. Further research is needed to reveal what geography teachers will stir to construct internal school-based examinations with more complex test items.

The results of this study point to a need to conduct additional research providing insight what teachers need to assess more complex geographical knowledge and cognitive processes in internal school-based examinations in prevocational secondary education in the Netherlands and how this can be accomplished. Which other formats for test items assessing more complex geographical knowledge and cognitive processes can be developed and implemented in internal school-based examinations? And also, what kind of instruments do teachers need to construct more complex test items and score reliably the responses on these test items to give more attention to meaningful learning?

Chapter 3

Geography teachers' practices regarding summative assessment: A study of pre-vocational education in the Netherlands*



* 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.

Abstract

To start a teacher professional development programme on the relationship between classroom summative assessment and learning, the current practices and dispositions of geography teachers towards internal school-based examinations in pre-vocational education in the Netherlands were investigated. A questionnaire provided data on how teachers construct these examinations and how they perceive the extent to which they use test items in these examinations that appeal to distinct cognitive processes. The data were statistically analysed to explore teachers' practices regarding the construction of the examinations and the correlation with their perceptions on test items appealing to distinct cognitive processes. The results showed that teachers rarely construct test items themselves; instead, they rely to a considerable degree on test items created by outside sources. In particular, older teachers and teachers with greater teaching experience tend to use more test items from outside sources. According to the respondents, about two-thirds of the test items appeal to higher cognitive processes. When teachers do construct test items themselves, however, they perceive to use more test items that appeal to higher cognitive processes. Furthermore, teachers' dispositions regarding the purpose of the internal school-based examinations seem to be highly influenced by high-stakes tests, such as the national exam.

3.1 Introduction

3.1.1 Teachers' practices regarding summative assessment

The relationship between teachers' practices on classroom assessment and learning has been studied extensively in recent decades. Several reviews and studies on this relationship have been published (Black & Wiliam, 1998a; Black, Harrison, Hodgen, Marshall, & Serret, 2010, 2011; Harlen, 2004a, 2004b; Harlen & Deakin Crick, 2002, 2003). Some of the reviews or studies focused on the relationship between assessment and learning (Black et al., 2011; Black & Wiliam, 1998b); others focused on the relationship between summative assessment and teachers' practices (Black et al., 2010; Harlen, 2004b) or on the relationship between students' learning and motivation (Harlen & Deakin Crick, 2002).

A central issue in research on the relationship between assessment and learning is the effect of assessment—formative as well as summative—on meaningful learning. Meaningful learning is generally defined as a learning process in which learners actively construct knowledge by integrating new information with existing knowledge. This concept has also been equated with cognitive processes that transcend rote learning (Anderson, Krathwohl, et al., 2001). Cognitive processes transcending rote learning include processes such as understanding, applying, analysing, evaluating and creating.

The relationship between assessment and meaningful learning is not always positive. In their review on classroom formative assessment, Black and Wiliam (1998a) revealed that classroom evaluation practices tend to encourage rote learning instead of meaningful learning. Others emphasized a similar tendency of teachers to focus classroom summative assessments on fact recall and to a lesser extent on critical thinking or other complex and demanding cognitive skills (Harlen, 2005; James & Gipps, 1998).

The tendency of teachers to use test items in classroom summative tests that focus on recall and memorization is often strengthened by the impact of high-stakes tests (Klenowski & Wyatt-Smith, 2011). In high-stakes tests such as national exams, test items in general put more emphasis on rote learning instead of meaningful learning (Harlen, 2005; James & Gipps, 1998). This tendency is enforced by the requirement to use test formats that produce reliable test results (Harlen, 2005; Stimpson, 2006). Reliable test results are

best achieved using test items that can be readily and reliably marked (Harlen, 2005). Teachers frequently imitate these formats for test items in their classroom summative assessments (Black & Wiliam, 1998a) and train their pupils how to answer these specific test items (Harlen, 2005). This phenomenon is also referred to as 'teaching to the test.'

The same practices can also be found in geography education. A study on assessment practices in K-12 classrooms and large-scale assessments in the USA revealed that a majority of the assessments test students' recall of geographic facts (Wertheim et al., 2013), while less than a third of the test items appealed to higher cognitive processes and the application of geographical skills. The tendency to give priority to test items that seem to ensure reliable test scores has been observed by others as well (Stimpson, 2006; Weeden, 2013)

The impact of high-stakes tests as such undermines teachers' attention to validity issues in summative assessments (Black et al., 2010). Teachers seem to be more concerned with issues of reliability and accountability than whether the assessments are in line with subject specific objectives. To achieve that teachers will pay more attention to the issue of validity, Black et al. (2010, p227) suggest 'that an appeal to the belief and values that underlie their commitment to their subjects can be a way to make validity a more salient feature of their work'. Besides reflection on their beliefs and values, attention should be given to teachers' assessment skills.

Teachers' practices in classroom summative assessment are not only influenced by high-stakes tests but also by teachers' dispositions regarding summative assessment. Teachers' dispositions can be described as the ability of teachers to apply knowledge and skills attuned to their beliefs and values. Jo and Bednarz (2014, p. 199) defined teachers' dispositions as "the tendencies of a teacher's behaviour employing particular knowledge and skills to achieve certain teaching goals".

Teachers' dispositions are the interplay between three separate yet interconnected domains; the intellectual, the cultural and the moral domain (Schussler, 2006; Schussler et al., 2010). These three domains reflect the subject specific content, the identities of teachers and their values. For teachers it is not sufficient to have the knowledge and skills when they are not willing to employ or enact them in their classroom (Jo & Bednarz, 2014). The

will to enact pedagogical content knowledge in the classroom is affected by teachers' personal beliefs and values.

Teachers' tendency to use test items in classroom summative tests that focus on recall and memorization are influenced, therefore, by teachers' dispositions towards summative assessment as well. Teachers, however, are not always aware of the fact that validity is under pressure because of this tendency. The review by Black and Wiliam (1998a) revealed that there seems to be a lack of consistency between the teachers' classroom practices and their perceptions of learning. Teachers are often unaware that they focus on rote learning. According to the authors, teachers often emphasize that they want to develop understanding as part of meaningful learning with their students.

Another important finding of previous research on teachers' practices regarding summative assessment is that teachers rarely discuss or share their practices with colleagues in the same school (Black & Wiliam, 1998b). Teachers are not only unaware of their colleagues' practices but do not trust assessment results obtained from their colleagues either. However, working together with their colleagues can improve teachers' practices. Harlen (2005) showed a promising effect on teachers' practices when they share their understanding of assessment procedures.

A lack of professional collaboration on assessment practices was also noted in previous research from the USA. Cizek, Fitzgerald, and Rachor (1996) showed a substantial variation in assessment practices between teachers. First, the use of primary sources to develop minor and major assessments varies between teachers. For major tests, teachers rely more on test materials from private publishers than for minor tests. However, not all teachers rely on test items from outside sources to the same degree. For instance, the use of test items from private publishers is influenced by characteristics of a teachers' work experience. Beginning teachers used fewer test items from outside sources and developed more minor tests themselves than did more experienced teachers.

3.1.2 Context: Geography education in the Netherlands

This paper is one part of a research effort designed to explore how summative assessment in pre-vocational geography education in the Netherlands—and

more specifically, internal school-based examinations—can contribute to meaningful learning. Furthermore, this research design explores how geography teachers can be trained and scaffolded to construct, judge and mark test items for internal school-based examinations in pre-vocational geography education in the Netherlands that contribute to meaningful learning.

The examination programme in pre-vocational education in the Netherlands consists of two parts; the first part pertains to about two-thirds of the objectives of the examination programme and is assessed summatively with internal school-based examinations, while the second part pertains to the other one-third of the objectives and is assessed summatively with a national external end-of-school (exit) examination. Both parts, however, contribute equally (50% each) to the overall result.

A previous study at the first stage of the research design provided some insight of how the objectives for the internal school-based examinations are assessed (Bijsterbosch, Van der Schee, & Kuiper, 2017). This study was based on a content analysis of internal school-based examinations in pre-vocational education in the Netherlands. The study revealed that the test items on the internal school-based examinations in pre-vocational education in the Netherlands tend to focus on rote learning. Over 60% percent of the test items appeal to some type of remembering, almost 30% to understanding and only 10% to higher cognitive processes such as applying, evaluating and creating. Most test items reflected the formats for test items that were used in former national external end-of-school (exit) examinations. The outcomes of this first study are in line with previously mentioned outcomes from the literature; that is, teachers tend to focus their classroom assessments on rote learning.

Still, little is known about the practices, beliefs and values of geography teachers regarding the construction of their internal school-based examinations in the Netherlands. This study attempts to provide some insight into what the current practices and dispositions of geography teachers in pre-vocational education in the Netherlands regarding internal school-based examinations are. The main issues with respect to teachers' practices that are investigated here are to what extent do teachers construct test items

themselves, whether they work collaboratively on the examinations and if and how they use instruments such as taxonomies or test matrices.

This study will also provide some insight concerning teachers' dispositions regarding the purpose of these internal school-based examinations and the relationship with the external end-of-school (exit) examination. Moreover, teachers' perceptions of the extent to which they use test items that appeal to higher cognitive processes will be explored. This perception is closely related to one of the three domains of teachers' dispositions: the intellectual domain. Intellectual dispositions have been defined by Schussler et al. (2010, p. 352) as "teachers' inclination to process knowledge of content and pedagogy, their awareness of what the educational context requires for desired learning outcomes to be reached, and their inclination to put their knowledge and awareness to use accordingly in the classroom". The outcomes on teachers' perceptions of the extent to which they use test items that appeal to higher cognitive processes should provide some insight into teachers' intellectual dispositions. The relationship between teachers' perceptions and practices towards test items that appeal to higher cognitive processes will be explored as well.

Finally, the relationships between teachers' practices and some background characteristics will be investigated to explore whether teachers' practices are influenced by age or teaching experience. The research questions for this study are as follows:

- *What are the current practices, beliefs and values of geography teachers in pre-vocational secondary education in the Netherlands regarding internal school-based examinations?*
- *What is the relationship between geography teachers' practices in pre-vocational secondary education in the Netherlands and their perceptions of test items that appeal to distinct cognitive processes in their internal school-based examinations?*
- *What is the relationship between the background characteristics of geography teachers in pre-vocational secondary education in the Netherlands and their practices regarding the construction of school-based examinations?*

3.2 Methodology

This study analysed the responses to a questionnaire that included 21 questions as well as some items concerning the background characteristics of the respondents (such as their age and teaching experience). The questionnaire was divided into four parts.

In the first part, respondents were asked to provide information about the content of the internal school-based examinations in relation to the objectives of the examination programme. This part addresses the content validity of the internal school-based examinations. The second part contained questions about the application of instruments such as test matrices or taxonomies. We also asked the teachers whether they constructed test items themselves or collaboratively and whether they used test items from outside sources such as textbooks. The respondents were also asked to indicate the percentage of test items in their school-based examinations that are self-constructed or come from outside sources. The third part was about the teachers' perceptions with regard to distinct cognitive processes in the internal school-based examinations. The respondents were asked to give an indication of the percentage of test items in their school-based examinations that appeal to distinct cognitive processes. The final part contained questions about the beliefs and conceptions of the teachers concerning the objectives and goals of internal school-based examinations in the examination programme.

The questionnaire was piloted in 2014 on 4 geography teachers and adjusted based on their feedback. Subsequently, an invitation to fill in the questionnaire was sent to geography teachers who worked in pre-vocational education. The questionnaire was published online, and a letter with a hyperlink was distributed by networks of secondary education teacher training institutions, by the newsletter of the Royal Dutch Geographical Society (KNAG) and by a national online community of geography teachers. The data were collected in October and November 2014.

Out of a total number of 729 schools offering the theoretical programme of pre-vocational secondary education in the Netherlands where geography was one of the possible subjects, 74 respondents filled out the online questionnaire. These figures roughly indicate that approximately 10% of the teachers with a group of pupils in the examination programme responded to the questionnaire. Of the 74 respondents, 45 were male, and 29 were female.

All the respondents worked as geography teachers in the theoretical programme of pre-vocational secondary education (VMBO-gt) in the Netherlands. The mean age of the respondents was 40 (*SD* = 10.95). The mean number of years of teaching experience was 13 (*SD* = 9.06). Approximately 81% of the respondents had a bachelor's degree in geography education, and 14% had a master's degree in geography education. The other respondents either had a bachelor's degree in primary education or held no valid qualification (yet).

Questionnaire responses were first analysed on a descriptive level. Then, several tests were performed to explore the correlations between variables (Pearson's PMCC, Spearman's RCC, Chi-Square and Cramer's V) and to explore differences in the means of variables (t-tests, ANOVA). Correlations and differences in means were regarded as significant when $\alpha < 0,05$.

3.3 Findings

This section presents the results for the three research questions. The first part of this section reports the findings concerning geography teachers' practices with respect to the construction of internal school-based examinations in pre-vocational education in the Netherlands and their beliefs and values. The second part reports the findings concerning the relationship between these practices and their perceived appeal to distinct cognitive processes. The third and final part reports the findings concerning the relationship between teachers' practices and their background characteristics.

3.3.1 Current practices, beliefs and values of geography teachers

First, the current practices of geography teachers regarding the construction of school-based examinations were investigated. Respondents were asked to answer questions about the origin of the test items they use and the perceived percentage of test items from different sources. Second, the respondents were asked which instruments they use to determine the content of the internal school-based examinations. Finally, the respondents were asked whether they work collaboratively on the construction of internal school-based examinations.

To the question about the origin of test items in the internal school-based examinations, most teachers responded that they use multiple sources. The

respondents could choose between tests attached to the textbook, self-constructed test items or other sources such as older exams. The results show that respondents use more than one source to compose the school-based examinations. Most respondents use test items from tests attached to the textbook (88%) and also self-constructed test items (73%) (Figure 3.1).

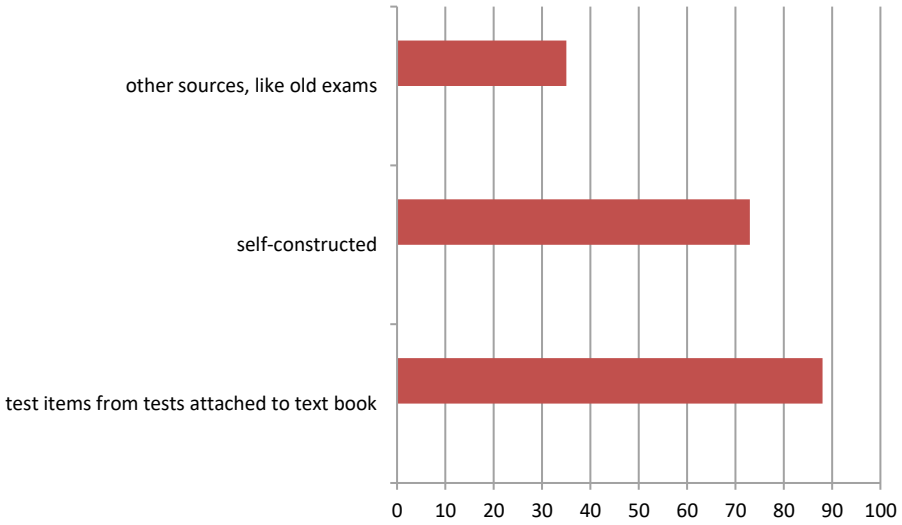


Figure 3.1 Percentage of teachers that use test items from different sources (N=74).

The respondents were also asked to give an indication of the percentage of test items in their school-based examinations related to various sources. Teachers responded that they perceive that almost 45% of the test items in the school-based examinations come from the tests attached to the textbooks and that only 17% are self-constructed test items (Table 3.1).

Table 3.1 Perceived percentage of test items related to the origin of test items (n= 74).

	% test items from textbook	% test items self-constructed	% test items from older exams	% other sources
Mean*	45	17	29	5
SD	30,3	21,5	10,7	8,6

*The sum of the perceived percentages is 96 because the respondents had to fill out an estimated percentage for the various categories and the sum of the estimation of some respondents was less than 100 percent.

To know how geography teachers determine what the content of the internal school-based examinations will be, the questionnaire contained some questions about whether teachers use instruments to determine the content of the examinations, especially a taxonomy or test matrix.

The responses indicate that 78% use some type of taxonomy to construct internal school-based examinations. More than half the respondents use a taxonomy that has become well known in the Netherlands, the so called RTTI taxonomy that consists of four categories: remembering (R), executing a familiar task (T1), implementing an unfamiliar task (T2) and comprehension (I). The other teachers use taxonomies such as Bloom's taxonomy. Approximately 22% of the teachers use no taxonomy.

Responses regarding the application of a test matrix (Table 3.2) show that 38% of the teachers do not use a test matrix to construct the internal school-based examinations. This percentage is slightly higher than the percentage of teachers who do not use a taxonomy, which indicates that some teachers use a taxonomy without a test matrix. The teachers who do use a test matrix rely on the matrix from the instructor's textbook (28%), construct a test matrix by themselves (18%) or construct a matrix in collaboration with their colleagues (10%).

Table 3.2 Percentage of teachers who use a test matrix when constructing school-based internal examinations (N=74).

Test matrix:	Percentage
No test matrix	38%
Textbook matrix	28%
Self-made matrix	18%
Matrix made in collaboration with colleagues	10%
Other	2%
No response on the question	4%

When asked whether teachers work collaboratively on the construction of the school-based examinations, 65% of the respondents reported that they

collaboratively decide what subject specific content in relation to the objectives of the examination programme will be assessed in the internal school-based examinations. Furthermore, 52% of the teachers work collaboratively on the construction of the test items; about two-thirds of the respondents decide in collaboration with their colleagues what the caesura of the internal school-based examinations will be, and 40% work collaboratively on the correction of the school-based examinations.

3.3.2 Teachers' beliefs and values and cognitive processes

To investigate geography teachers' beliefs and values regarding the internal school-based examinations the respondents were asked what the content of the internal school-based examinations should be and what they thought the purpose of the internal school-based examinations should be. Almost all (90%) responded that the objectives of the examination programme that belong to the internal school-based examinations as well as the final exam should be assessed in the internal school-based examinations. Furthermore, 74% of the teachers responded that one purpose of the internal school-based examinations is to prepare the pupils for the external end-of-school (exit) examination by using the same formats for test items in the internal school-based examinations as in the final exam.

To explore the perceptions of teachers as to what extent they use test items that appeal to distinct cognitive processes, the respondents were asked to give an indication of the percentage of test items on their internal school-based examinations related to one of the following cognitive processes: (1) remembering, (2) understanding, (3) applying, and (4) other higher cognitive processes such as evaluating and creating. The results are summarized in Table 3.3.

According to the respondents, the test items in the cognitive dimension were rather equally divided over three of the cognitive processes, with a mean of 34% (*SD* 13.6) of the test items referring to remembering, 32% (*SD* 9.5) to understanding and 26% (*SD* 11.4) to applying. The perceived percentage of test items related to higher cognitive processes such as evaluating and creating was 12% (*SD* 6.5). The range between the minimum and maximum scores is rather wide, especially for remembering, understanding and applying.

Table 3.3 Scores of teachers on the perceived percentage of assessed cognitive processes in internal school-based examinations.

Cognitive processes	N	Minimum	Maximum	Mean*	SD
remembering	74	10	75	34	13.6
understanding	74	10	70	32	9.5
applying	74	5	70	26	11.4
other higher cognitive processes such as evaluating and creating	74	0	35	12	6.5

*The sum of the perceived percentages is 104 because the respondents had to fill out an estimated percentage for the various categories and the sum of the estimation of some respondents was more than 100 percent.

3.3.3 Relationship between current practices and cognitive processes

To explore the relationship between teachers' practices and their perceptions as to the extent to which they use test items that appeal to distinct cognitive processes, several correlation tests were performed. A positive correlation was found between the perceived percentage of test items from tests attached to the textbook and the perceived percentage of test items related to remembering (r = 0.29, p < 0.05, two tailed) and a negative correlation between the perceived percentage of test items from tests attached to the textbook and the perceived percentage of test items related to higher cognitive processes (r = -0.33, p < 0,01, two tailed). A converse pattern was found for the correlation between the perceived percentage of self-constructed test items and remembering (r = - 0,24, p < 0, 05, two tailed) and the perceived percentage of self-constructed test items and higher cognitive processes (r = 0,38, p < 0,01, two tailed).

A t-test for differences in means between the group of respondents who construct test items themselves and the group of respondents who do not shows a significant difference for the perceived percentage of test items related to higher cognitive processes (t(72) = -2,05, p = 0,044). The respondents who construct test items themselves perceive the percentage of test items related to higher cognitive processes as being almost twice as high (M = 9,6, SD = 7,9) as the respondents who do not (M= 5,5, SD = 5,9).

A t-test for differences in means between the group of respondents who use test items from the tests attached to the textbook and the group that does not gives a similar result with respect to the perceived percentage of test items related to higher cognitive processes. The respondents who use test items from the tests attached to the textbook perceive the percentage of test items related to higher cognitive processes as lower (M = 7,8, SD = 7,3) than those who do not (M = 17,5, SD = 6,1). The t-test for this difference in means is also statistically significant ($t(72) = 3{,}18$, $p = 0{,}002$).

No statistically significant correlations were found between the use of a test matrix or taxonomy on the one hand and the perceived percentage of test items related to distinct cognitive processes. Additionally, no correlation was found between the collaborative construction of school-based examinations and the perceived percentage of test items related to distinct cognitive processes, nor for the relationship between the collaborative construction of school-based examinations and the use of instruments such as a test matrix or taxonomy.

3.3.4 Relationship between background characteristics and teachers' practices

Finally, the relationship between some background characteristics such as age and teaching experience and teachers' practices were investigated. To test these relationships, the respondents were clustered into four categories by age and teaching experience. Correlation tests were run to test the relationship between age and teaching experience on the one hand and the origin of test items on the other. The results show a slight positive correlation for age and the number of respondents who use test items from tests attached to the textbook ($r_S = 0{,}24$, $p < 0{,}05$, two tailed) and a slight negative correlation between age and respondents who construct test items themselves ($r_S = -0{,}24$, $p < 0{,}05$, two tailed).

A correlation was also found between teaching experience and the perceived percentage of test items related to the different sources. The results show a slight positive correlation for teaching experience and the perceived percentage of test items from tests attached to the textbook ($r_S = 0{,}25$, $p < 0{,}05$, two tailed). The negative correlation between teaching experience and the perceived percentage of self-constructed test items is somewhat stronger ($r_S = -0{,}32$, $p < 0{,}05$, two tailed).

To explore in more detail the differences for age and the perceived percentage of self-constructed test items, an analysis of variance (ANOVA) was performed. The results show a significant difference by age class (F (3, 70) = 3,37, p = .023). Respondents between 22 and 30 years old perceive the percentage of self-constructed test items as higher than do respondents who are 51 years of age or older (Table 3.4). Post hoc tests show that this difference is statistically significant (p = .016). The difference in perception of self-constructed test items can be explained by 13% for these two age groups (η^2 = .13).

Table 3.4 Age in years and perceived percentage of self-constructed test items.

Age	N	Mean	SD
22-30	20	27	27,7
31-40	19	20	21,0
41-50	19	15	18,7
51+	16	6	6,5
Total	74	17	21,5

An analysis of variance for teaching experience and the perceived percentage of test items from the tests attached to the textbook also shows a significant difference between these groups (F (3, 70) = 3,73, p = .015). Teachers with 25 years or more of teaching experience perceive the percentage of test items from the tests attached to the textbook as higher than 69% (Table 3.5). Post hoc tests note a significant difference between the group of respondents with between 5 and 14 years of teaching experience and the group with 25 years or more of teaching experience (p = .016). The difference between these groups can be explained by 14% for years of teaching experience (η^2 = .14). The difference between the group with between 15 and 24 years of teaching experience and the group with 25 years or more of teaching experience is also significant. Post hoc tests for these two groups show a significant difference (p = .037). The difference between these groups can be explained by 14% for years of teaching experience (η^2 = .14).

Table 3.5 Teaching experience in years and perceived percentage of test items from tests attached to the textbook.

Teaching experience in years	N	Mean	SD
0-4	12	51	28,1
5-14	38	38	29,2
15-24	13	36	30,1
25+	11	69	26,1
Total	74	45	30,3

The relationship between age and teaching experience and the perceived percentage of test items related to distinct cognitive processes was also explored. No significant correlation between age and the perceived percentage of test items related to distinct cognitive processes was found. Additionally, no statistically significant correlation between teaching experience and the perceived percentage of test items related to distinct cognitive processes was found.

Although no statistically significant correlation was found, an analysis of variance revealed that there is a significant difference between the classes on teaching experience and perceived percentage of test items related to higher cognitive processes (F (3, 70) = 3,47, p = .021). Teachers with between 5 and 14 years of teaching experience perceive the percentage of test items related to higher cognitive processes as higher (Table 3.6). Post hoc tests show that the difference between the groups with between 5 and 14 years of teaching experience and the group with 25 years or more of teaching experience on the perceived percentage of test items related to higher cognitive processes is significant (p = .034). The difference between these groups can be explained for years of teaching experience by 13% (η^2 = .13). An ANOVA test on age by class and the perceived percentage of test items related to higher cognitive processes showed no significant results.

Table 3.6 Teaching experience in years and perceived percentage of higher cognitive processes.

Teaching experience in years	N	Mean	SD
0-4	12	6	6,4
5-14	38	11	8,3
15-24	13	5	6,6
25+	11	7	4,0
Total	74	9	7,6

3.4 Conclusions and discussion

3.4.1 A small percentage of self-constructed test items

This study revealed some interesting features about the practices of geography teachers regarding summative school-based examinations in pre-vocational education in the Netherlands and their relationship with meaningful learning. First, teachers rely largely on test items sourced from tests attached to the textbooks in constructing school-based examinations. Almost half the test items originate from these tests, and only approximately 30% of the test items come from other sources such as older national exams. According to the respondents, only 17% of the test items are self-constructed.

These results show that a rather small percentage of test items for internal school-based examinations devised by geography teachers in pre-vocational education in the Netherlands are self-constructed. From this study, it is unknown what the reason might be for this; it could be due to a lack of time, a lack of knowledge, or a low sense of self-efficacy regarding the construction of suitable test items. Although this is unknown for geography teachers in pre-vocational education in the Netherlands, results from research in other subjects in England suggest it might be a combination of lack of skills and lack of confidence (Black et al., 2010).

This is an important issue for further research because this study found some evidence that construction of test items by teachers does have a positive

effect on their perception of test items' contributions to meaningful learning. Teachers who use more self-constructed test items perceive the percentage of test items contributing to meaningful learning as higher. Prior research showed that more than 60 percent of the test items in school-based examinations in pre-vocational education in the Netherlands are related to a type of remembering and, consequently, test items appealing to cognitive processes that transcend rote learning are less common in these examinations (Bijsterbosch et al., 2017). Because teachers perceive the percentage of test items that contribute to meaningful learning as higher when those test items are self-constructed, self-construction of test items might be a promising principle changing teachers' practices with respect to classroom summative assessment.

In this respect, the fact that teachers with the longest teaching experience use more test items from tests attached to the textbooks is not a positive finding. The same relationship can be noticed with respect to the average age of geography teachers in pre-vocational education in the Netherlands. The older teachers become, the more they appear to rely on the tests attached to the textbooks.

Again, from this study it is unknown why older teachers and teachers with more teaching experience use more test items from tests attached to the textbook. Still, assuming that older teachers may function as role models for their younger colleagues, this is not a hopeful result. Functioning as a role model could be an important aspect for stimulating self-efficacy (Bandura, 1989; Schunk, 2003). When older teachers and teachers with more teaching experience rarely construct test items themselves, using them as models becomes problematic. These rather disappointing findings are even stronger when we consider that teachers who construct more test items themselves perceive the percentage of test items that contribute to meaningful learning as higher.

A note of caution with regard to these results is due here because it is unknown whether the teachers who use more self-constructed test items truly construct more test items that contribute to meaningful learning. The respondents were asked only to give an indication of the percentage of test items they thought were related to distinct cognitive processes. It might be the case that teachers who use more self-constructed test items think they

use more test items that contribute to meaningful learning but in reality use the same percentage of test items that appeal to higher cognitive processes as teachers who perceive this percentage to be much lower.

Caution concerning the above results is also imperative because only 74 geography teachers in pre-vocational education responded to the questionnaire. Furthermore, the sample was not fully random. Although the invitation to fill in the questionnaire was published in multiple ways the chosen procedure might have affected the representativeness of the respondents.

Still, the construction of test items by teachers seems to have a positive effect on teachers' practices, beliefs and values regarding classroom summative assessments and their relationship with meaningful learning. This relationship with meaningful learning seems not to be affected, however, by other practices. No evidence was found for the effect to use instruments such as a test matrix or a taxonomy. Additionally, no evidence was found for a positive effect of working collaboratively, although we did expect some positive effects from such collaboration based on the literature (Harlen, 2005).

3.4.2 The impact of high-stakes tests

In this study, some evidence was found for the impact of high-stakes tests on the internal school-based examinations. Almost three quarters of the teachers are convinced that they should use the same formats for test items in their internal school-based examinations as are found in the external end-of-school (exit) examination because they feel that helps to prepare the students for the external end-of-school (exit) examination. Additionally, most teachers find it important to assess the objectives for the external end-of-school (exit) examination as well as the objectives for the internal school-based examinations. From these results the conclusion seems to be justified that a majority of the geography teachers in pre-vocational education are influenced by high-stakes tests such as the external end-of-school (exit) examination, not only with respect to the content validity of the internal school-based examinations but also in regard to formats for test items and the corresponding construct validity.

The results from this study suggest that teachers choose formats for test items that can be considered to give reliable test results. An approach to overcome

these constraints between reliability and validity could be a dependability approach that emphasizes the reinforcing effect of both reliability and validity (Harlen, 2005). Dependability, in this sense, is the sum of reliability and validity and is meant to optimize reliability while ensuring validity, although this is not a calculable sum. This approach ensures the construct validity of the assessment while aiming at the highest possible reliability of the assessment scores (Harlen, 2004a; Wiliam, 1993). The application of test items and having well-specified criteria used to judge them is crucial when applying the concept of dependability (Harlen, 2005).

Black et al. (2011) confirm that this can be a helpful approach, especially when an appeal is made to the beliefs and values of teachers with respect to the purpose of summative classroom assessments. This approach of dependability, however, is not an easy one. Construct validity and reliability are often considered competing concepts, although classical test theory emphasizes that test validity can be reached only when the test scores are to some extent reliable (Van Berkel, Bax & Joosten-ten Brinke, 2014).

3.4.3 Teacher professional development

The results from this study have given some input about how to accomplish change in geography teachers' practices, beliefs and values regarding the purpose and content of classroom summative assessment in pre-vocational education. To change teachers' practices, an enhancement of teachers' knowledge and skills on the relationship between assessment and meaningful learning seems to be needed. Furthermore, a change in the beliefs and attitudes of teachers regarding the purpose of summative classroom assessment seems necessary as well.

To stimulate change in teachers' practices and their beliefs, a professional development programme could be useful (Klenowski & Wyatt-Smith, 2011). Teacher professional development can be achieved in multiple ways. Guskey (1986, 2002) stressed the importance to start with teacher practices. When their practices can be changed, teachers will ultimately change their beliefs and attitudes as well.

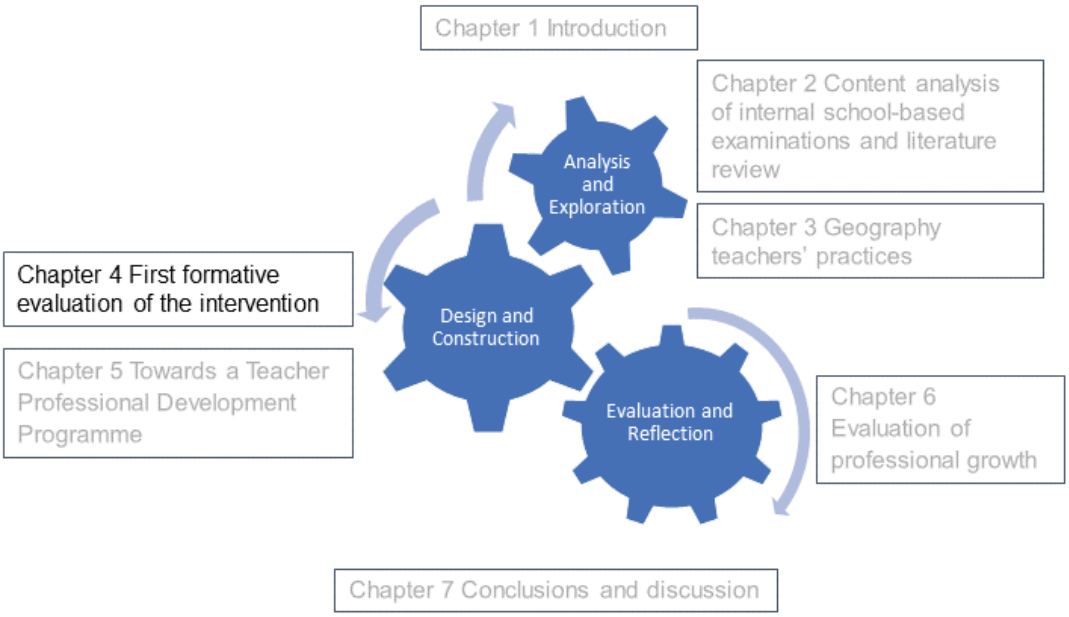
By way of contrast, Clarke and Hollingsworth emphasized multiple pathways to achieve professional growth or development (Clarke & Hollingsworth, 2002). In their interconnected model of professional growth, it does not seem

to be necessary to start with teachers' practices. Change in practices or in beliefs and attitudes can be fostered by stimuli from outside and by reflection and enactment. Enactment of new classroom practices can stimulate change in teachers' beliefs but, on the other hand, a reflective change in beliefs and attitudes can result in new classroom practices as well.

To achieve the professional growth of geography teachers in pre-vocational schools in the Netherlands with respect to their practices, beliefs and values regarding classroom summative assessment further research is needed. Research into a teacher professional development programme in pre-vocational geography education in the Netherlands should provide insight into how to accomplish professional growth for teachers' practices, beliefs and values regarding classroom summative assessments and their relationship with meaningful learning.

Chapter 4

Characteristics of test items focusing on meaningful learning and of the criteria used to judge and mark them: A case study in pre-vocational geography education in the Netherlands*



* Bijsterbosch, H., Béneker, T., Kuiper, W., & Van der Schee, J. A.. Characteristics of test items focusing on meaningful learning and of the criteria used to judge and mark them: A case study in pre-vocational geography education in the Netherlands. Submitted to *European Journal of Geography*, April 2017.

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. To meet these criteria, alignment of learning objectives, instruction and assessment is essential. Furthermore, teachers and students should be scaffolded with strategies to manage the test items and scoring rubrics, and teachers should be involved more closely with the construction of test items and scoring rubrics. In 2016, a small-scale study was conducted with six geography teachers of pre-vocational education to examine which characteristics of test items and accompanying scoring rubrics can contribute to meaningful learning and which strategies can scaffold both teachers and students towards summative assessment that contributes to meaningful learning; the study also asked which test items and rubrics are feasible and practical. The results showed that teachers were most positive about pre-structured test items and the principle of testing what a student knows instead of whether the student knows or can do a predetermined thing. Both teachers and students were also positive about the application of a flow chart to scaffold students in answering these test items.

4.1 Introduction

4.1.1 Summative assessment and meaningful learning

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 et al., 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 tendency of assessments to focus on the recall of knowledge not only affects students' learning but also their motivation for learning. Students who prefer to learn more actively are discouraged by tests that mainly assess their recall of knowledge (Harlen, 2005; Harlen & Deakin Crick, 2002). Effects on learning and motivation seem to be stronger when assessments are used for summative purposes, and especially when they are used for purposes of accountability (Butt et al., 2006). However, some summative assessments, such as those designed for internal use in schools, may have the potential to enhance students' learning processes (Black et al., 2010; Black & Wiliam, 2012).

The learning process also 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). The extent to which these criteria are specified is, according to Harlen, a key variable.

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, especially when it aims to contribute to progress in learning.

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 et al., 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. Before presenting the content of the toolkit and the outline and results of the study, the next section will first provide a brief overview of the literature on the relationship between meaningful learning, summative assessment and students' levels of performance.

4.1.2 Meaningful learning, assessment and levels of performance

Considering the requirement to develop specified criteria for use in judging and marking students' responses on assessment tasks that contribute to students' ability to learn in a meaningful way (Harlen, 2005), several models have been developed to judge students' levels of performance on assessments of learning and progression. 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. At the first stage, logically immature individuals, as they were referred to, tend to answer tautologically. At the second stage, the individual is dominated by the content, and only at the third stage is the individual able to think beyond the given content to evoke possible hypotheses from own experience.

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. To reveal students' capabilities of systematic analysis – i.e., "to identify significant elements, note key relationships and achieve a reasoned explanation" (Rhys, 1972, p. 186) – students were provided with cases that contained information and data that

were not directly related to their own experience. At the lowest level, students were unable to comprehend the geographical environmental context of the case, and they referred to personal experience. At the second level, students were able to refer to essential features of the given case but only used some circumstantial evidence to underline their points. Students who came up with a more adequate analysis based on a combination of several relevant factors performed at the third level. Finally, at the highest level, students came up with a positive judgement or assertion deduced from a guiding hypothesis. From this study, it appeared that students' performances were more closely related to their chronological age and less to their mental age.

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. At the first level of performance in the SOLO taxonomy, students are not able to answer in a structured way (pre-structural); at the second level, students' answers relate to one relevant feature (unistructural); at the third level, students' answers contain multiple features (multistructural), but these are not related to each other. At the fourth level, students' answers reflect relational thinking (relational). At the fifth level, students are able to combine the given information with prior knowledge to deduce more abstract principles and apply these in another situation (extended abstract).

In past decades, some geography educators attempted to make the SOLO taxonomy more subject-specific. Stimpson (1992) turned the SOLO taxonomy into a framework with criterion-referenced questions that were related to typical geographical questions, such as 'What?", "Where?" and "Why there?". Leat and Nichols (2000) compared the stages of the SOLO taxonomy with the observed stages of students working on geographical mysteries. More recently, the SOLO taxonomy has been used to score students' responses on a geographical essay (Munowenyu, 2007).

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. At the lowest level of the hierarchy, the level of mentioning, students are only able to provide incoherent bits of information without a structure. At the level of describing, students are able to give brief descriptions of the topic, which they've derived from the provided material. When students are able to relate, they give a personal explanation but without supportive arguments. At the level of explaining, students use relevant evidence to come up with structured arguments. At the highest level – that of conceiving – students offer individual conceptions that they have developed through continuing reflection (Entwistle & Smith, 2002).

This hierarchy has been reduced by Smith (2002), for modelling purposes, to three levels of understanding: unconnected understanding, descriptive understanding and explanatory understanding. At the level of unconnected understanding, students know facts but do not know how to relate them. When students do bring facts together to form a description, they act on the level of descriptive understanding. Finally, at the highest level of explanatory understanding, students bring facts and descriptions together to form explanations (Smith, 2002).

Table 4.1 presents an overview of the models described above, which have been designed to judge and mark students' levels of performance. These models are compared with each other in an attempt to distinguish general levels of performance. Although the various models did not contain a uniform number of levels, it seems possible to identify five that reflect students' levels of performance. At the lowest level, identified as level 0 in Table 4.1, students are not able to respond correctly. At the first level, students merely repeat what is already stated by information in the test item. When students show, to some extent, the ability to relate this newly provided information to what they have already learned, this could be classified as the second level. In contrast with the second level, students only show an ability to present well-argued reasoning at the third level. At the fourth and highest level, the students show the ability to evaluate or generalize.

Table 4.1 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, comprehen-sive 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	Conceiving: students show individual conceptions, which they've developed through continuing reflection	

4.2 Design of the toolkit and provisional design principles

4.2.1 The toolkit as part of the intervention

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.

4.2.2 The toolkit; three sections

The first section of the toolkit contains examples of test items that appeal to distinct cognitive processes related to meaningful learning. 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 (Appendix D), 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).

Alignment is important to accomplishing educational goals and, in the context of this study, to ensure that test items in summative assessments will reflect the learning objectives and the various cognitive processes that are supposed to stimulate meaningful learning. An important element of the assessment formats of the examples, therefore, is that the test items must contain new information. As noted in their handbook: “If assessment tasks are to tap higher-order cognitive processes, they must require that students cannot answer them correctly by relying on memory alone” (Anderson, Kratwohl, et al., 2001, p. 71). To require that students construct new knowledge and give meaning to it, the assessment formats of the examples were chosen to tap students’ reasoning. To tap students’ reasoning, assessment formats such as

constructed response tasks or essays are highly suitable (Brookhart, 2010). Although other assessment formats, such as multiple choice, can be used to assess higher-order cognitive processes (Brookhart, 2014), these formats are less effective at assessing students' reasoning and at discovering what the learner knows, understands or can do instead of assessing if the learner knows, understands or can do a predetermined thing. Therefore, assessment formats such as multiple-choice questions are not part of the first section of the toolkit. The first section contains examples such as:

- 'Constructed response tasks' that appeal to different types of understanding (Appendix E, example A).
- '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 (Appendix E, example B).
- '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 (Appendix E, example C).
- 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 (Appendix E, example D).
- '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 (Appendix E, example E).

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. To implement materials that are meant to stimulate curricular change, it is necessary to focus on elements that are crucial and vulnerable (Thijs & Van den Akker, 2009). In this toolkit, the section containing scoring rubrics and prescriptions to judge and mark test items is considered to be such an element. 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.

Based on several different approaches from other researchers to identifying levels of performance (Table 4.1), a model was designed to assess, judge and mark students' levels of performance in pre-vocational geography education in the Netherlands (Table 4.2). 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. 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 4.2 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, 2012). 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 4.3) 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 4.3 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.

Classroom discussions between teachers and students are assumed to be helpful in overcoming potential disparities between teachers' target understanding and students' personal understanding (Entwistle & Smith, 2002; Smith, 2002). When assessments tend to test 'if a student knows, understands or can do a predetermined thing' and the student has a different understanding of 'what he knows, understands or can do', classroom discussions can help to discover these disparities and resolve any discrepancies in 'understanding'. In this respect, classroom discussions could also be helpful in stimulating several types of divergent assessments.

4.3 Method

4.3.1 Expert appraisal

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.

A first main outcome of the evaluation was the partial revision of the theoretical background to the taxonomy table of the revised taxonomy of Bloom in the first section. The instructions had been written in rather abstract terms. It was determined that more concrete examples would be helpful to illuminate the theoretical notifications. In addition, the experts suggested extending the explanation of the taxonomy table with more examples from the examination program. Furthermore, an exercise was added in which teachers classified some of the objectives in the taxonomy table.

Overall, the experts were positive about the other elements of the toolkit, particularly about the examples of differential items. The differential items have the potential to differentiate, and not to discriminate, among students'

answers. Furthermore, when meaningful learning and divergent assessment are important principles in the construction of test items for summative assessment, the differential items are seen by the experts as an interesting and promising application of these principles.

The experts were also positive about the model with scoring rubrics. The criteria in this model reflect the principles of constructing and scoring test items that contribute to meaningful learning. Another positive element, according to the experts, was the section of the program with proposed strategies. Yet, the experts suggested placing more emphasis on these strategies. Scaffolding teachers and students was considered quite essential to accomplishing the aims of the program.

4.3.2 Outline of the case study

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.

4.3.3 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 (N=18) 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.

4.4 Results

To answer the research question, the following three sub-questions were used to evaluate the results from the survey, the lesson observations and the interviews. The questions are consistent with the three sections of the toolkit:

- What type of test items appealing to meaningful learning are feasible and why?
- To what extent are the scoring rubrics feasible?
- To what extent are the instruments and strategies for both teachers and students feasible and practical?

4.4.1 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.

4.4.2 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)

4.4.3 Feasibility and practicality of the instruments and strategies

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 4, 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 5, 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.

4.5 Conclusions and discussion

This study was designed to contribute to solving two problems: first, test items in internal school-based examinations in pre-vocational geography education tend to stimulate rote learning instead of meaningful learning; and, second, teachers encounter problems when constructing test items focused on meaningful learning. A toolkit was designed to scaffold teachers to construct and score test items in internal school-based examinations that focus on meaningful learning. To ensure that these examinations contribute to meaningful learning, feasible test items, corresponding scoring rubrics, instruments and strategies were assumed to be essential.

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. 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, 2005b). 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 et al., 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.

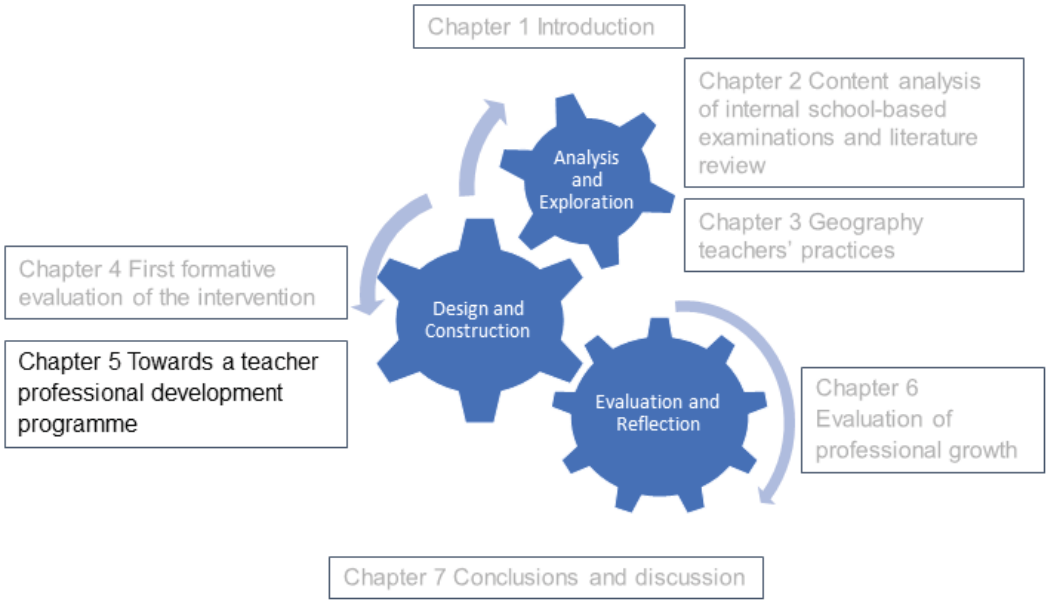
Some final remarks should be made about this study. The current study is limited in several ways. First, only six teachers in pre-vocational education participated. With this small sample size, caution must be applied to the results. Additionally, the selection of participating teachers and students 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.

Chapter 5

Teacher professional growth in summative assessment and meaningful learning: A case study in pre-vocational geography education in the Netherlands*



* Bijsterbosch, H., Kuiper, W., Béneker, T., & Van der Schee, J. A.. Teacher professional growth in summative assessment and meaningful learning: A case study in pre-vocational geography education in the Netherlands. Submitted to *Teacher Development*, April 2017.

Abstract

Teachers' classroom assessment practices tend to encourage rote learning instead of meaningful learning. To enhance teachers' classroom assessment practices with respect to the construction and judgement of test items that contribute to meaningful learning, teacher involvement in assessment construction appears necessary. To foster teacher professional growth in relation to this issue, a teacher professional development programme on summative assessment and meaningful learning in pre-vocational geography education in the Netherlands was designed. In 2016, a prototype of the programme was tested and evaluated in a small-scale case study. The results suggest that the programme contributed to change in teachers' knowledge, skills and practices through the mediating processes of reflection and enactment. The programme, therefore, appeared to be feasible and practical to stimulate teacher professional growth in summative assessment and meaningful learning.

5.1 Introduction

The relationship between teachers' classroom assessment practices and students' learning has received considerable critical attention (Black et al., 2010, 2011; Black & Wiliam, 1998a, 1998b, 2012; Harlen, 2004a, 2005; Harlen & James, 1997). A central issue regarding this relationship is that teachers' classroom assessment practices tend to encourage rote learning instead of meaningful learning. Teachers are not always aware of this tendency (Black et al., 2010, 2011; Harlen, 2004a), which seems to indicate a discrepancy between teachers' knowledge and beliefs and their practices.

Teachers' assessment-related knowledge, beliefs and practices are all part of teachers' assessment literacy. Assessment literacy is a set of competencies including knowledge and skills related to educational assessment and the purposes of assessment (Brookhart, 2011; DeLuca et al., 2016; Xu & Brown, 2016). These competencies have been translated into several standards for assessment literacy, which are usually supposed to serve as a guide for teachers and teacher trainers (Brookhart, 2011; DeLuca et al., 2016).

To enhance teachers' literacy, the closer involvement of teachers in assessment construction appears to be necessary (Harlen, 2005). When teachers are involved in the construction of test items and the corresponding criteria to judge and mark, this will positively affect not only their practices but also their knowledge, beliefs and values. Teachers who are more involved in assessment construction become more aware of the issue of validity with respect to summative assessment (Black et al., 2010).

To date, however, little attention has been devoted to the assessment literacy of geography teachers in the Netherlands and its relationship with students' learning. Previous research by the authors on internal school-based examinations in pre-vocational geography education in the Netherlands has provided some insight into the knowledge, beliefs, attitudes and practices of geography teachers (Bijsterbosch et al., 2017; Bijsterbosch et al., 2016). These studies showed that:

- A majority of test items assess a kind of remembering.
- Teachers rely heavily on test items from external sources, such as tests attached to the textbooks, in the construction of internal school-based

examinations. This tendency appears to be stronger when teachers are older and have more teaching experience.
- A negative correlation exists between the use of test items from external sources and the estimated percentage of test items that contribute to meaningful learning.
- Teachers appear to overestimate the percentage of test items contributing to meaningful learning.
- Teachers' conceptions of the content and purpose of internal school-based examinations appear to be highly influenced by high-stakes tests, especially the external end-of-school (exit) examination. The results indicated that teachers use the same formats in their internal school-based examinations as in the external examinations because they believe that these test items give the most reliable results. Therefore, their constructed internal examinations appeared to be characterised by an emphasis on test items that can be reliably marked at the expense of construct validity.

These two studies were part of an overall research design that aimed to support the professional growth of geography teachers in pre-vocational education with regard to summative assessment and meaningful learning. Professional growth refers to a more-than-temporary change in teachers' knowledge, beliefs and practices regarding the relationship between summative assessment and meaningful learning. This intended change is supposed to have a positive effect on teachers' assessment literacy and teachers' practices in earlier years of pre-vocational education and as such contribute to meaningful learning in geography education.

This paper reports on the second phase of the design research: a formative evaluation of a second prototype of a teacher professional development programme (TPDP) on summative assessment and meaningful learning. The research question for this study is:

How practical and feasible is a teacher professional development programme on internal school-based examinations and meaningful learning in pre-vocational geography education to foster teacher professional growth?

The aim is to examine and to evaluate to what extent this programme and its components are feasible and practical. First, this study should provide insight regarding which instruction materials, instruments and strategies are considered practical. Second, this pilot study should provide insight into the

feasibility of the outline of the whole programme. Third, because this prototype was implemented for the first time in a case study with teachers, this study should also provide insight regarding to what extent the professional growth intended by the programme and the constituting components could be identified.

The outline and the constituting components of the programme, such as instruction materials, instruments and strategies, were implemented and evaluated in a pilot study with six geography teachers in pre-vocational education in the Netherlands in the autumn of 2016. This paper reports on the results of the evaluation of the programme. Before the results are reported, the next section will first give a brief overview of the literature on teacher professional development.

5.2 Teacher professional development

To set up a TPDP, several models have been introduced to develop, analyse and stimulate teacher professional development. In 1986, Guskey proposed a model for professional development that was reaffirmed and slightly adjusted in 2002 (Guskey, 1986, 2002). Crucial to this model (Figure 5.1) is that change in teachers' attitudes and beliefs does not come first; rather, it is altered by the successful implementation of new practices and the consequent improved learning outcomes of students. Or, as Guskey stated: "it is not the professional development per se, but the experience of successful implementation that changes teachers' attitudes and beliefs" (Guskey, 2002, p. 383).

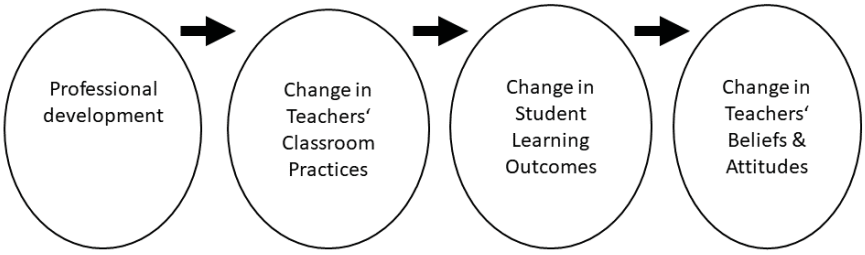


Figure 5.1 A model of teacher change (Guskey, 2002).

Guskey's linear model of teacher change was incorporated by Clarke and Hollingsworth (2002) in their Interconnected Model of Professional Growth

(Figure 5.2). This model contains four domains: the personal domain, the domain of practice and the domain of consequences as teacher-related domains one the one hand and the external domain on the other. Unlike Guskey's model, the Interconnected Model of Professional Growth is non-linear. Professional growth can be achieved in multiple growth pathways when lasting changes in and among the personal domain, the domain of practice and the domain of consequence can be fostered. Changes in these domains are initiated by enactment and reflection and are directed by information or stimuli from the external domain. These stimuli from the external domain are considered crucial in directing teacher learning (Voogt et al., 2011) when they focus on curricular enactment (Ball & Cohen, 1996).

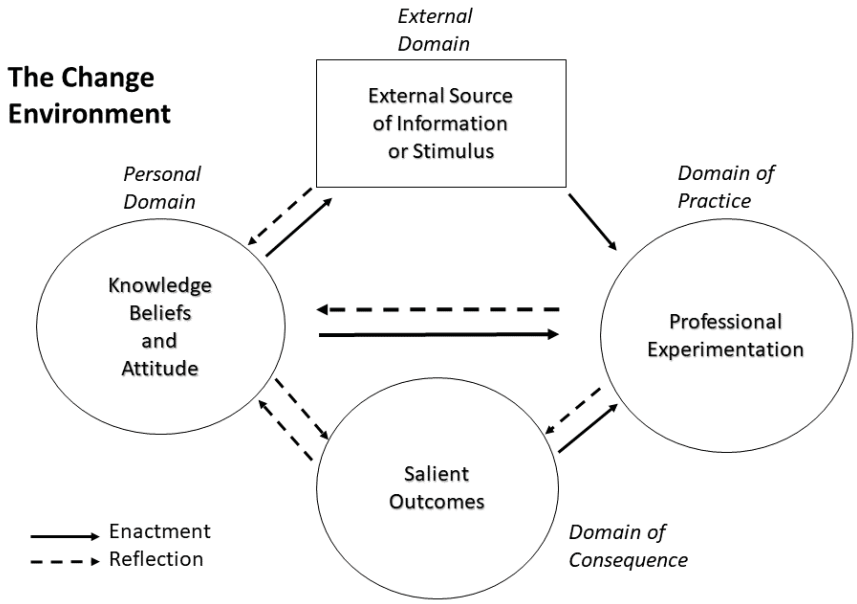


Figure 5. 2 The interconnected model of professional growth (Clarke & Hollingsworth, 2002, p. 951).

In addition, it should be noted that professional growth is a complex process that can be accomplished only when teachers learn (Clarke & Hollingsworth, 2002). Teachers learn in the Interconnected Model of Professional Growth as active learners. Active in this sense means that professional growth is not something that is done to teachers but is the outcome of active engagement and reflective participation. Learners shape their professional growth through

enactment and reflection. Enactment in this sense is distinguished from just 'acting', by deliberately translating a belief or pedagogy into action.

In addition to this fundamental characteristic of teacher learning, some other characteristics can be identified. First, teacher learning can be accomplished best when the learning takes place in authentic situations (Putnam & Borko, 2000; Whitcomb et al., 2009), with a focus on subject specific knowledge (Garet et al., 2001). However, the authentic situation is often influenced by contextual factors that might hinder professional growth, such as a perceived lack of time. It is important to acknowledge these factors as much as possible at the beginning of a professional development programme (Clarke & Hollingsworth, 2002).

Second, teacher learning is positively affected by the collaborative design of curriculum materials. Collaboration implies autonomy on decisions about the construction of materials. Collaborative teacher learning becomes more effective when learning is accompanied and scaffolded by a trainer or coach (Bransford et al., 1999).

Third, teacher learning in teacher development teams is more effective when it is stretched over time and when teachers have the opportunity to implement intended changes (Bransford et al., 1999; Penuel et al., 2007). To implement intended changes, reflection on intended outcomes is essential. This reflection becomes more effective when learners reflect not only on their learning goals and learning strategies but also on their beliefs and values (Korthagen, 2004).

Having defined the essentials of teacher development and teacher learning, the final part of this section addresses ways to plan, diagnose and evaluate the professional growth of teachers. To diagnose and evaluate professional growth, it is important first to define the desired outcomes of a TPDP. Subsequently, it is necessary to determine what the training components should be in order to achieve these desired outcomes. To identify the relationship between desired outcomes and training components, an analytical matrix can be helpful (Harland & Kinder, 1997). Joyce and Showers (2002) developed such a matrix (Table 5.1) to determine how the content of a TPDP can be designed based on desired training outcomes.

Table 5.1 Training components and attainment of outcomes in terms of percentage of participants (Joyce & Showers, 2002).

Components	Outcomes		
	Knowledge thorough	Skill strong	Transfer (executive implementation)
Study of Theory	10	5	0
Demonstrations	30	20	0
Practice	60	60	5
Peer Coaching	95	95	95

Following Joyce and Showers' matrix, the components can be study of theory, demonstration, practice and (peer) coaching. These components of the TPDP are strongly related to the desired outcomes. The percentages in Table 5.1 reflect the percentage of participants likely to attain the desired outcomes when the successive training components are applied. These percentages are an extrapolation made by Joyce and Showers based on research and their experience. Although the estimates are very rough, they give "rules of thumb for estimating the product of training" (Joyce & Showers, 2002, p. 78). When the desired product of training is a transfer to classroom practices, a TPDP should contain peer coaching in addition to instruction materials (theory), demonstration by a trainer and the collaborative practice of participants.

5.3 Goals and provisional design principles for the TPDP

A TPDP was designed to foster teacher professional growth in relation to summative assessment and meaningful learning. The goal of this TPDP is to support the professional growth of teachers regarding the construction and scoring of test items in school-based examinations in pre-vocational geography education in the Netherlands that stimulate meaningful learning. To accomplish this goal, the desired training outcomes of this TPDP are 1) a change in teachers' knowledge and skills, 2) a change in teachers' practices

regarding school-based examinations and 3) a change in teachers' beliefs and values.

To attain these desired training outcomes, it is important to know how lasting changes with respect to teachers' knowledge, beliefs and practices can be realised. After all, although the TPDP aims at changes in teachers' knowledge, beliefs and practices in relation to internal school-based examinations, it is also desirable for the programme to affect their conceptions and practices regarding summative assessment in earlier years of pre-vocational education. The ultimate goal is to stimulate meaningful learning in geography education. Changing conceptions and practices regarding summative assessment – and, more specifically, internal school-based examinations – are means to accomplish this goal.

To achieve these aims, the TPDP was set up in line with the Interconnected Model of Professional Growth by Clarke and Hollingsworth (2002). This model incorporates change in teachers' knowledge, beliefs and practices through reflection and enactment, and it is, therefore, highly applicable to designing a TPDP on professional growth with respect to assessment literacy.

Following the aims, the TPDP in this study has the following characteristics:

- It provides teachers with instruction materials and instruments that enhance teachers' knowledge and skills with respect to the relationship between summative assessment and meaningful learning.
- It stimulates teachers' core reflection on their beliefs, values and existing knowledge regarding summative assessment and meaningful learning.
- It stimulates enactment of new knowledge and skills regarding summative assessment and meaningful learning through theory, demonstration, collaborative practice and peer coaching.
- It contains strategies to stimulate teacher learning through active collaboration.
- It is situated in teachers' classroom practice but also provides teachers with the opportunity to work collaboratively with colleagues in other contexts outside their own school in order to enable future transfer of knowledge.

5.4 Methodology

This study is part of an educational design research (EDR). In the first phase of the EDR, a content analysis on test items in internal school-based examinations, a questionnaire among geography teachers in pre-vocational education and panel interviews with experts and geography teachers were used to analyse and explore the problem. The outcomes of the first phase, and a literature study on this issue, were used to design the TPDP. In the spring of 2016, a first prototype of the TPDP was evaluated with four experienced teacher educators for expert appraisal. The evaluation was formative and focused on the relevance, consistency and practicality of the first prototype. The outcomes were used to finetune the prototype. The main outcomes of the expert appraisal were:

- The experts agreed that intended professional growth can be accomplished not by a fixed sequence of events but through the interplay of stimuli from the external domain, teachers' beliefs and practices, and students' learning outcomes.
- The experts emphasised the importance of the initial stage of the programme. Teachers should become aware of the goals of the programme and should share the goals of the programme. In addition, the experts suggested devoting more attention in the initial stage of the programme to the problem, i.e., that in school-based examinations, a relatively high percentage of test items assess a kind of reproduction.
- The experts were positive about the sequence of phases in which core reflection on teachers' beliefs was followed by demonstration, collaborative practice and peer feedback.

The amended, second prototype was tested in a pilot study with six geography teachers in pre-vocational education from September to December 2016. All teachers worked in the third grade of pre-vocational education. The first author recruited teacher participants by sending e-mails with an invitation to teachers working in pre-vocational education in the vicinity of the institute of the first author. Teachers were asked to participate in a TPDP on internal school-based examinations and meaningful learning. About 50 teachers were directly invited to participate. Six teachers responded positively to the invitation and participated in this programme.

5.5 Outline of the TPDP

In this TPDP, professional growth was pursued by instruction materials from the external domain, which were supposed to stimulate an extension of teachers' knowledge and skills, and by instruments and strategies that were supposed to stimulate enactment and reflection. To accomplish the goals of the TPDP, the programme was executed in three successive phases (Table 5.2).

Phase I: Acknowledgement of pre-existing knowledge and core reflection

Before the first meeting was held, the participating teachers were asked to fill in an inventory about their conceptions of 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 outcomes of the inventory were discussed with the teachers in the first meeting, and the teachers were asked to reflect on their deepest beliefs and values related to the outcomes. For this reflection, the phase model of core reflection (Korthagen & Vasalos, 2005) was used.

A second important element of the first meeting was the activation of pre-existing knowledge regarding summative assessment and meaningful learning. This step is essential for teachers to be able to extend their knowledge based on new information. Teachers will either add this new information to their existing knowledge or, in case of pre-existing misconceptions, revise their concepts.

The next step was instruction on how summative assessment could contribute to meaningful learning. To support the instruction, examples of test items were aligned with the learning objectives by a taxonomy table based on the revised Bloom's taxonomy (Anderson, Kratwohl, et al., 2001). This taxonomy table was somewhat adapted to the context of geography education in the Netherlands.

After the first meeting, the teachers had two weeks to practice at home with some of the test items and align them with the objectives. To help them align the objectives with the test items, the taxonomy table was provided as an instrument with the intention to stimulate teachers' enactment. Attention to

enactment is vital to accomplish change in teachers' practices (Ball & Cohen, 1996).

Table 5.2 Outline of the TPDP.

Phase I Acknowledgement of pre-existing knowledge and core reflection
1st meeting:
• Activation of pre-existing knowledge • Reflection on beliefs and values regarding summative assessment and geography education • Instruction and demonstration on the relationship between test items and learning
Practice and intended enactment (two weeks, at home)
Phase II Extend and internalise knowledge
2nd meeting:
• Demonstration and instruction on test items and cognitive processes of understanding and applying • Analysis of examples of 'good practices' • Collaborative practice • Introduction of scoring rubrics
Practice and intended enactment (three weeks, at home)
3rd meeting:
• Demonstration and instruction on test items and cognitive processes, focusing on evaluating and creating • Analysis of examples of 'good practices' • Demonstration and instruction on pre-structured test items • Introduction of strategies as a flow chart for students
Phase III Application in authentic context
Practice and intended enactment, and peer feedback (six weeks, at home)
4th meeting:
• Evaluation and discussion on test items • Reflection on teachers' knowledge, beliefs and practices

Phase II: Extend and internalise knowledge

The next step was to extend and internalise teachers' knowledge. To support teachers in extending and internalising their knowledge, the participating teachers were provided with materials, instruction and demonstration on test items focusing on understanding and applying at the beginning of the second meeting. These materials consisted of some theory with respect to summative assessment and meaningful learning and good practices of test items and scoring rubrics.
Modelling, demonstration and practice were key principles in this second meeting. These key principles were applied in three consecutive steps. First, teachers analysed existing examples of test items and scoring rubrics that were supposed to contribute to meaningful learning. Second, the teachers practised in constructing test items themselves. They could use a flow chart for teachers regarding how to construct these items. Third, these test items were discussed at the end of the second meeting with all participants to attain a mutual understanding with respect to this issue.

In between the second and third meetings, the teachers had three weeks to practise with the test items and scoring rubrics. The teachers were supposed to give each other feedback on the items before the items were discussed at the beginning of the third meeting. Peer feedback was meant to stimulate teachers' self-efficacy by encouraging them through feedback, mastery and vicarious experiences (Bandura, 1989; Schunk, 2003).

In the third meeting, pre-structured test items were introduced as examples of items with the potential to appeal to several cognitive processes, including higher-order thinking processes, in a more structured way for students. Furthermore, the teachers practised with more open and complex test items focusing on cognitive processes such as evaluating and creating. Finally, the teachers received instruments and strategies to scaffold students, such as a flow chart.

Phase III: Apply in an authentic context

In the third and final phase, the participating teachers constructed pairwise test items for their internal school-based examinations. To align the test items with the objectives and the cognitive processes appealing to meaningful learning, the teachers were supposed to use the taxonomy table. The

participants were also supposed to provide each other with peer feedback on their self-constructed test items in order to stimulate reflection and enactment. Teachers' reflection and enactment were supposed to be stimulated when they discussed their choices with a peer.

In this phase, the teachers also started to practise with their students. Part of the TPDP were strategies to scaffold students to cope with these items. The teachers used these strategies in their classroom practices to prepare the students for the internal school-based examination. Students' performance in classroom practices was supposed to affect teachers' beliefs towards these test items and consequently their practices.

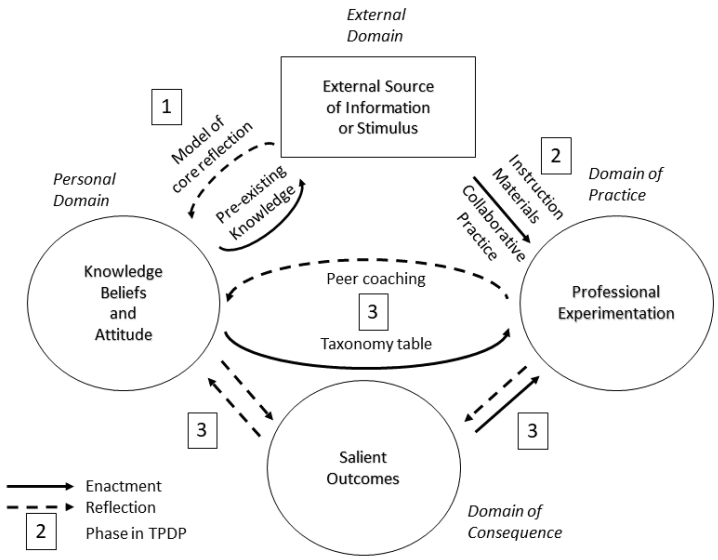


Figure 5.3 Outline of the TPDP in line with the interconnected model of professional growth (model adapted from Clarke & Hollingsworth, 2002).

The constructed test items were discussed with the whole group in the fourth and final meeting. In that meeting, the outline of the TPDP and its constituting components were evaluated with the teachers as well. Before the meeting, the teachers were asked to fill in a survey on the programme and its constituting components. The outcomes of this survey guided the group interview on these issues in the final meeting.

As noted before, the TPDP in this study was set up in line with the Interconnected Model of Professional Growth by Clarke and Hollingsworth

(2002). Figure 5.3 gives an overview of the relationship between this model and the outline of the TPDP in this study. This figure stresses that the TPDP in this study aims not only at change in the three teacher-related domains but also the deliberate stimulation of reflection and enactment.

5.6 Data collection

To collect the data to answer the research question, the teachers were asked in a survey about their perception of the practicality of the instruction materials, the instruments, the strategies and the feasibility of the outline of the programme. Materials, instruments, strategies and outlines are considered practical when teachers perceive an element as realistically usable. "Practicality refers to the extent that users … consider the intervention as clear, usable and cost-effective in 'normal' conditions in the settings for which it has been designed and developed" (Van den Akker, 2010, p. 47).

The teachers filled in the survey anonymously. For each part of the instruction materials, each instrument and each strategy, teachers indicated on a 1-to-5 point Likert scale to what extent the item was practical in relation to the intended outcomes. Second, the teachers were asked why they scored these instruction materials as practical or not practical. At the end of the survey, the teachers were asked to give a score for the whole programme on a 10-point scale and to share their opinion on the programme and the outline of the programme.

The outcomes of the survey were collaboratively discussed and evaluated in a group interview with the participating teachers afterwards. The teachers received an overview of the outcomes before the group interview started. The teachers were asked to reflect together on the outcomes of the survey. The group interview was semi-structured and used the research question as a guideline for the interview. The results from this group interview were also used to answer the research question.

The qualitative data that came from the survey and the group interview were analysed by the first author. Statements of the teachers that referred to the research question were selected. The selection and analysis of the data were approved by another researcher.

Part of the instruction materials was a toolkit with examples of test items, scoring rubrics and strategies for students. An evaluation of the practicality of these geographical test items, scoring rubrics and strategies was reported separately in another paper. The outcomes of this evaluation were only used for triangulation when necessary.

To answer the research question, the following three sub-questions were used to analyse the data from the survey and the group interview:

- To what extent are the instruction materials, instruments and strategies in this TPDP practical?
- To what extent is the outline of this TPDP feasible?
- To what extent could elements of teachers' professional growth, initiated by the instruction materials, instruments or strategies, be identified?

5.7 Results

5.7.1 The practicality of the TPDP instruction materials, instruments and strategies

The teachers reported positive perceptions of the instruction materials, especially the examples of test items and scoring rubrics appealing to meaningful learning. The teachers reported in their responses and during the interview less consensus about the theory in the instruction materials. For some teachers, the theoretical background was interesting but time-consuming and less practical. Others mentioned that the theoretical background was practical for them because it helped them become aware of what they were doing. One of the teachers summarised this issue as follows:

"… you are not just practicing, but you also become aware of why you are doing this. And to see lots of examples, that worked for me… You connect (the theory) with your practice. For me, that was a good balance."
(Teacher A in group interview)

In the programme, some instruments and strategies were used with the aim to stimulate teachers' reflection and enactment. One of the instruments to stimulate teachers' enactment was a taxonomy table meant to align the objectives for the internal school-based examinations with instruction and testing. The taxonomy table was something the teachers had not used before. The teachers reported that the taxonomy table was practical but complex.

One of the teachers reported that the taxonomy table is a practical instrument but not intuitive to apply. The information in the taxonomy table is, according to this teacher, 'overwhelming'. Other teachers mentioned in the group interview that the taxonomy table was helpful to elucidate the objectives.

Another instrument that was introduced to the teachers was a flow chart to construct test items. The teachers reported in the survey that this instrument was practical. One teacher explained in the group interview how he had used the flow chart to align the test items with the objectives for the internal school-based examinations. Although the teachers perceived this instrument as helpful, they also admitted that the application of the flow chart in the construction of test items made this process time consuming and therefore less practical.

In addition to these instruments, which were meant to stimulate teachers' enactment, some instruments and strategies were used in the programme with the purpose of stimulating teachers' reflection, such as the inventory and model of core reflection at the beginning of the programme. Some teachers argued that these instruments were practical to apply in the programme, while others were less positive. One of the teachers stated in the survey that the purpose of the whole programme could have been made clearer.

The group discussion at the beginning of the first meeting revealed discrepancies between the outcomes of the inventory and what teachers mentioned when reflecting on their deepest beliefs. Teachers' beliefs about the aim of geography education and the purposes of summative assessment seemed not to be perfectly aligned. All teachers agreed or strongly agreed that problem solving is important in geography education, but in the group discussion, they varied strongly in their valuing of higher cognitive processes, such as evaluating and problem solving, as part of summative assessment.

Regarding the questions about teachers' awareness of limitations with regard to their aims, one of the steps in the model of core reflection, the teachers only mentioned limitations that could be labelled as external factors. Most teachers mentioned a lack of time, students' lack of motivation or a lack of equipment, mainly due to limitations caused by a limited number of atlases or inappropriate classrooms. Other striking limitations mentioned were the influence of high-stakes tests, such as the national exit exam, and language as a barrier for students in pre-vocational education. The teachers strongly

agreed that the national exit examination has a huge influence on the way they construct their internal school-based examinations. Finally, none of the teachers mentioned limitations that could be labelled as related to their own personal knowledge or beliefs.

A common view amongst the teachers was that the strategy of peer feedback was important. Peer feedback was given on multiple occasions. One such occasion was during practice in between the meetings, and another was when teachers constructed test items for the test. They also gave feedback to one another during the discussion on test items in the final meeting. The teachers highly valued peer feedback and discussion on these occasions with respect to their learning:

"To see each other's test items is very useful, according to me… To give each other peer feedback is also helpful."
(Teacher B, responding to the survey)

"Discussion about alignment of test items causes a deepening of your own thinking."
(Teacher C, responding to the survey)

The teachers were less positive about the strategy of the collaborative construction of test items. Some teachers responded that they constructed test items by themselves rather than in collaboration with others. One teacher stated that although he did not construct test items in collaboration with others, this strategy was practical to him.

5.7.2 The feasibility of the outline of this TPDP

The teachers were asked about the extent to which the outline of the programme was feasible. There was a sense of agreement amongst the teachers regarding the feasibility of the programme as a whole. Asked to score the programme on a 10-point scale, all teachers, individually of each other, gave the programme an eight, indicating that the teachers were overall satisfied with the programme. In one case, a teacher explained this score with the following statement:

"I have learned a lot, and not only me but also my students benefit from this."
(Teacher D, responding to the survey)

In the group interview, the teachers shared their opinion on the programme and its constituting components. The teachers highly valued the meetings, especially the second meeting. One of the reasons the teachers mentioned for why they highly valued the meetings was that these meetings gave them a feeling of 'structure'. Another reason mentioned was that these meetings gave them an opportunity to discuss with each other the outcomes of their practice. Therefore, the meetings had, as one teacher mentioned, 'an added value' for the programme.

Although there was less consensus about the constituting components of the programme (see results on sub-question 1), the teachers had only a few suggestions for changing the programme as a whole. A common view amongst the teachers was that they had spent less time on the programme than expected and, therefore, suggested inserting elements to facilitate collaborative work and peer feedback, such as an online community to provide each other with feedback. In the opinion of the teachers, such a community could serve as a 'big stick' to stimulate them to work on the programme at home. The common opinion of the teachers was that peer pressure along with peer feedback would help.

5.7.3 Elements of teachers' professional growth, initiated by the instruction materials, instruments or strategies

To examine to what extent elements of teachers' professional growth could be identified, teachers' responses from the survey and the group interview were analysed using the Interconnected Model of Professional Growth by Clarke and Hollingsworth. Professional growth refers to lasting changes in at least two of the three teacher-related domains through the mediating processes of reflection and enactment. To examine which elements indicate change in one of the domains or indicate reflection or enactment between the domains, all elicitations were analysed and coded with a scheme that reflected change in a domain or the reflective or enactive links between the domains (Clarke & Hollingsworth, 2002).

A number of issues were identified. First, change in the personal domain was expressed several times. Change in the personal domain was mainly suggested with respect to teachers' knowledge and skills:

"The will to work on construction of test items was already present. The knowledge how to do this was less prevalent."
(Teacher C, responding to the survey)

Change in the personal domain was also reflected by more awareness amongst the teachers of the content and purpose of summative assessment and the relationship with this assessment and students' learning. The teachers' comments below illustrate these changes:

"I have become more aware what good summative assessments are like and how to construct them."
(Teacher D, responding to the survey)

"I have become more aware to assess what a student knows, but I have also found ways now how to do this."
(Teacher B, responding to the survey)

A second issue that emerged from the analysis was that elements of change in the domain of practice could be identified. In response to the question about to what extent their tests had changed due to the professional development programme, one teacher reported:

"The whole test has been changed."
(Teacher D, responding to the survey)

This statement suggests that this teacher deliberately changed the test, indicating change in the domain of practice. Responses from other teachers also indicated that teachers had changed their summative assessments. The teachers not only reported this change in practice but in some instances also showed that their assessment practices had been changed deliberately, based on external information, thus indicating enactment between the external domain and the domain of practice:

"I continuously ask myself if students have to relate their pre-existing knowledge with the information in the test item. When they have to do so, this means to me that this is a kind of meaningful learning. The next step is to categorise the test item (understanding, evaluating etc.)."
(Teacher B, responding to the survey)

Teachers showed not only enactive links between their domain of practice and the external domain but also reflective links between the external domain and the personal domain. Some of their responses indicated that the teachers attached value to the information and materials from the external domain, which appeared to influence their beliefs. One of the teachers mentioned:

"Most examples in the materials are useful."
(Teacher E, responding to the survey)

This teacher perceived most examples in the materials as useful. Statements such as this suggest reflective links between the external domain and the personal domain.

On the other hand, this teacher mentioned:

"The last example was more difficult because during the classroom discussion with students, confusion arose about the right answer. For some students, this test item was difficult."
(Teacher E, responding to the survey)

The teachers' beliefs on the practicality of the examples appeared to be affected by students' difficulties with one example, as well. This statement, therefore, indicates that teachers' beliefs are affected simultaneously through reflective links with both the external domain and the domain of consequence.

A third important issue that emerged from the analysis, therefore, was the relationship between the domain of consequence, i.e., how students perceived and valued this new way of testing, and the personal domain or the domain of practice through reflection. These reflective links appeared to work in two ways. First, students' difficulties with certain test items appeared to decrease teachers' enthusiasm to apply these test items (domain of practice). A common view amongst teachers was that not all test items, especially the ones appealing to evaluating or creating, were applicable and practical for students. One teacher reported, with respect to this issue, that these test items are also more demanding for less literate students. Second, other responses indicated reflective links between the domain of consequence and the personal domain. Reflection on student performance appeared to stimulate teachers' appreciation of elements from the external domain. One

teacher referred to the strategies students could use in answering the test items:

"This (the flow chart) is probably more directing, but it makes it more attainable for students to come up with a good answer."
(Teacher C responding to the survey)

Other teachers reported that the use of the flow chart helped students structure their answers and, therefore, improved their performance. Feedback to students on the application of this strategy appeared to improve the quality of their answers, as perceived by the teachers. Some teachers mentioned that they had started to use the flow chart in other classes for this reason. The way students responded to the flow chart, therefore, seemed to have influenced teachers' conceptions positively.

5.8 Conclusions and discussion

In this paper, the aim was to assess to what extent a designed TPDP and its constituting components are feasible and practical to support teacher professional growth regarding summative assessments and meaningful learning in pre-vocational geography education. The design of the programme and instruction materials was based on the outcomes of the analysis and exploration phase of the design study, which showed that teachers hardly construct test items themselves. Second, a content analysis showed that a majority of test items in internal school-based examinations test recalling knowledge. Moreover, when teachers construct test items themselves, they appear to overestimate the percentage of test items appealing to meaningful learning rather than recalling knowledge. The TPDP in this study, therefore, aimed to contribute to a solution to two problems. The first is that in internal school-based geography examinations in pre-vocational education in the Netherlands, a high percentage of test items appeal to recalling knowledge and hardly appeal to the cognitive processes associated with meaningful learning. The second is that teachers hardly construct test items themselves and appear to have problems constructing test items contributing to meaningful learning. Teacher professional development on this problem, therefore, appeared to be necessary.

The results from this study suggest that change in teachers' knowledge, skills and practices through the mediating processes of reflection and enactment, affected by the designed TPDP, could be identified and thus support teacher professional growth. The outcomes of this study suggest that change in the personal domain and the domain of practice is influenced by teachers' interpretation and valuing of student performance.

Teacher professional development in this study was regarded as a non-linear process that depends on the interplay between change in teachers' knowledge, dispositions and practices on the one hand and students' learning outcomes on the other (Clarke & Hollingsworth, 2002). Through reflection and enactment, intended change in one of these domains could foster change in the other domains without a fixed sequence in the TPDP. However, a sequence in the professional growth of the teachers, in line with Guskey's model of professional growth (Guskey, 2002), was apparent. The suggested sequence in this study was a change in teachers' dispositions and practices through the value the teachers attached to student performance.

In their elaboration of the model of professional growth, Clarke and Hollingsworth suggested that a sequence in accordance with Guskey's model could be one of the growth pathways. Nonetheless, according to Clarke and Hollingsworth, teacher professional growth can be accomplished in multiple ways, and "teacher change often involves multiple and cyclical movements between the analytical domains of the teachers' world" (Clarke & Hollingsworth, 2002, p. 961). Therefore, although change in teachers' knowledge, skills and practices in this study appears to have been influenced by student performance, this does not necessarily indicate that this growth pathway should determine teachers' professional growth in future cases. Rather, it suggests that a future TPDP on the relationship between summative assessment and meaningful learning in pre-vocational geography education should contain multiple cycles and offer participants the opportunity to accomplish professional growth "consistent with individual inclinations" (Clarke & Hollingsworth, 2002, p. 962).

In line with these implications for a future professional development programme, it seems important to devote more attention to teachers' individual beliefs and values in the initial stage of the programme. Teachers should not only share the goals of the programme but also become aware of

the identified problem in relation to their own beliefs and values. In addition, teachers should be given the possibility to find ways to address this problem themselves. When these possibilities are offered, teachers will become more actively engaged in the TPDP, which is an important prerequisite for teacher learning.

In addition to reflection on teachers' deepest beliefs and values, it seems to be important to address teachers' awareness of limitations and their core qualities to realise the ideal or desired situation at the beginning of the programme. Because teachers in this study only mentioned limitations that could be identified as external factors, it might be advisable to pay more attention to teachers' core reflection over a longer period of time. Other instruments, such as a weekly or biweekly logbook, might contribute to more teacher reflection on their limitations and core qualities related to their own knowledge and skills.

Another important outcome of this study suggested that scaffolding students with strategies, such as a flow chart, was important for teachers' change in their personal domain. These strategies were perceived by the teachers to help the students answer test items that appealed to cognitive processes belonging to meaningful learning. Furthermore, by providing these strategies, teachers perceived the quality of the given answers to be higher.

A final important result of this study was that, in general, the participating teachers regarded the outline of the TPDP and its constituting components as feasible and practical. However, according the teachers, there is room for improvement. One of the improvements suggested by the teachers is to use an online community to stimulate exchange of constructed test items and feedback between the teachers instead of pairwise construction of test items and peer feedback. Another suggestion was to use this community in order to facilitate a kind of peer pressure to stimulate teachers to work on the programme. Support by a trainer in the collaborative design process of teams could be helpful (Huizinga, Handelzalts, Nieveen, & Voogt, 2014).

The current study showed some promising results regarding the professional development of teachers in relation to summative assessment and meaningful learning and the contribution of the programme to this professional development. A note of caution is due here since only six teachers in pre-vocational education participated in this study. The findings may be limited by

the small sample size. Furthermore, the selection of participating teachers and students was not fully random.

Another source of uncertainty is the type of research applied in this study. The results depended on what teachers reported in the survey and the interviews. Although this method is suitable to explore the reasons for teachers' remarks, more research is needed to verify the results from these groups.

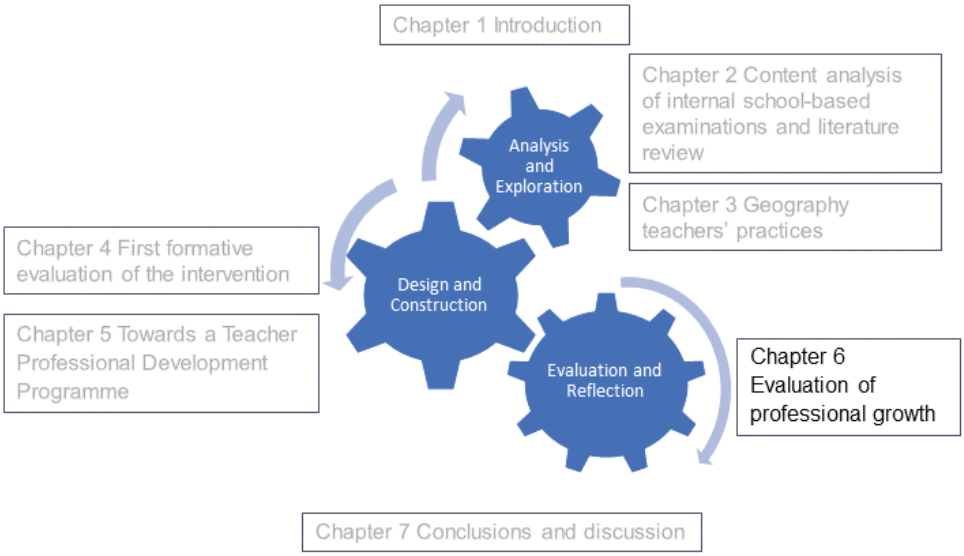
The outcomes of this small-scale study will be used to redesign the programme. The redesigned programme will be tested and evaluated with another group of teachers to examine how and to what extent this TPDP fosters teachers' professional growth.

Acknowledgements

The authors of this paper would like to thank prof. dr. Joke Voogt for her contribution to this chapter.

Chapter 6

Evaluation of a teacher professional development programme on assessment literacy: A case study of pre-vocational geography education in the Netherlands*



* Bijsterbosch, H., Béneker, T., Kuiper, W., & Van der Schee, J. A.. Evaluation of a teacher professional development programme on assessment literacy: A case study of pre-vocational geography education in the Netherlands. Submitted to *The Teacher Educator*, November 2017.

Abstract

Teachers should have the necessary assessment knowledge and skills to contribute to students' learning. To achieve higher mastery levels of assessment literacy, Xu and Brown proposed the TALiP (Teacher Assessment Literacy in Practice) framework. This study provides insight into how a professional development programme designed for teachers contributed to the achievement of higher mastery levels. The outcomes of this study support the value of the TALiP framework and show how the programme evoked a change in teachers' practices and conceptions. A reflection on educational goals and teachers' conceptions, collaborative practice and peer feedback played an important role in fostering higher mastery levels in assessment literacy.

6.1 Introduction

Teachers' classroom assessment practices play a vital role in students' learning. However, students' learning seems to be limited by the same practices, particularly when the assessments are summative. Studies over the past two decades have provided important information about teachers' classroom assessment practices which, formative and summative, tend to stimulate rote learning instead of meaningful ways of learning (Black & Wiliam, 1998a, 1998b; Harlen, 2004b, 2005).

Teachers are not always aware of the fact that their classroom assessment practices have this impact on learning (Black et al., 2010, 2011; Harlen, 2004a). Teachers perceive their assessment practices as being more in line with their educational goals. These educational goals often reflect the intended outcomes in terms of active learning and more demanding cognitive skills than in terms of recalling knowledge.

This discrepancy between teachers' perceptions of their classroom assessment practices and students' learning is of extreme importance because this might indicate a lack of teachers' assessment literacy (Xu & Brown, 2016). Teachers' assessment literacy is mainly referred to as a mixture of knowledge and skills to construct, score and administer assessments in order to use students' results to make decisions (Brookhart, 2011; DeLuca et al., 2016; Xu & Brown, 2016). Several standards for teachers' assessment literacy have been introduced, one of which, introduced by the American Federation of Teachers, National Council on Measurement in Education and National Education Association (1990), has been highly influential (Brookhart, 2011). Brookhart (2011) recently proposed a new set of updated standards for educational assessment knowledge and skills. These standards include knowledge and skills in regard to formative assessment in order to bring teachers' assessment practices in line with learning intentions for students, and to be able to communicate about these practices with students and parents.

Xu and Brown (2016) stressed the importance of in-service teacher education to enhance teachers' assessment literacy. This in-service training should not only focus on a solid knowledge base, but should also pay attention to teachers' beliefs and values because "…the effectiveness of assessment training might be offset by teachers' conceptions, emotions, needs, and prior

experiences about assessment" (Xu & Brown, 2016, p. 155). Similarly, Levy-Vered and Nasser-Abu Alhija (2015) emphasised the relationship between teacher training in assessment literacy in terms of knowledge and skills, and teachers' conceptions of assessments. The results of their study indicated that "… a high degree of assessment literacy is associated with positive conceptions of assessment" (p. 393). Furthermore, the results of this study pointed out that teacher training in assessment literacy had a direct, positive effect on teachers' self-efficacy. In addition, Koloi-Keaikitse (2016) demonstrated that in-service teacher training had a positive effect on teachers' assessment practices.

To inform future teacher-training programmes concerning assessment literacy with a more comprehensive aim, Xu and Brown (2016) developed a conceptual framework for teacher assessment literacy in practice (TALiP). This model includes six components, starting with

1) A knowledge base,
2) Teachers' conceptions of assessment,
3) Institutional and socio-cultural contexts,
4) Teachers' compromises considering their knowledge, their conceptions and the contexts,
5) Teachers' learning, and
6) Teachers' identity (re)construction as assessors.

The model integrates pre- and in-service teacher training and can be used to identify three levels of mastery, namely

1) A basic mastery of educational assessment knowledge,
2) An internalised set of the understanding and skills of the interconnectedness of assessment, teaching, and learning, and
3) A self-directed awareness of assessment processes and one's own identity as an assessor (Xu & Brown, 2016).

Achieving a higher level of mastery is not simply a matter of acquiring more knowledge and skills. Teachers should become more aware of their knowledge, skills and conceptions by reflective practice and active participation in a community (Xu & Brown, 2016). These two elements, reflective practice and active participation, are essential in order for teachers to learn and to reconstruct their identities as assessors.

Teachers' professional development in the TALiP-model, as described above, is in line with other frameworks or models aimed at teachers' professional development. These models, such as Guskey's (1986, 2002) model of teacher change, or the interconnected model of professional growth by Clarke and Hollingsworth (2002), also acknowledge that teachers' professional development should focus on teachers' beliefs and practices in addition to their knowledge and skills. To accomplish their professional growth, teachers should be able to reflect on their beliefs and enact their extended knowledge into practice while paying considerable attention to students' learning outcomes.

To date, little is known about teachers' assessment literacy in the Netherlands and how professional development in this area can be accomplished. To examine how professional growth in terms of teachers' knowledge, beliefs and practices with regard to summative assessment could be accomplished, a teachers' professional development programme (TPDP) was designed. This paper reports on the results of this TPDP. The aim of the TPDP was to accomplish professional growth with regard to teachers' abilities to construct and score school-based examinations in pre-vocational geography education in line with the objectives of these examinations and, as such, contribute to meaningful learning. An evaluation of the TPDP should provide insight into how professional growth in terms of teachers' assessment literacy could be fostered.

To gain insight into how teachers' assessment literacy can be fostered, the design of the TPDP was based on Clarke and Hollingsworth's (2002) interconnected model of professional growth. This model contains four change domains:

1) The external domain, including information and stimuli,
2) The personal domain, incorporating knowledge, beliefs and attitudes,
3) The domain of practice, including teachers' practices, and
4) The domain of consequence including salient outcomes.

The last three domains are considered to be teacher-related domains. Each domain is a 'change domain'. "It is change in external stimuli, change in practice, change in salient outcomes, and change in knowledge or beliefs that constitutes the domain, not information, practice, outcomes, or knowledge per se" (Clarke & Hollingsworth, 2002, p. 953).

Professional growth can be accomplished when lasting changes in at least two of the three teacher-related domains occur. Change in one of these domains can cause change in another through two mediating processes: reflection and enactment. When lasting changes in one domain induce lasting changes in one of the others, this can be identified as a growth network. Growth networks indicate that teachers learn. Teachers learn as active learners and become agents who shape their professional growth through reflection and enactment.

Professional growth in terms of teachers' assessment literacy, therefore, reflects lasting changes regarding teachers' knowledge and skills, their deeply held beliefs and values, their practices and the relationship with students' learning outcomes. To identify professional growth, it is necessary to measure the extent to which change in the personal domain and in the domain of practice has been realised. To identify what evoked these changes and how change was accomplished, it is necessary to gain insight into how these changes were mediated by reflection and enactment because these change sequences, together with the mediating processes of reflection and enactment, constitute the growth networks that are fundamental for teachers' professional development (Clarke & Hollingsworth, 2002). Thus, the aim of this study is to identify if and how these growth networks were evoked by the TPDP. The research question for this study is:

How can a designed teachers' professional development programme on summative assessment and meaningful learning in pre-vocational geography education in the Netherlands contribute to the professional growth of teachers in terms of changes in teachers' practices, knowledge, skills and beliefs through reflection and enactment?

6.2 The teacher professional development programme

6.2.1 Redesign of the programme

As stated previously, the aim of the TPDP in this study was to foster teachers' professional growth with regard to teachers' assessment literacy; more specifically, their competency in constructing and scoring test items that contribute to meaningful learning in pre-vocational geography education. Professional growth was supposed to be attained when changes in teachers'

knowledge, beliefs and practices by reflection and enactment and reflection on students' learning outcomes can be identified.

A previous prototype of the TPDP was tested and evaluated in a pilot study. The pilot study aimed to gain insight into the feasibility and practicality of the outline of the TPDP and its constituting components, such as the external materials, demonstration, collaborative practices and peer feedback. The results of this pilot study were used to redesign the outline and materials of the programme.

One of the most striking outcomes of the pilot study was that a change in teachers' practices or beliefs appeared to be initiated mainly by inferences drawn by the teachers concerning the students' performances. These reflective links seemed to be more dominant than were other reflective or enactive links. These dominant reflective links might have been induced by the outline of the TPDP, which paid less attention to reflection on the goals of the TPDP and the relationship to teachers' beliefs, as the teachers reported after the programme.

Another important outcome was that teachers reported having problems with collaborative practices. Teachers preferred to work individually on the construction of the test items. Nonetheless, they were highly appreciative of feedback from their peers regarding the constructed test items and of discussions about these test items with the entire group.

Teachers in the pilot study also encountered problems with enacting new knowledge and skills. One of the problems was with judging and marking more complex test items focusing on higher order cognitive processes. Despite being provided with a model to score these test items and a flow chart that was supposed to scaffold students, the teachers found it difficult to judge students' responses when the test items demanded more sound and complete answers. Overall, the teachers seemed to need more time to enact new knowledge and skills.

6.2.2 Outline of the programme

To enable teachers to enact new knowledge and skills over a longer period, the TPDP in this study, and in contrast to the pilot study, was designed with a

double loop (Figure 6.1). Both loops contained the three phases that were also used in the pilot study:

1) Acknowledgement of pre-existing knowledge and core reflection,
2) Extending and internalising new knowledge and skills through demonstration, collaborative practice and peer feedback, and
3) Professional experimentation and reflection.

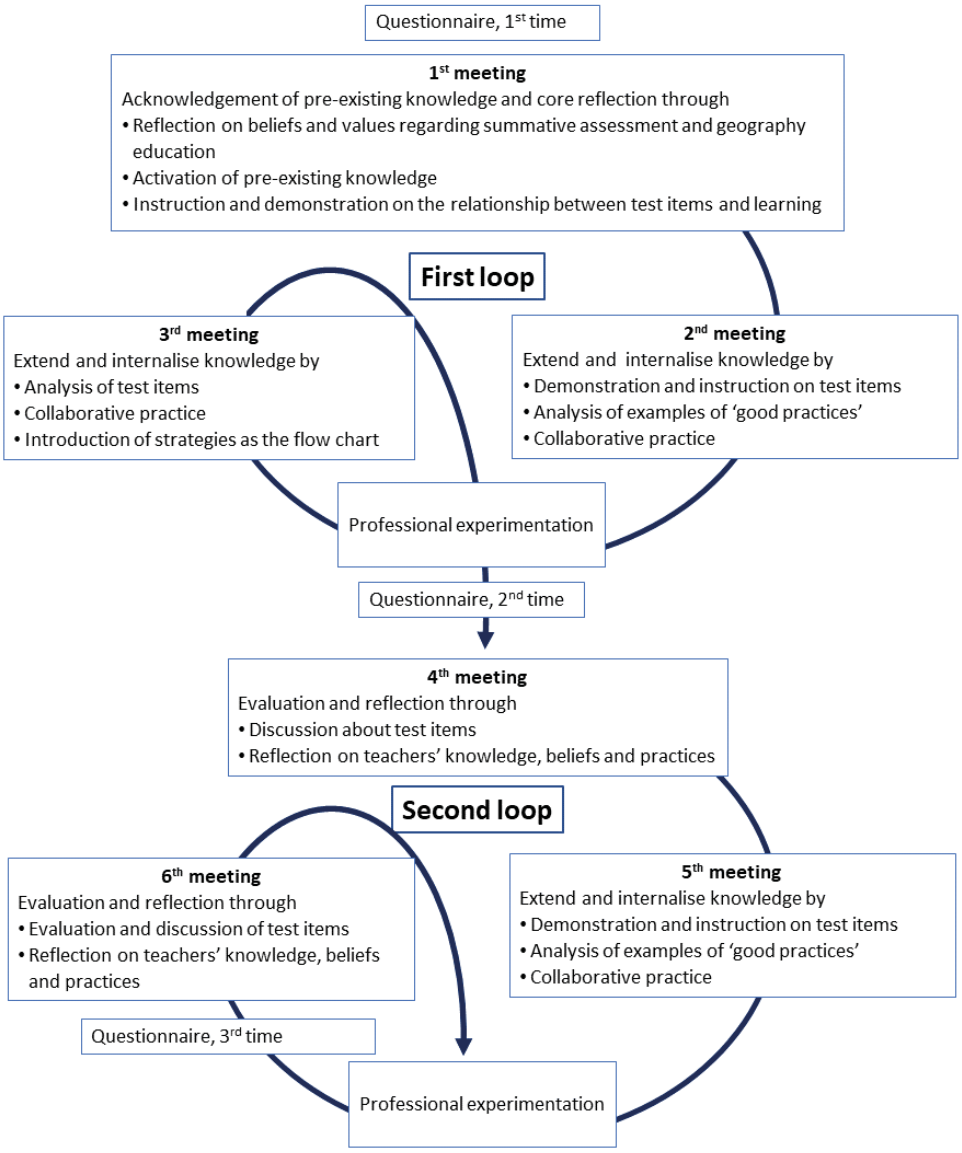


Figure 6.1 Outline of the TPDP.

The double loop was intended to have a positive effect on teachers' learning because it gave teachers the opportunity to implement and enact new knowledge and skills over a longer period. Teachers' learning benefits from an extended programme (Garet et al., 2001; Penuel et al., 2007). Therefore, the purpose was to provide teachers with the opportunity to experiment with new knowledge and skills in the first loop. Once the teachers had extended and internalised their knowledge and skills, the teachers' professional experimentation, enactment and reflection was supposed to be reinforced during the second loop.

In the first loop, the participating teachers attended three meetings. The meetings were led by the first author, who also designed the TPDP and materials for the meetings. In the meetings, the focus was on extending teachers' knowledge and skills with regard to the construction and scoring of test items via instruction, demonstration, collaborative practice and (peer) coaching and feedback. These components are supposed to realise the intended outcomes of a TPDP, based on the matrix by Joyce and Showers (2002).

The instruction and demonstration in these first three meetings were guided by materials from the external domain. These materials are crucial not only to extend teachers' knowledge and skills, but also to stimulate teachers' enactment (Ball & Cohen, 1996; Voogt et al., 2011). Instruction in and the demonstration of good practices were followed by collaborative practice during and between the meetings. The participants gave each other feedback on these first practices, which were then discussed collaboratively. Teachers were also asked to reflect on the external materials and their first practices to stimulate the extension of their knowledge and skills.

To foster a change in teachers' beliefs, reflection on their deepest beliefs was also included in the first meetings. Before the teachers attended the first meeting, they were asked to complete a questionnaire pertaining to their perceived knowledge and skills, as well as their dispositions towards the aim of geography education and summative assessment. Although dispositions in itself have multiple subdomains and it seems difficult to attain consensus on the exact meaning of dispositions (Schussler, 2006; Schussler et al., 2010), the disposition inventory that was used in this study was meant to reflect teachers' beliefs, values and attitudes towards the relationship between the

aim of geography education and summative assessment. Reflection on teachers' dispositions is essential because the implementation of new knowledge and skills depends not only on the quality and feasibility of the materials from the external domain, but also on teachers' dispositions to apply such knowledge and skills (Jo & Bednarz, 2014). To stimulate teachers' reflection on their dispositions, the questions in the core reflection model (Korthagen, 2004; Korthagen & Vasalos, 2005) served as a guide.

In between the third and fourth meetings, within a duration of four weeks, the participating teachers experimented with the construction of test items and the criteria to judge and mark these items. This phase between the first and second loops focused on the implementation of the teachers' knowledge and skills in practice in the internal school-based examinations. Before these examinations were held, the teachers experimented by using classroom practices with the students and scaffolded them via a stepwise flow chart. The students could use this chart to answer the test items.

In line with this chart, the teachers assisted the students to become aware of the criteria for judging and marking the test items. Mutual awareness and understanding between the students and teachers regarding the criteria for judging and marking are crucial in order to avoid misunderstandings and to achieve the intended learning outcomes (Black et al., 2011; Entwistle & Smith, 2002). Should the learning outcomes not meet the intended outcomes due to a lack of mutual understanding, this could have a negative impact on teachers' practices and on their beliefs.

In the fourth meeting, the teachers gave each other feedback regarding the self-constructed test items. Giving each other feedback was not only intended to stimulate the teachers' learning (Hattie & Timperley, 2007; Sluijsmans et al., 2013), it was also aimed to promote the teachers' awareness and self-efficacy. Teachers' self-efficacy will change positively when teachers have mastery experiences and see others having the same experiences (Bandura, 1989, 2001; Schunk, 2003). Teachers will also become more aware of the knowledge and skills they are supposed to master, and will act more purposefully, when their self-constructed materials receive critical attention from a peer (Borko, 2004; Whitcomb et al., 2009).

In the same meeting, students' results for the test items were evaluated collaboratively. A test analysis of test items in the examinations was

conducted to provide insight into how the students performed on the test items. These test items were analysed to determine their p-value and R_{IT}-value (correlation of question score to total examination score). The participating teachers were asked to reflect, individually and collaboratively, on the results of the analyses of students' performances and their scores. The evaluation of the students' results was aimed at stimulating teachers' reflections on the relationship between the results and their practices and beliefs.

Before the second loop began, the teachers were also asked to complete the questionnaire pertaining to their practices and dispositions for the second time. During the first meeting in the second loop, teachers were asked to reflect on the results of the questionnaire. The teachers' reflections were again guided by the model of core reflection.

In the fifth meeting, instruction materials with test items and scoring rubrics focusing on higher order cognitive skills, such as evaluating and creating, were demonstrated and discussed. Based on the demonstration and instruction, the participating teachers practiced constructing these types of test items and the criteria to judge and mark them. Consensus regarding what could be considered to be examples of good practices was important to guide the teachers in the next phase, namely professional experimentation for the upcoming internal school-based examinations. The participants were also asked to complete the questionnaire pertaining to their perceived practices, knowledge, skills and beliefs for the third time.

The teachers' professional experimentation ultimately resulted in newly constructed test items and criteria to judge and mark the next internal school-based examination. In the sixth and final meeting, the constructed test items were evaluated and discussed by the entire group of participating teachers. The group discussion focused on the quality of the constructed test items, why these items could potentially contribute to meaningful learning and how the teachers perceived these test items in relation to the goals of geography education. Similarities and differences in their responses were discussed in order to gain more insight into the teachers' deepest beliefs and values regarding the relationship between internal school-based examinations and meaningful learning.

6.3 Method

6.3.1 Participants and context of the study

As previously mentioned, the evaluation of the TPDP at this stage should provide insight into if and how a designed TPDP could foster teachers' professional growth with regard to summative assessment and meaningful learning in terms of a change in teachers' practices, knowledge, skills and beliefs. Therefore, the evaluation focuses not only on the effectiveness of the programme but also on how and why it works. To examine the effectiveness and impact of a redesigned intervention, a 'real world' try-out or a case study is most suitable (McKenney & Reeves, 2012; Nieveen, 2010).

This study was set up as a case study with a limited number of participants. This design of the intervention made it possible to examine how a TPDP can contribute to the professional growth of teachers regarding their assessment literacy. The limited number of participants allowed the possibility to use interviews to reveal how teachers responded to the programme and how they perceived the intended changes in more depth.

The designed intervention was evaluated from January through June 2017 using a case study of a group of eight geography teachers in pre-vocational education. All the teachers worked in the third grade of pre-vocational education. Two teachers worked at the same school and collaboratively constructed their school's internal school-based examinations. Geography was one of six subjects in the students' final examination. The subjects in the internal school-based examinations in the spring of 2017 varied among three areas of geography: sources of energy, poverty and wealth, and boundaries and identity. Table 6.2 shows an overview of the schools, the teachers and the subjects for the three internal school-based examinations.

Participating teachers were recruited in several ways by the first author. First, teachers working in pre-vocational education in the vicinity of the first author's institute received e-mail invitations to participate. They were asked to participate in a TPDP on internal school-based examinations and meaningful learning. The second method involved recruiting teachers from other teacher-training institutions in the Netherlands. The third method was to distribute the invitation via multiple online geography education communities in the Netherlands.

Overall, fourteen teachers responded positively to the invitation. Ten of these fourteen teachers actually joined the programme. Seven teachers worked in the vicinity of the first author's institute and attended the meetings at that institute. Three teachers came from other parts of the Netherlands and attended the meetings at a central location in the Netherlands. The two groups attended separate meetings, but the meetings were identical in terms of numbers and content.

One teacher in the group of seven had to withdraw during the first loop for personal health reasons. In the group of three teachers, one teacher left the programme after the first loop because of a change in tasks at school and pregnancy. Because these teachers were unable to complete the programme, the results of their school-based examinations and questionnaires at the beginning of the programme were not used in further analysis.

6.3.2 Data collection

To identify changes in teachers' practices, knowledge, skills and beliefs, a cross-sectional qualitative data analysis was performed. This type of analysis enables within- and between-case searches (Spencer, Ritchie, Ormston, O'Connor, & Barnard, 2014). Among the most useful instruments to collect data in this type of analysis are interviews, questionnaires and pre-/post-tests (McKenney & Reeves, 2012).

Changes in teachers' practices were examined by analysing the content of the internal school-based examinations the teachers constructed during the programme. From the participating teachers' seven different schools, an internal school-based examination before the start of the TPDP, after the first loop and after the second loop was analysed using the taxonomy table of the revised taxonomy by Bloom (Anderson, Krathwohl, et al., 2001). All test items in the internal school-based examinations in this table were scored by the first author. A random selection of forty-two test items was scored by another geography teacher educator to achieve intercoder agreement. An interrater reliability test showed that Cohen's Kappa was 0, 76 ($p<0.001$) for the scores of the test items in the taxonomy table for the cognitive dimension, which indicates a substantial agreement.

Changes in knowledge, skills and beliefs were measured via questionnaires and interviews. The teachers were asked to complete the questionnaire three

times: at the beginning of the programme, after the first loop and at the end of the programme. The questionnaire contained questions about teachers' perceived knowledge and skills regarding assessments, questions about their beliefs regarding the relationship between geography education and summative assessment, and three open-ended questions to identify the extent to which the teachers perceived changes in their practices and conceptions. The questionnaire had been used in the pilot study and was adapted after evaluation and discussion. Additional open questions were added to examine the extent to which teachers perceived a change in their practices, knowledge, skills or beliefs.

At the end of the programme, the participating teachers were interviewed individually. The interview questions were based on the results of the questionnaires and aimed to reveal the extent to which and why the teachers perceived changes in their practices, knowledge, skills and beliefs in more depth. The interviews were interactive in the sense that the follow-up questions were largely driven by what the interviewee had already said (Spencer et al., 2014). The teachers were asked to elaborate on their responses in the questionnaire and to reflect on their practices and beliefs. The interviews were fully transcribed. Fragments and statements in the qualitative data from the interviews were selected when the statements referred to a change in practices, knowledge, skills or beliefs. A selection of these statements was used to illustrate and exemplify the results via quotes, which were translated from Dutch.

6.3.3 Data analysis: potential growth networks

As mentioned earlier, it is also important to identify changes that are more than momentary. Professional growth is reflected by more lasting changes mediated by reflection and enactment. Therefore, at the end of the programme, a non-cross-sectional approach was used to identify potential growth networks of individual teachers. This approach allows researchers to search for common patterns that are specific to particular cases within the whole data set (Spencer et al., 2014).

To identify potential growth networks of individual teachers, the data from the questionnaires and interview for each individual teacher were combined to search for change sequences in the data set. Change sequences were defined as changes in two domains mediated by reflection or enactment.

Table 6.1 gives an overview of the coding scheme that was used to identify these change sequences.

Table 6.1 Coding scheme change sequences.

Change sequence	Label	Illustration
Change in practices through enactment of materials from the external domain.	EP	"I learned to apply the principle that a test item must contain new information in order to tap higher order cognitive processes. Therefore, in a test about living environments in Groningen, I added four photographs to the test item, to stimulate students' reasoning about what they saw on the photographs. "
Change in practices through reflection on students' performances.	SP	"When students' reactions are positive, this gives you the energy to make adaptions to the next test as well."
Change in practices through reflection on or enactment of knowledge, skills and beliefs.	KP	"I have adapted my tests and use more test items that focus on applying or creating."
Change in knowledge, skills and beliefs through reflection on the external domain.	EK	Knowledge: "I have learned that another taxonomy is more useful in geography education." Beliefs: "The trigger for me were the continuous questions from you about my aims with geography education. "
Change in knowledge, skills and beliefs through reflection on practices.	PK	"Well, I simply started to act. I reviewed my own tests critically and compared the test items with other potential items."
Change in knowledge, skills and beliefs through reflection on students' performances.	SK	"In one case, a student used to give very short answers. This time, he referred to the information accompanying the test item and he performed better than on previous tests."

When the change sequences were reported more than once, these more lasting changes were identified as growth networks. This definition is in accordance with Clarke and Hollingsworth's definition (2002, p. 958): "Where data have demonstrated the occurrence of change that is more than momentary, then this more lasting change is taken to signify professional growth. A change sequence associated with such professional growth is termed a growth network".

Figure 6.2A provides an overview of all potential change sequences with reflective and enactive links within the interconnected model of professional growth. Figures 6.2B and 6.2C represent two examples of potential growth networks. Figure 6.2B illustrates a teacher who reported lasting changes in practice through reflection on (extended) knowledge and skills and the enactment of (new) materials from the external domain. Figure 6.2C illustrates a change in the beliefs of a teacher after reflecting on students' outcomes and classroom practices.

The coding of change sequences was done by the first author, and the outcomes were discussed with another experienced teacher educator. Only change sequences that had full agreement between the two were used for further interpretation. The interpretation of potential growth networks was subsequently also discussed with the other teacher educator.

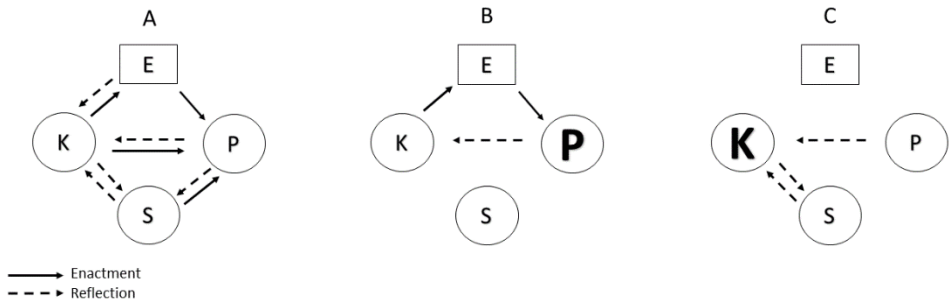


Figure 6.2 Potential Growth Networks (E = external domain; P = professional experimentation; S = salient outcomes; K = knowledge, skills and beliefs). (Adapted from Clarke & Hollingsworth, 2002, p. 959).

6.4 Results

6.4.1 Teachers' practices

To identify the extent to which teachers' practices changed in such a way that they constructed more test items focusing on meaningful learning, the test items in the three internal school-based examinations were scored in the cells of the taxonomy table. Table 6.2 displays the number of test items for each internal school-based examination and a calculated percentage of test items focusing on remembering, understanding, applying, evaluating or creating. The percentages were calculated by multiplying the number of test items per cognitive category by the maximum score that students could achieve for these test items. This number was divided by the total score a student could achieve for the test.

An analysis of these test items showed that the second and third examinations contained a higher percentage of test items that focused on meaningful learning (Table 6.2). The second examinations contained more test items that focused on understanding and applying and fewer test items that involved recalling knowledge. The third examinations also contained more test items that focused on evaluating and creating. Further statistical analyses showed that median score rating of the percentage of test items focusing on the recall of knowledge dropped from 71 per cent at the beginning to 55 per cent after the first loop, and to 42 per cent after the second loop. The difference in time for test items focusing on the recall of knowledge was statistically significant; $\chi^2(2) = 10.286$, $p = 0.006$.

Table 6.2 illustrates that the shift away from the recall of knowledge towards test items with a focus on more meaningful ways of learning mainly resulted in a higher percentage of test items that focused on understanding. Test items that focused on understanding mainly included test items that asked for an explanation or an interpretation of a geographical relationship, using maps or figures in the test items. Another characteristic of these test items is that students were required to use the knowledge they had learned for the test in order to answer the test items correctly.

Although some teachers constructed more test items that focused on the higher cognitive processes, such as evaluating or creating, after the second loop, the number and percentage of these test items was still limited. When

the teachers constructed these test items, the maximum score for the test items in most cases did not really differ from the scores for the other test items, which was the case in examination 3 in school E. Only in a few cases - examination 3 in school A and in school C – did the examination contain one or two test items that focused on evaluating or creating with a considerably higher maximum score, which resulted in a higher percentage of test items focusing on these cognitive processes (12 and 15 per cent, respectively). In these cases, students had to express their own opinions about an issue by using what they had learned and incorporating the new information or figures accompanying these test items.

Table 6.2 (Numbers) and *calculated percentages* of test items (N=558) in school-based examinations focusing on remembering/understanding/applying/ evaluating/creating.

	Examination 1	Examination 2	Examination 3
School A/ Teacher A1	poverty and wealth (13/5/1/0/0) *71/25/4/0/0*	poverty and wealth (13/12/2/2/0) *49/39/5/7/0*	boundaries and identity (9/9/1/1/2) *42/39/3/3/12*
School B/ Teacher B1	sources of energy (19/4/3/0/0) *73/15/12/0/0*	boundaries and identity (18/7/8/0/0) *55/21/23/0/0*	boundaries and identity (26/12/5/1/0) *60/31/8/0/0*
School C/ Teacher C1	poverty and wealth (14/5/1/1/0) *65/24/7/7/0*	sources of energy (11/9/0/0/0) *50/50/0/0/0*	boundaries and identity (8/9/0/0/1) *31/54/0/0/15*
School D/ Teacher D1	sources of energy (21/4/0/0/0) *88/10/0/0/0*	sources of energy (11/5/1/0/0) *63/33/4/0/0*	boundaries and identity (9/9/1/0/0) *42/53/6/0/0*
School E/ Teacher E1 Teacher E2	sources of energy (10/4/0/0/0) *71/29/0/0/0*	sources of energy (14/8/2/0/0) *64/29/7/0/0*	boundaries and identity (7/11/0/1/1) *31/56/0/6/6*
School F/ Teacher F1	sources of energy (17/11/1/0/0) *59/38/3/0/0*	boundaries and identity (13/11/3/0/0) *46/42/13/0/0*	boundaries and identity (21/25/1/1/0) *44/52/2/2/0*
School G/ Teacher G1	sources of energy (22/10/2/1/0) *63/29/6/3/0*	sources of energy (21/6/0/0/0) *78/22/0/0/0*	boundaries and identity (16/21/4/0/0) *39/51/10/0/0*

In the questionnaires and interviews, the teachers were asked their opinions about the extent to which their internal school-based examinations had changed and about the possible reasons for such changes. The teachers' responses concerning their perceptions of changed practices showed that most of the teachers believed that their internal school-based examinations had changed, particularly after the second loop. Most of the teachers commented that these changes had occurred because they had more knowledge and skills and, secondly, because they had become more aware of the content of their assessments. Some teachers also mentioned that their practices had changed not only with regard to the internal school-based examinations for the third grade, but also in other classes and during their classroom practice. As one teacher mentioned:

"I did not construct my summative assessments consciously… I never really asked myself what the test items focused on, what type of pre-existing knowledge the students needed to answer the test items. But I do now. I became more aware and I noticed that I ask more from the students during classroom practices"
(Teacher F1 in the interview).

Other teachers also reported experimentation in other classrooms and with different grades. As another teacher reported:

"Did I change my practices? Yes, I tried to change in other grades as well. I also applied these changes in classes with other grades".

After being asked to elaborate on what he meant by these changes, he added:

" When the students have to use new information, and combine this with their pre-existing knowledge, this generates a more meaningful test item wherein the students really have to be active"
(Teacher D1 in the interview).

A change in teachers' practices could also originate from a change in teachers' self-efficacy in terms of summative assessments. The comment below (by the same teacher) illustrates that his practices mainly changed via a combination of more knowledge, awareness and increased self-efficacy:

"I have learned that you can ask students questions that give them the opportunity to express their own opinions instead of reproduction",

With regard to the question about what helped to change his assessment practices, he replied:

"…the demonstration, pedagogies and knowledge in the programme, the ideas behind it and the confidence to put this in practice"
(Teacher D1 in the interview).

6.4.2 Teachers' knowledge and skills

To determine the extent to which teachers perceived a potential change in their knowledge and skills, the participating teachers were asked to complete the questionnaire pertaining to their perceived knowledge and skills regarding summative assessment and the construction of test items three times. The responses revealed that, in general, most teachers perceived a slight improvement in their skills. A change in perceived knowledge was less obvious. Three teachers reported no change at all in their knowledge during the programme, with the others reporting a slight improvement since the beginning.

The interviews revealed that most teachers perceived an increase in their knowledge and skills in terms of more awareness. Most teachers reported that they had become more aware of the knowledge and skills that are needed to construct test items. During one interview, a teacher revealed that his perceived skills in constructing test items had dropped from 4 to 2 on a 1-to-10-point scale after the first loop, and increased to a 7 after the second. The teacher explained that the drop after the first loop was the result of increasing awareness concerning that which he felt was necessary.

Therefore, according to most of the teachers, it was not merely an extension of their knowledge and skills that contributed to a change in their assessment practices, but also awareness of ways of putting the desired changes into practice. Most of the teachers responded that they had become more aware of the possibilities via materials that were part of the external domain, such as the taxonomy table for the revised taxonomy by Bloom. For example, one of the teachers, when asked if her knowledge and skills concerning summative assessment had changed, responded:

"Yes, very much… I became more critical of the test items in the assessments… I am more aware of the (revised) taxonomy by Bloom. I have more knowledge now about, well, how reproductive my assessments were" (Teacher A1 in the interview).

Another interviewee responded:

"I always thought that students could be challenged more, but I found it difficult to do this. That is the reason I joined this programme. This programme certainly helped, I really started to change my practices. After years of thinking 'I would like to put this more into practice', I noticed now that I am able to do so, to judge students on what they know and can do, instead of punishing them for what they don't know. The most important thing I learned in this programme is that such a system exists to put this into practice"
(Teacher C1 in the interview).

Other teachers also reported about the external domain. For them, the instruments and strategies that were part of the instruction materials were very helpful. In particular, the model with steps to construct and score test items focusing on meaningful learning was mentioned and used by the teachers. The comment below illustrates the practicality of this model:

"I received many good examples (in the TPDP), these models really worked … I like to have structure, so the model provides you with control. Does this count, is this what I want? Okay, then this is what I can do, this is how I can do it"
(Teacher F1 in the interview).

6.4.3 Teachers' beliefs and values

Apart from a change in knowledge and skills, most teachers reported a change in their beliefs and values. In the interviews, most teachers acknowledged that the responses of students contributed to this change in their beliefs and values. When discussing this issue, one interviewee said:

"You become more enthusiastic…what is so nice about these test items, is that students' responses give more energy to you as a teacher as well. The

programme helped in this sense"
(Teacher E1 in the interview).

This view was echoed by another teacher, who related students' responses to the new test items to students' learning:

"I see that the students appreciate it. And, eh, ehm… you notice in terms of application that students understand better … in their long-term-memory …. otherwise the students learn for a single moment, can hold it just for a few days. To focus on the long-term memory, I think that is very important"
(Teacher B1 in the interview).

Another important reason for the change in the teachers' beliefs was awareness of the alignment of summative assessments with the aim of geography education. Most teachers mentioned that they changed their assessments to bring them more in line with their educational goals. The comment below illustrates this increasing awareness:

"What triggered me was the question 'what do you want to achieve with your geography education?' Well, and the answer is not that students should learn 50 concepts by heart"
(Teacher E1 in the interview).

Another teacher commented on this issue:

"I find this very important, learning concepts is important, but also that they are able to use these concepts and give their own opinions"
(Teacher E2 in the interview).

Although most teachers showed a change in their practices related to their beliefs and values, the teachers also reported the constraints on changing their assessment practices. All the teachers mentioned the lack of time as the most important constraint, and some teachers mentioned this issue multiple times. One of the teachers reported this issue in the following words:

"Well, I think more about how I could reformulate test items, so the students will respond more in a way I would like to see. I have improved in this respect. But to adjust my tests, well that's still quite difficult, because than you have to deal with the time. That is certainly a constraint for me, a lack of time"
(Teacher G1 in the interview).

Although the lack of time was mentioned multiple times, most of the teachers realised that, despite this constraint, it might be beneficial to persevere with changing their assessment practices, even when students' responses were not particularly positive at first. As one interviewee explained,

"Of course, it takes time. But on the other hand, when students become more enthusiastic, it is worth the effort."

Interviewer: "And what if the students do not respond enthusiastically?"

"Well, we are very enthusiastic about it, so the students have a bit of bad luck in that case, ha, ha ha. In that case we must continue practicing…And if the students don't have the confidence yet, we should practice more, also during the lessons"
(Teacher E1 in the interview).

Most of the teachers realised what it would take to persevere in the future, namely collaborative practice and discussion. One of the elements of the TPDP that the participating teachers appreciated most was the meetings in which their self-constructed test items were discussed collaboratively. These discussions and peer feedback helped the teachers to reflect on the content and purpose of the test items.

6.4.4 Potential growth networks

The teachers in this study showed different growth networks with regard to changing practices, knowledge, skills and beliefs. Teacher F1's responses, for example, illustrated a change in practice, not only through the enactment of new knowledge, but also via reflection on classroom practices (Figure 6.3A). For this teacher, professional experimentation in the classroom was clearly part of professional growth.

Statements from Teacher A1, Teacher C1 and Teacher D1 illustrated different growth networks (Figure 6.3B). In these networks, reflective and enactive links between the personal domain and the external domain played a more important role. These teachers discussed how their knowledge was extended via materials from the external domain, and they said this made them more aware of what they did and what they wanted to accomplish.

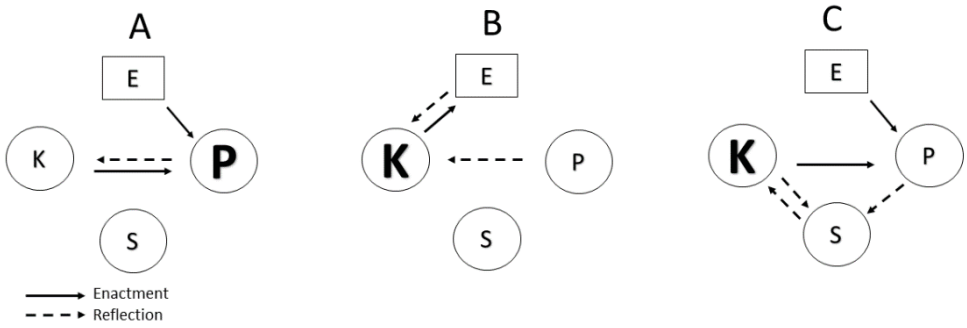


Figure 6.3 Growth Networks of teachers in this TPDP (E = external domain; P = professional experimentation; S = salient outcomes; K = knowledge; skills and beliefs). (Adapted from Clarke & Hollingsworth, 2002, p. 959).

Teachers B1, E1 and E2 indicated growth networks that are better illustrated by Figure 6.3C. For these teachers, changes in their beliefs were connected via reflective links with the domain of salient outcomes and enactive links between the external domain and the domain of practice. In their responses, these teachers showed that the TPDP was a trigger for them to make more connections between the intended objectives and the students' learning outcomes.

For one teacher, Teacher G1, it was more difficult to identify growth networks. Although the teacher constructed more test items focused on understanding in the third school-based examination, this teacher reported the least changes with regard to practices, knowledge, skills or beliefs. This teacher repeatedly referred to a lack of time as a constraining factor to change practices.

Some teachers also expressed that their growth pathway was not always linear. This was illustrated, for example, by Teacher E2 who made it explicit that the first loop in the programme was necessary in order to realise what seemed to be a gap in knowledge and skills. The enactment of the extended knowledge and skills, therefore, seemed to depend on the possibility of a subsequent loop after the first one.

6.5 Conclusions and discussion

The aim of the present research was to examine if and how a designed TPDP could foster teachers' professional growth regarding summative assessment and meaningful learning in terms of changes in teachers' practices, knowledge, skills and beliefs. Teachers' professional growth is considered to enhance their assessment literacy and to contribute to more meaningful ways of testing students' learning. Furthermore, teachers' changing practices should reflect the educational goals of geography education.

This study has identified that the designed TPDP induced a change in teachers' practices. Teachers not only constructed more test items that focused on a type of meaningful learning in their internal school-based examinations, they also showed a change in their classroom practices. An increasing awareness of the content of the test items and the relation to processes in the cognitive dimension caused the teachers to realise that more focus on types of meaningful learning is necessary to align the test items with their educational goals. The TPDP seemed to help the teachers to become more confident to change their practices in line with these goals. These findings are in agreement with the results of the study by Levy-Vered and Nasser-Abu Alhija (2015), which showed that teacher training in assessment literacy had a direct and positive effect on teachers' self-efficacy.

A change in teachers' practices was also related to a perceived change in knowledge and skills. A change in knowledge and skills was evoked by enactment of instruments, such as the taxonomy table for the revised taxonomy by Bloom. The model for constructing and scoring test items also seemed to help the teachers to enact new knowledge and skills. These instruments, which might assist in the constructive alignment of the goals and the assessment practices, are not only helpful for enacting new knowledge and skills, but also to stimulate the teachers' reflections.

Reflection on the teachers' educational goals appeared to be extremely important in order to evoke a change in the teachers' beliefs and values, in addition to a change in practices. The participating teachers showed a change in their beliefs and values regarding summative assessment and meaningful learning that was strongly related to a reflection on the constructed test items in relation to the goals. This reflection on their goals seemed to be even stronger than was the reflection on the students' responses to the test items.

Reflection will not occur spontaneously. To evoke change sequences in teacher professional growth mediated by reflection, the teachers in this TPDP were asked to keep a logbook during the entire duration of the programme. The teachers were asked to fill in the logbook every two weeks. The logbook was guided by questions that focused on the reflection on teachers' goals. Most teachers found it very difficult to regard this logbook as an instrument to stimulate their reflection. Instead, most teachers regarded this instrument as an obligatory exercise that was part of the programme. For them, the collaborative discussions during the meetings were more helpful to stimulate their reflection on the purpose of the test items and the alignment of these test items with their goals.

The teachers' learning also seemed to be hindered by a perceived lack of time. Most of the teachers acknowledged that a change in their assessment practices would take time, for themselves and for their students. Practicing and scaffolding students over a longer period seems to be necessary to achieve these changes. This TPDP was set up with a double loop to stretch out the teachers' learning over time. It appeared that a period of six months could create a change in the teachers' practices, knowledge, skills and beliefs, but this still seems to be a short period in which to accomplish sustainable changes. To achieve these sustainable changes in the longer term, perseverance seems to be another quality that is necessary for teachers' professional growth, in addition to more knowledge, skills and reflections. In this regard, collaborative practice and discussion appeared to be a prerequisite for teachers to persevere.

This research may help us to understand how a TPDP can contribute to the professional growth of teachers with regard to their assessment literacy. Teachers' assessment literacy cannot be enhanced by simply extending teachers' knowledge and skills. Training programmes to equip teachers with the knowledge and skills to meet the standards for assessment literacy are not sufficient.

To achieve professional growth with regard to teachers' assessment literacy, it seems to be important that teachers have the opportunity to enact new knowledge and skills in daily practice. Reflection on these practices helps to further extend, internalise or revise their knowledge and skills. To change teachers' practices, collaborative practice in meetings also seems to be very

important. Similarly, collaborative discussions about these practices, combined with reflection on the purpose of the summative assessments, appeared to stimulate a change in teachers' beliefs.

The findings of this study therefore support the TALiP framework by Xu and Brown (2016) to enhance teachers' assessment literacy via in-service education. As argued by Xu and Brown, an increase in teacher assessment literacy is not simply a matter of acquiring more knowledge and skills. Reflective practice and active participation in real-world settings are key elements for achieving a higher mastery level with regard to teachers' assessment literacy.

The outcomes of this study also support the assumptions underlying the Interconnected Model of Professional Growth by Clarke and Hollingsworth (2002). An important assumption is that professional growth can be accomplished alongside multiple growth pathways. Change in and between domains through reflection and enactment can cause changes in other domains, without a fixed sequence.

Taken together, teachers' professional growth in terms of summative assessment and meaningful learning in pre-vocational geography education appeared to depend on the interplay among the four domains, mediated by reflection and enactment, without a single dominant sequence in changes, and enforced by multiple loops. The outcomes of this study, therefore, strengthen the idea that teacher professional growth can best be achieved via an integrated programme focusing on changes in all teacher-related domains through reflection and enactment. Nonetheless, teachers should have the opportunity to achieve professional growth via their own growth pathways. In order to offer teachers the opportunity to realise growth via their own pathways, the results of this study suggest implementing multiple integrative loops. A programme with multiple loops also offers the possibility of aiming at successively higher mastery levels in assessment literacy in each consecutive loop.

These findings suggest that a future TPDP on summative assessment and meaningful learning in geography education should include multiple integrative loops. However, the focus in each loop could be different depending on teachers' mastery levels. A first loop could focus explicitly on extending teachers' basic knowledge and on collaborative reflection

concerning the educational goals. To achieve mastery level 2, a second loop could focus on the constructive alignment of the educational goals and the teachers' extended knowledge and skills. Instruments such as the taxonomy table for the revised taxonomy by Bloom and a model to construct and score test items could play an important role in a second loop. To achieve the third level of mastery in the TALiP-framework, a third loop could be implemented to stretch out the programme more in time and to focus on the relationship with students' performances and learning. The teachers' responses in this study supported the idea that it is important for teachers to continue to work in groups and to discuss their summative assessments in relation to the students' learning and external demands in order to achieve a reconstruction of their identities as assessors.

The professional growth of geography teachers seems to be necessary in order to bring their assessment practices more in line with educational goals, also from an international perspective. There is no doubt in the global geography education community that summative geography assessments should contain test items that examine the full range of cognitive processes, including evaluating and creating (Bourke & Lane, 2017). Compared to the science framework in the Trends in International Mathematics and Science Study (TIMMS) tests, geographers place even more emphasis on the importance of higher order cognitive processes in tests (Bourke & Lane, 2017).

The study in this paper revealed that a TPDP on summative assessment is a promising method to support geography teachers in constructing more test items that focus on meaningful learning. However, these test items were mainly focused on understanding and applying and placed less emphasis on evaluating and creating. As is known from the previous study of the TPDP, teachers find it difficult to judge students' answers on open test items that focus on evaluating or creating. The teachers' ability to judge and score test items seemed to depend on their conceptual knowledge of the geographical content. To be able to judge students' answers on these test items, a good level of conceptual knowledge is required.

Insufficient conceptual knowledge might have affected the teachers' practices in this TPDP. The participating teachers did not construct many test items that focused on evaluating or creating. To increase the percentage of test items pertaining to these higher order cognitive processes, it seems to be important

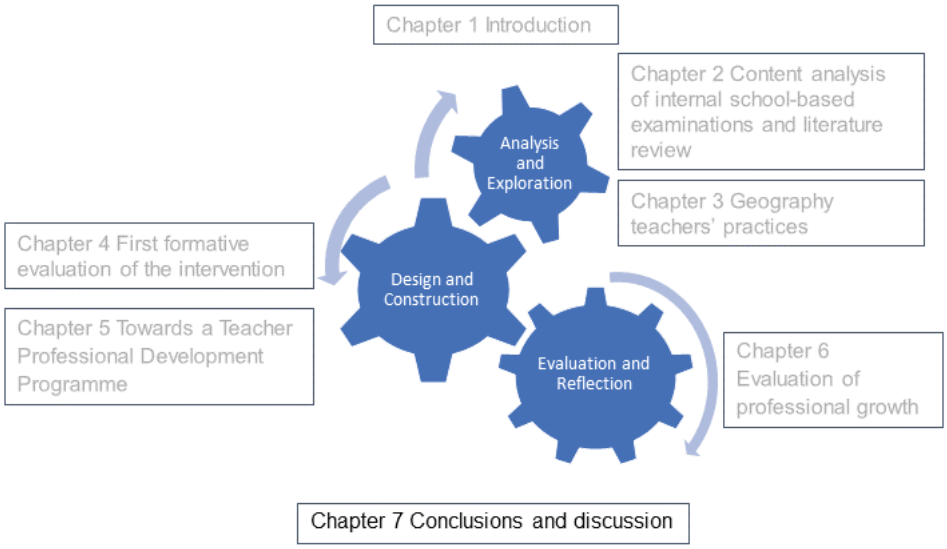
that teachers feel confident and enhance their conceptual knowledge in order to be able to construct and score these test items.

An important limitation of this study is that only eight teachers from seven different schools participated. The findings may be limited by the small sample of the study. A higher number of participants would have strengthened the outcomes. Secondly, some of the data must be interpreted with caution because they were derived from questionnaires and interviews. These qualitative data are useful to reveal teachers' thoughts or beliefs in more depth. However, the findings may have been affected by the interpretation of the statements. More research with a larger number of participants is needed to verify the results of these groups.

More research is also needed to understand better how teachers' mastery levels in assessment literacy develop over a longer period. It would be interesting to assess the effects of long-term collaboration among teachers in a small group with more than two loops in order to examine the potential growth of teachers' assessment literacy. A longitudinal study of teachers in pre-vocational geography education might reveal how this could be accomplished. More research is also required to investigate the effects of this TPDP with larger groups of participants and in other contexts, within and outside of geography education.

Chapter 7

Conclusions and discussion



7.1 Introduction

The aim of this thesis is to examine how teacher professional development regarding the relationship between summative assessment and meaningful learning in geography education can be fostered. Teacher professional development has been defined as lasting changes in teachers' knowledge, skills, beliefs and practices through reflection and enactment. In this research, fostering teacher professional development regarding teachers' assessment literacy was based on the Interconnected Model of Professional Growth by Clarke and Hollingsworth (2002) and the framework of Teacher Assessment Literacy in Practice (TALiP) developed by Xu and Brown (2016). Both models reflect multiple growth pathways and levels in teacher learning with regard to teachers' knowledge, skills and conceptions through practice and reflection. The context of this research is internal school-based examinations in the third grade of pre-vocational geography education in the Netherlands.

The main research question was:

How can geography teachers' professional growth in secondary pre-vocational geography education in the Netherlands be fostered with regard to their practices, knowledge, skills and beliefs in relation to school-based examinations and meaningful learning?

The main research question was divided into the following sub-questions:

1a. What kind of geographical knowledge and which cognitive processes are prevalent in test items in school-based geography examinations in pre-vocational secondary education in the Netherlands?

1b. What kind of beliefs, attitudes and conceptions do geography teachers in pre-vocational secondary education in the Netherlands have upon the school-based geography examinations?

2a. What are the current practices, beliefs and values of geography teachers in pre-vocational secondary education in the Netherlands regarding internal school-based examinations?

2b. What is the relationship between geography teachers' practices in pre-vocational secondary education in the Netherlands and their perceptions

of test items that appeal to distinct cognitive processes in their internal school-based examinations?

2c. What is the relationship between the background characteristics of geography teachers in pre-vocational secondary education in the Netherlands and their practices regarding the construction of school-based examinations?

3. 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?

4. How practical and feasible is a teacher professional development programme on internal school-based examinations and meaningful learning in pre-vocational geography education to foster teacher professional growth?

5. How can a designed teacher professional development programme on summative assessment and meaningful learning in pre-vocational geography education in the Netherlands contribute to the professional growth of teachers in terms of changes in teachers' practices, knowledge, skills and beliefs through reflection and enactment?

The research was conducted as an educational design research study (Chapter 1). To examine how teacher professional development can be fostered, the first step was to explore and analyse current practices of geography teachers in pre-vocational education, their conceptions, and their perceived knowledge and skills. Teachers' practices were analysed via a content analysis of their internal school-based examinations in the third grade. A questionnaire provided insight into the way in which teachers constructed the internal school-based examinations and their conceptions of these examinations. The results of the content analysis (Chapter 2) and the questionnaire (Chapter 3), as did the literature review, informed the design of a teacher professional development programme (TPDP), based on tentative design principles, with the aim of fostering teachers' professional growth. A first formative evaluation of the developed materials (Chapter 4) and the programme (Chapter 5) was conducted to identify how the materials and the programme could be improved. The redesigned intervention was evaluated to provide more insight

into the extent to which teacher professional growth could be identified (Chapter 6).

This final chapter summarises the most important findings of the studies based on the five sub-questions. The section thereafter contains the conclusions and the more conclusive design principles as one of the important yields of this design research. In this section, the main research question will also be answered. The following section contains reflections on the intervention and the design approach. In addition, the limitations of this research will be discussed. The subsequent section contains implications and recommendations with regard to teacher learning and assessment literacy in geography education. Furthermore, recommendations for a future teacher development programme and for further research, practice and policy are suggested. In the final section, the impact of contextual and institutionalised factors on geography teachers' assessment practices and consequences for future constructive alignment of the goals of geography education, the curriculum and assessment will be discussed.

7.2 Main findings

Sub-question 1a: What kind of geographical knowledge and which cognitive processes are prevalent in test items in school-based geography examinations in pre-vocational secondary education in the Netherlands?

Sub-question 1b: What kind of beliefs, attitudes and conceptions do geography teachers in pre-vocational secondary education in the Netherlands have upon the school-based geography examinations?

Geography teachers' summative assessment practices in pre-vocational geography education in the Netherlands appear to be in line with the findings from the literature that revealed that teachers' assessment practices, formative and summative, do not always initiate meaningful ways of learning (Black & Wiliam, 1998a, 1998b; Harlen, 2004b, 2005). More than 60 per cent of the test items in internal school-based examinations focused on the recall of knowledge (Chapter 2). Test items focusing on higher order cognitive skills, such as evaluating or creating, were rarely included in these examinations. These results are consistent with a study on geographical test items in the USA, which pointed out that more than half of the test items in large-scale

standardised assessments and classroom assessments in the USA tested the recall of geographic facts (Wertheim et al., 2013).

The content of the internal school-based examinations seemed to deviate from teachers' goals in geography education. During the panel interviews, the teachers confirmed that their goals go beyond the recall knowledge. Most teachers felt that geography education should aim at the development of understanding geography, and should scaffold students to become citizens who can make informed decisions about their world in the future. Teachers also admitted that, when constructing the examinations, they were influenced strongly by the way in which geographical knowledge is tested in the end-of-school (exit) examinations. This impact of high-stakes tests match those observed in earlier studies (Harlen, 2005; Klenowski & Wyatt-Smith, 2011).

Sub-question 2a: What are the current practices, beliefs and values of geography teachers in prevocational secondary education in the Netherlands regarding internal school-based examinations?

Sub-question 2b: What is the relationship between geography teachers' practices in pre-vocational secondary education in the Netherlands and their perceptions of test items that appeal to distinct cognitive processes in their internal school-based examinations?

Sub-question 2c: What is the relationship between the background characteristics of geography teachers in pre-vocational secondary education in the Netherlands and their practices regarding the construction of school-based examinations?

Teachers' responses from the questionnaire to questions about the purpose of the internal school-based examinations confirmed the outcomes of the panel interviews with regard to the impact of the end-of-school (exit) examinations (Chapter 3). The majority of the teachers believed that this practice would help the students when they applied the same type of test items in their school-based examinations. Teachers' responses during the panel interviews and the questionnaire indicated that the teachers seemed to be more concerned with reliable test results than they were with the validity of their school-based examinations. These concerns affirmed that which was already known from the literature, namely that test items that can be marked

reliably are assumed to fulfil teachers' needs (Harlen, 2005), even when this creates doubt concerning the validity of the test results (Black et al., 2010).

Perhaps the most striking result from the questionnaire was that the teachers rarely constructed the test items themselves. The teachers estimated that 17 per cent of the test items were self-constructed, although the variance between (groups of) teachers was considerable. Most test items in the internal school-based examinations were taken from other sources, such as textbooks. In this regard, another interesting outcome from the questionnaire was that senior teachers and teachers with more experience estimated the percentage of self-constructed test items to be lower than did the younger teachers who had some experience.

Another outcome from the questionnaire was that teachers estimated that their internal school-based examinations contained fewer test items that focused on the recall of knowledge than revealed by the content analysis of the internal school-based examinations. The teachers who constructed more of the test items themselves perceived this percentage to be lower than did the other teachers, which might suggest that teachers who constructed test items themselves for their internal school-based examinations focused less on the recall of knowledge. Stimulating teachers to construct more test items themselves, might help to make the summative assessments more meaningful, as Harlen (2005) suggested previously.

Phase of design and construction of the intervention

The outcomes of the content analysis of the internal school-based examinations, together with the questionnaire and the literature study, resulted in tentative design principles and, consecutively, a first prototype of an intervention intended to foster teacher professional growth in terms of the construction and scoring of test items that focus on meaningful learning. The first prototype for the intervention was based on tentative design principles, derived from the phase of analysis and exploration, describing the function and characteristics of a toolkit (Chapter 4) and a TPDP (Chapter 5). The toolkit contained instruction materials, examples of test items, strategies for students to address these test items, and strategies and instruments for teachers to construct and score the test items. The tentative design principle for the toolkit that consisted of these materials was 'meaningful learning'; in other words, all the test items, instruments and strategies were aligned and

intended to contribute to meaningful learning through a focus on cognitive processes transcending rote learning, the integration of new information and prior knowledge, and the principle of divergent assessment.

The outline of the TPDP was based on the Interconnected Model of Professional Growth by Clarke and Hollingsworth (2002). In this programme, the designed toolkit had the function of the external domain with materials and stimuli to enhance teacher professional growth. The design of the programme was guided by the tentative design principles, with the aim of changing teachers' knowledge, skills, beliefs and practices through enactment of new materials, reflection on these materials and students' performances, and collaborative practice.

The designed prototype was discussed with experts to evaluate the internal structure of the intervention, which is the first step when testing a prototype (McKenney & Reeves, 2012). The expert appraisal focused mainly on the soundness and feasibility of the intervention. The outcomes of the expert appraisal were used to redesign the intervention.

The redesigned intervention was tested and evaluated as part of a small-case study with a group of six geography teachers in pre-vocational education. The evaluation focused on the feasibility and practicality of the toolkit (Chapter 4), and on the feasibility and practicality of the TPDP and its components (Chapter 5).

Sub-question 3: 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?

The evaluation of the toolkit (Chapter 4) showed that the participating teachers perceived pre-structured test items focusing on understanding or applying as being more feasible than were test items focusing on evaluating and creating. The teachers mentioned that they encountered problems in scoring the test items that focused on evaluating and creating. They also mentioned that students had difficulty answering these test items because these were more demanding in terms of literacy and the ability to structure answers.

To scaffold students to structure their answers, teachers could provide a flow chart, which was part of the toolkit, for the students. Both teachers and students were positive about this flow chart. It helped the students to structure their answers, and the teachers perceived that the quality of students' answers increased when they used the flow chart. Although this instrument appeared to be feasible, the participating teachers still expressed a low sense of self-efficacy concerning the scoring of open test items focusing on evaluating or creating.

In line with these outcomes, a model with scoring rubrics to scaffold teachers in scoring the test items was not perceived as being feasible by the teachers. The teachers mentioned that they encountered problems when judging and marking students' answers to more open test items when using the model. However, the teachers did value the constitutive principles of the model. In particular, the principle of divergent assessment (Pryor & Crossouard, 2008) was valued highly. This principle seemed to be promising in terms of bringing teachers' summative assessment practices more in line with their formative classroom assessment practices.

Sub-question 4: How practical and feasible is a teacher professional development programme on internal school-based examinations and meaningful learning in pre-vocational geography education to foster teacher professional growth?

The practicality and feasibility of the TPDP and its components was evaluated with the same group of teachers (Chapter 5). The most important outcome was that the evaluation results suggested that the TPDP seemed to support teachers' professional growth. Results from the questionnaires and interviews suggested that teachers' knowledge, skills and beliefs changed, as well as their practices. Statements from the interviews also showed reflective and enactive links between the newly provided instruments and materials on one hand, and teachers' practices, knowledge, skills and beliefs on the other. Furthermore, reflective links among students' performances and teachers' practices, knowledge, skills and beliefs could be identified.

Some important reflective links between the domain of consequences, in other words teachers' perceptions of students' performances, and the personal domain appeared to be prevalent in a change of teachers' beliefs. These results seemed to confirm Guskey's (2002) model of professional

growth, which emphasised the importance of a change in students' learning outcomes in altering teachers' beliefs and attitudes. This change sequence, which is also one of the possible growth pathways identified by Clarke and Hollingsworth (2002), seemed to be important for the teachers in the programme. However, this does not preclude the possibility of other growth pathways. Statements in the questionnaires and interviews with the teachers also seemed to confirm these alternative growth pathways, particularly the changes in knowledge, skills and practices through enactment. According to the teachers, these changes were mainly enforced by discussions during the meetings of the TPDP and peer feedback.

Sub-question 5: How can a designed teacher professional development program on summative assessment and meaningful learning in pre-vocational geography education in the Netherlands contribute to the professional growth of teachers in terms of changes in teachers' practices, knowledge, skills and beliefs through reflection and enactment?

The evaluation with the first group of teachers was used to redesign the intervention. To give teachers more time to internalise the aims and principles of the programme, the redesigned TPDP was extended over time and consisted of a double loop. In the first loop, the focus was on constructing and scoring test items that focused on understanding and applying and the second loop concentrated on more open and complex test items that focused on evaluating and creating. The outline of the programme containing this double loop was intended to have a positive effect on teachers' self-efficacy by providing a sequence for the introduction of new knowledge and skills, and more time to adopt the constitutive principles.

In line with the suggestions from teachers who participated in the aforementioned case study, the TPDP also focused more on the aims of the programme in relation to the aims of geography education and the purposes of summative assessment. To support teachers' reflections on these aims, other supportive instruments were introduced in the programme, such as a logbook. Furthermore, by extending the programme via a second loop, the teachers had more time to discuss their self-constructed test items collaboratively and to give each other feedback. These strategies were also supposed to stimulate teachers' professional growth.

A group of eight teachers attended the full programme for six months from January 2017. An evaluation of the effectiveness of the programme (Chapter 6) showed that teachers' professional growth with regard to teachers' assessment literacy seemed to have been accomplished. Firstly, teachers' practices had changed. A content analysis of their internal school-based examinations revealed that teachers applied fewer test items that focused on the recall of knowledge in their examinations. The focus in the examinations shifted somewhat to test items focusing on understanding and, to a lesser extent, applying. As anticipated, the participating teachers did not construct more test items focusing on evaluating or creating after the first loop, but did so after the second loop, although the number and percentage of test items focusing on evaluating or creating were still limited. During the interviews, the teachers confirmed this analysis. Most of the teachers mentioned that their practices had changed, particularly after the second loop, which seemed to confirm the expectation that teacher professional growth would benefit from a programme incorporating a second loop. Most of the teachers also reported increased self-efficacy with regard to the construction of test items focusing on meaningful learning.

The results from the interviews and questionnaires indicated that teachers' knowledge, skills and beliefs also appeared to have changed. Teachers mentioned that they had become more aware of the purpose of their internal school-based examinations in relation to the aim of geography education. Reflection on this aim and the purpose of summative assessment seemed to be important in order to alter teachers' conceptions. More focus on constructive alignment (Biggs, 1996) during the programme might have enforced this change in teachers' conceptions.

An analysis of the interviews with the teachers indicated that teachers demonstrated professional growth along several distinctive pathways. For some teachers, a growth pathway triggered by the extension of their knowledge and their enactment of this knowledge in practice seemed to prevail, while others indicated a pathway that was extensively enforced through reflection on students' outcomes. It seemed to be important, however, that teachers were stimulated to enact new knowledge and skills on one hand, and to reflect upon their practices and students' outcomes on the other in integrative consecutive cycles in order to foster changes in their practices and beliefs. These outcomes confirmed the expectation that teacher

professional growth is not linear and can be evoked through multiple pathways. What seems to be important is that teachers are given the opportunity for reflection and enactment concerning instruction materials, strategies, instruments, professional experimentation and reflection on students' learning outcomes, via an integrated programme enforced by multiple loops to prompt teacher professional growth.

7.3 Conclusions

This research showed that a TPDP concerning the construction and scoring of meaningful test items for internal school-based examinations evoked a change in teachers' practices. During the programme, the median score rating for the percentage of test items in the internal examinations that focused on the recall of knowledge dropped from 71 to 42 per cent. These test items were generally replaced by test items that focused on understanding. Therefore, it can be argued that the designed intervention contributed to the solution to the problem of internal school-based examinations in pre-vocational geography education in the Netherlands tending to focus on the recall of knowledge.

However, in this research, the designed intervention was not only intended to contribute to the solution to this problem, but also to inform about the characteristics of the intervention; in other words, how and why the intervention worked and how professional development of teachers could be fostered. These outcomes can be used to stimulate teacher professional development in other situations and contexts. To identify how and why the intervention worked, the main research question had to be answered.

Main research question:

How can geography teachers' professional growth in secondary pre-vocational geography education in the Netherlands be fostered with regard to their practices, knowledge, skills and beliefs in relation to school-based examinations and meaningful learning?

The results of this research supported the idea that professional growth with regard to teachers' knowledge, skills, beliefs and practices could be fostered via diverse growth pathways through reflection and enactment. Reflection and enactment appeared to be stimulated by supportive instruments and

strategies that were designed as part of the external domain in the programme in this research. The outcomes therefore support the Interconnected Model of Professional Growth by Clarke and Hollingsworth (2002) in order to enable professional growth in the teacher-related domains via various pathways.

The outcomes of this research also support the idea that higher mastery levels in assessment literacy can be achieved through in-service teacher professional development programmes, with a focus on teachers' conceptions and practices, in addition to the extension of knowledge and skills. To accomplish these higher levels, it seems to be important that such a programme contains multiple consecutive cycles. Each cycle is integrative in nature, which means that teachers are offered the opportunity for reflection and enactment based on instruction materials, strategies, instruments and professional experimentation in each cycle. Nonetheless, the focus in each cycle could be different in order to scaffold teachers to achieve higher mastery levels stepwise in line with the TALiP framework (Xu & Brown, 2016).

Finally, teachers' professional growth seemed to be stimulated through active participation and collaborative practices related to real-life settings. The teachers valued the discussions and practices during the group meetings in the programme highly. In these meetings, peer feedback appeared to be an important mediator for changes in teachers' conceptions. Teachers' concerns seemed to be enhanced by the fact that peer feedback was given for test items in their internal school-based examinations with a high impact on students' performances.

In conclusion, the results of this research seemed to provide evidence that the professional growth of teachers in assessment literacy can be fostered if the following principles are taken into account:

1) In the initial phase of the programme, a TPDP should focus on reflection on the educational goals in order to achieve a change in teachers' beliefs. Awareness of alignment between the goals with summative assessment practices seems to be an important prerequisite to altering teachers' conceptions towards assessment.
2) In order to accomplish a change in teachers' practices, a focus on the constructive alignment of goals, instruction and assessment practices through instruments that bring these goals in line with the practices are

needed, such as a taxonomy table, a model to construct and score test items, examples of 'good practices' and a flow chart for students.
3) Collaborative practice, peer feedback and discussions about self-constructed test items in group meetings are important elements of a TPDP in order to evoke changes in teachers' knowledge and skills with regard to their own assessment practices.
4) An extended programme with multiple cycles is necessary to accomplish lasting changes in teachers' knowledge, skills, beliefs and practices. To achieve higher mastery levels in assessment literacy, these consecutive cycles are needed to extend the knowledge base and reflect upon it, to align the assessment practices with the goals and instruction and, finally to bring the assessment practices in line with students' learning.
5) To foster teacher professional growth, a TPDP should enable multiple growth pathways by integrating the provision of external materials, the enactment of new knowledge and skills, teachers' professional experimentation and reflection upon students' outcomes in each cycle of a TPDP.
6) A TPDP should be situated in teachers' classroom practice in order to stimulate teachers' self-directed awareness by seeking compromises between their practice and institutional requirements.

7.4 Reflections

7.4.1 The intervention: strengths and weaknesses

One of the more significant findings to emerge from this research was that reflection on the goals of geography education appeared to be extremely important in order for teachers to alter their beliefs or conceptions regarding summative assessment. Teachers reported having become more aware of how to align these goals with their assessment practices. These outcomes support the assumption that enhancing teachers' assessment literacy is not merely a matter of extending knowledge and skills. According Xu and Brown (2016), "yet, the knowledge base is insufficient because these principles only serve as decontextualized guidelines and are not ready-made solutions to problems that arise within complex and diverse classroom assessment scenarios" (pp. 155-156).

In their TALiP framework (Section 6.1), Xu and Brown (2016) also suggested the achievement of higher mastery levels of assessment literacy through reflective practice and active participation by (re)constructing teachers' conceptions and assessor identity in relation to their practice. This process of seeking compromises between external requirements and classroom assessment practices is an iterative process in which teachers continuously question their deepest beliefs and their professional identities as assessors. To stimulate reflection by teachers concerning their deepest beliefs and their assessor identities, one of the tentative design principles that guided the design of the intervention was that the TPDP was intended to stimulate this type of reflection through core reflection (Korthagen & Vasalos, 2005). To reflect on their beliefs that related to their professional identities, the teachers were asked to reflect numerous times during the programme via various instruments. Although, in the interviews and questionnaires, the teachers showed reflective links regarding their beliefs on one hand and their practices, their educational goals and students' performances on the other, connecting these reflections to reflection on teachers' deepest beliefs and values related to their assessor identities was not straightforward.

One of the instruments that was intended to stimulate the teachers' core reflection was a logbook. The teachers were supposed to send in their written logbooks before the next meeting, thus providing the opportunity to discuss certain signalled dilemmas collaboratively. However, most of the teachers encountered several problems in the use of this logbook. The teachers were uncertain of the purpose of this logbook, and some of the teachers seemed to consider the task of completing the logbook as 'homework for the teacher' that they had to perform.

To coerce the teachers to reflect on their beliefs and their identities as assessors, it would appear necessary that a future TPDP should be extended to include an extra cycle or stage that focusses on teachers' assessor identities. In this stage, the focus should be on compromises between external requirements and teachers' practices, as well as on the relationship of these compromises with teachers' deepest beliefs and their professional identities. To achieve the highest mastery level of assessment literacy, this type of reflection seems to be inevitable.

An extra cycle or stage might also be needed to accomplish more awareness among teachers with regard to how their assessment practices can stimulate higher order cognitive processes, such as evaluating and creating. The strength of the designed TPDP in this research seems to be that teachers became more aware of the fact that the majority of the test items in their internal school-based examinations were focused on the recall of knowledge, and that this deviated from their educational goals. However, the teachers changed their examinations mainly by replacing a number of test items that focused on remembering with test items that focused on understanding and, to a lesser extent, on applying. For most teachers, constructing and scoring test items that focused on evaluating and creating seemed to be difficult to accomplish.

One of the causes of this problem was that teachers reported encountering difficulty in scoring these test items. These difficulties seemed to be caused by uncertainty regarding how to score students' diverging responses. To scaffold the teachers to judge and score students' responses, a model to identify levels of performance, which was in line with the criteria to judge and mark, was provided in the instruction materials. The proposed model was based on other models identifying levels of performance (see Chapter 4), which mainly focused on levels that had a high degree of resemblance to the cognitive dimension. Nonetheless, the teachers found it difficult to judge the diverging responses with regard to the marking criteria and the geographical content. The current challenge is thus to redesign a model for judging and scoring these test items in a way that is practical and feasible for the teachers. More practice with these test items and a model for judging and scoring the test items also seems to be necessary.

A promising finding was that both teachers and students valued a flow chart to scaffold students to respond to test items that focused on meaningful learning highly. This instrument appeared to be promising in terms of scaffolding the students and providing them with feedback concerning their responses. The same instrument could be used to challenge students to judge their responses by themselves or as peers in classroom practices. The outcomes could then be discussed in the classroom setting. Therefore, instruments such as the flow chart can scaffold the students to construct responses that are more structured and of higher quality, as the outcomes from this research suggest (Chapter 4). This instrument could also bring

summative assessment practices more in line with formative assessment, and stimulate mutual awareness between teachers and students regarding attainable levels of performance.

7.4.2 Reflection on the research design

The research in this thesis was set up as an EDR. The main goal of an EDR is to develop and implement a solution to an educational problem. (McKenney & Reeves, 2012). In the case of this research, the intervention should contribute to the problem of teachers' use test items with a focus on the recall of knowledge in their internal school-based examinations. Evaluation of the designed intervention should also provide knowledge about how and under what conditions a designed TPDP regarding this problem works. This type of EDR can be described as a development study (Plomp, 2010).

An important characteristic of a development study is that the research is conducted in real-world settings. In the case of this research, the design was tested and evaluated in three consecutive stages of development related to the real-world settings of the teachers: The first prototype was evaluated with experts, the second prototype was tested and evaluated in a small-case pilot-study with six teachers and the third prototype was tested and evaluated in a case study with eight teachers. The focus of the evaluation shifted from the soundness of the intervention to the practicality and feasibility thereof and, finally, to effectiveness of the intervention. This approach is in line with the three stages of design research, namely alpha, beta and gamma testing (McKenney & Reeves, 2012).

By drawing on the concept of a development study in real-world settings, some attention must be paid to the situative perspective of this research. As explained in the introduction (Chapter 1), situative in this context means that teacher learning is considered to be embedded in multiple participative contexts in daily practice and as teachers' individual knowledge construction within these practices. This situative perspective was advocated by Borko (2004) in order to study teacher professional development with a simultaneous focus: a focus on collecting evidence for teachers' individual change in knowledge and practices, and a focus on the contribution of types of collaborative and participative activities to teacher professional development. As Borko pointed out, "to explore the connections among professional development activities and processes on the one hand, and

individual teachers' knowledge and instructional practices on the other, researchers must use the multiple conceptual frameworks and units of analysis that situative perspectives provide" (Borko, 2004, p. 8).

Due to this simultaneous focus on the learning of individual teachers and the relationship to activities in the development programme, the research in this thesis is typical of a Phase I research activity, as identified by Borko (2004). Phase I research activities focus on professional development at a single site and leave the role of the facilitator of the programme and the context unstudied. Consequently, the research in this thesis does not provide evidence that the designed TPDP can be enacted with integrity in other contexts or with other facilitators. Future research at multiple sites with other facilitators should provide evidence concerning if and how the TPDP can be enacted with integrity in other settings, such as other types of secondary education. Only when evidence in a Phase 2 study - a similar TPDP at multiple sites with multiple facilitators - is collected, can a Phase 3 study - a comparative field study of multiple programmes in multiple contexts - be conducted. To facilitate the transfer of insights from this research to other contexts or to a Phase 2 study, a description of characteristics of the intervention is required. These characteristics, or design principles, as described in Section 7.3, provide insight into how and why the intervention works and can be used in other contexts.

Some comments also have to be made regarding the sample size in this research and the nature of the evaluation. The intervention in this research was tested and evaluated with relatively small groups of teachers in a pre-vocational educational setting. To collect evidence concerning how and why the interventions works, a distinction between a formative and summative evaluation of the intervention must be made. As pointed out by Nieveen,

> the function of *formative evaluation* is 'to improve'. It focuses on uncovering shortcomings of an object during its development process with the purpose to generate suggestions for improving it. The function of *summative evaluation* is 'to proof'. A summative evaluation is carried out to gain evidence for the effectiveness of the intervention and find arguments that support the decision to continue or terminate the project... However, it is not always possible to draw a sharp line between formative and summative evaluation. (2010, pp. 92-93).

In line with the statement by Nieveen, the function of the evaluation in this research was firstly to improve the designed intervention by focusing on the soundness, feasibility and practicality thereof. Therefore, the evaluation with the experts and the first group of teachers was mainly formative. The evaluation with the second group of teachers was more summative, aiming to find some evidence for the effectiveness of the programme.

To evaluate the effectiveness of a teacher professional development programme, Guskey (2000) drew a distinction between 'proof' and 'evidence'. In evaluating teacher professional development, the ultimate goal is to prove that the intervention works by conducting an experimental, controlled intervention. However, in authentic educational situations, it is difficult, and perhaps even impossible, to meet the criteria for setting up such an experimental condition to collect proof (Guskey, 2000). According to Guskey, the real-world setting for interventions aiming at the professional development of teachers is too complex to provide proof. Instead of collecting data to prove that an intervention works, the evaluation should focus on the collection of data to gather 'good evidence' regarding whether the intervention works or not. To collect data for evidence of the extent to which an intervention works, Guskey suggested using a model with five levels. Table 7.1 provides an overview of these five levels of evaluation for interventions on teacher professional development. In this table, the five levels are related to the evaluation of the research in this thesis. An extra column was added to indicate where this thesis reported on this level.

As shown in Table 7.1, the evaluation of the intervention in this research focused initially on the improvement of the intervention. Expert appraisal and a pilot study were conducted to gather information about the soundness, feasibility and practicality of the programme and its components (Evaluation Level 1). In the pilot study, information pertaining to how students perceived the materials used by the teachers and how the teachers responded to students' reactions was also gathered (Evaluation Level 5). The evaluation in this phase was mainly formative.

Table 7.1 Evaluation levels by Guskey (2000) and the designed TPDP in this research.

Evaluation Level	What Questions Are Addressed?	How Will Information Be Gathered?	What is Measured or Assessed?	How Will Information Be Used?	Thesis
1. Participants' Reactions	How feasible and practical are the designed, materials, instruments and strategies? How feasible is the TPDP?	Question-naires Interviews Group interviews	The feasibility and practicality of the designed intervention and its components.	To improve the TPDP and its components.	Chapter 4 Chapter 5
2. Participants' Learning	To what extent could professional growth be identified?	Interviews Question-naires	A potential change in teachers' knowledge, skills and beliefs.	To collect evidence for the effectiveness of the intervention.	Chapter 6
3. Organisational Support and Change					
4. Participants' Use of New Knowledge and Skills	To what extent did the content of internal school-based examinations change?	Content analysis of school-based examinations	A change in teachers' practices.	To collect evidence for the effectiveness of the intervention.	Chapter 6
5. Students' Learning Outcomes	How did the students experience the test items, instruments and strategies?	Interviews with students Classroom observations	Feasibility of test items, instruments and strategies.	To improve test items, instruments and strategies.	Chapter 4

In the next phase, the evaluation with another group of teachers focused on the characteristics of the intervention and the effectiveness thereof (Evaluation Levels 2 and 4). Therefore, the evaluation was more summative in nature. However, although the information provided some evidence of teachers' professional growth, 'proof' that the TPDP works, in the sense proposed by Guskey, was not found. To prove the effectiveness of the

designed TPDP, research in an experimental or quasi-experimental setting with substantially more participants is required.

One of the evaluation levels for the effectiveness of a TPDP is the students' learning outcomes. Students' outcomes are sometimes even regarded as the ultimate aim of a TPDP, and an evaluation of the effectiveness of a programme should therefore take the students' progress into account (Maandag et al., 2017). In this research, students' outcomes were evaluated in terms of students' perceptions of the practicality of the instruments and strategies (Chapter 4) and the teachers' perceptions of the students' responses (Chapters 5 and 6). An evaluation of the students' learning progress in terms of measured student performances was not part of this research. Future research to identify the impact of the designed TPDP on students' performances is therefore needed in order to evaluate the effectiveness of this programme on students' learning.

One final note must be made about the absence of information at the third level of evaluation. The relationship of teacher professional development to the school organisation was not part of this research. As mentioned earlier, this research can be regarded as typical Phase I research, as defined by Borko (2004). This type of research usually leaves the context of the intervention unstudied, which was also the case in this research. The effect of organisational support, or the lack of support, on teacher professional development in assessment literacy is an issue that should be studied in more depth in the future. Some of the outcomes of this research, mainly the perceived lack of time for the teachers to work on their assessment skills, gives rise to the urgency of investigating this issue.

7.4.3 Reflection on the role of researcher

As in other educational design research, the researcher also was in the role of the designer of the intervention. The roles of the designer and researcher might be in conflict when collecting the data. This might lead to a methodological bias because the researcher might interpret the data less objectively and participants might respond in favour of the intervention because they know that the researcher designed it (McKenney & Reeves, 2012; Nieveen, 2010).

Several arguments, pro et contra, can be raised regarding the combined roles of designer and researcher. McKenny and Reeves (2012) described these arguments in terms of the advocate and the critic. For the advocate, the most important argument for combining the roles of designer and researcher might be to gain deeper insight into how the intervention works and why. Observations and interviews are helpful instruments to reveal how participants perceive the intervention in more depth. Because the researcher has designed the intervention, the researcher is better able to ask follow-up questions during interviews, for example. Observations regarding how the intervention works might help in the redesign of the intervention. For the researcher, first-hand, detailed understanding might be beneficial.

On the other hand, from the perspective of the critic, this combination might lead to a bias, as mentioned previously. Respondents might respond differently to the researcher and respond in ways that are more socially desirable or less critical. The researcher might be less receptive to criticism or may interpret responses differently because the researcher is attached to the design too closely.

To overcome these biases, triangulation of methods and data sources are important, if not essential. Evaluation of the intervention should be based upon data from multiple instruments. Another possibility is to use "unobtrusive data collection methods" (McKenney & Reeves, 2012, p. 150).

In this research, both strategies to decrease the methodological bias have been applied. As shown in Table 7.1, multiple instruments have been used at all stages of testing and evaluation to meet the requirements of triangulation. Interview schemes and observation schemes were used in combination with questionnaires or a content analysis. Furthermore, the participating teachers could complete the questionnaires anonymously, which can be regarded as an unobtrusive data collection method. It is important to bear in mind, however, that the number of participating teachers was limited, which might have affected the unobtrusiveness.

7.4.4 Limitations

As indicated previously, this research was conducted with relatively small numbers of participants in authentic situations. Consequently, the findings in this research may be limited and make these findings less generalisable to

other contexts. More research is required to determine the effects of the designed TPDP on teacher professional growth with more substantial groups that include larger numbers of participants. This research should also be extended to other types of secondary education and to other subjects in secondary education in order to develop the research from a Phase 1 study to a Phase 2 study.

The outcomes of this research may also be limited by the diverse composition of the participants, respondents or samples in the various stages of the research. In the first phase of the research, the content analysis was based on 49 internal school-based examinations that were sent in by geography teachers from 13 schools. The questionnaire was completed by 74 respondents, all of whom were teachers in pre-vocational geography education. However, the extent to which these respondents correspond with the teachers from the 13 schools that sent in the internal school-based examinations is unknown. This makes it difficult to compare the outcomes from both instruments at this stage of the research. Furthermore, the level of the prior knowledge, skills, beliefs and practices of participating teachers in the programme compared to those of other teachers in pre-vocational education is also unknown, as is the extent to which these groups of participants reflected the groups of respondents in the first stage of the research in terms of knowledge, skills, beliefs and practices.

In the third stage of this research, teachers participating in the TPDP responded positively to the invitation from the researcher. The teachers all participated voluntarily in the TPDP. However, it is important to bear the possible bias in this participation in mind because most of the participants were former students who were acquainted with the researcher. The potential effect of this bias is unknown. Teachers' motivations might also have been affected because the programme was approved as a formal activity for which the participating teachers could register their professional development in the new register for teachers.

Finally, the results need to be interpreted with caution because a substantial number of the results are based on qualitative data mainly derived from interviews. Although the interpretations of the data were approved by others, the findings may have been affected by the interpretations of the comments.

Further research with larger numbers of participants aiming to collect more evidence is therefore needed.

7.5 Implications and recommendations

The outcomes of this research inform future teacher professional development programmes and how teacher learning within these programmes could be fostered. To implement a future programme with regard to teachers' assessment literacy, the findings of this study have some implications related to four dimensions:

1) The knowledge base
2) Pedagogical content knowledge
3) Students' personal understanding
4) The emotional dimension.

Implications regarding these four issues will be discussed in the next sections. Having discussed these issues, in the section thereafter, a framework for future teacher professional development programmes with regard to teacher assessment literacy will be proposed.

7.5.1 The knowledge base with regard to assessment literacy

An implication of this research is that a TPDP concerning assessment literacy should stimulate teachers to extend their knowledge base regarding assessment and constructive alignment. In addition to extending knowledge, reflection by teachers upon their conceptions, students' learning outcomes and institutional requirements is needed in order to achieve higher mastery levels in assessment literacy. Concerning the knowledge base, several standards have been proposed. The standards proposed by Brookhart (2011) are probably the most comprehensive and up-to-date standards, which have taken the more recent insights into account. According to Brookhart (2011), these standards could serve as

> guiding teacher educators as they plan and implement teacher preparation programs, guiding teacher professional developers as they plan and implement in-service programs; guiding teacher self-assessment; and guiding educational measurement specialists in their conceptualization of

> student assessment for a range of research and development purposes. (pp. 10-11)

However, these standards, as is the case with other proposed standards, do not draw a distinction between levels of competency for pre-service teachers and in-service teachers, or for teachers at different stages of their careers. The outcomes of the research in this thesis, however, indicate that teachers' knowledge base or assessment competency after pre-service education, or even after years of teaching experience, is not fixed and is subject to improvement.

To extend teachers' knowledge base, it seems to be important to provide teachers with the opportunity to extend their acquired knowledge and skills via in-service education. This in-service education should focus on the alignment of educational assessment skills and knowledge with subject-specific content and goals. Aligning the goals with the assessment practices requires a deeper understanding on the part of the teachers regarding the relationship between the students' learning processes and their assessment practices. Instruments, such as the taxonomy table in the revised taxonomy of Bloom, could serve as guides for teachers in the construction of assessments that focus more on students' learning and are thus more in line with the educational goals.

Therefore, in contrast to the standards, it seems to be more realistic to draw a distinction between teachers' assessment competency as novices or experts. This distinction in levels of competency concerning assessment literacy standards provides opportunities to distinguish between teachers with basic competence regarding assessment literacy and teachers with higher or even excellent qualifications. Differentiating among teachers in terms of competency or excellence was already promoted in 2011 by the Education Council of the Netherlands. The Education Council of the Netherlands (2011) advised that schools should designate five per cent of their teachers as excellent teachers and role models for other teachers. These teachers should be scaffolded over time to work on their professional development and that of their colleagues with the aim of innovation and the improved quality of their education.

The outcomes of this research can therefore be used to develop targeted interventions aimed at designating more competent and even excellent

teachers with regard to assessment literacy. Teachers who have participated in a TPDP concerning assessment literacy and have accomplished a higher mastery level successfully could be designated as experts and role models for their colleagues. These teachers could belong to the five per cent of excellent teachers within schools and function as curriculum leaders within their departments, designing the curriculum by paying specific attention to the alignment of the educational goals, instruction and assessment. It is recommended that schools invest in teacher professional development in assessment literacy with the aim of designating excellent teachers within school departments who can function as role models for their colleagues.

7.5.2 Pedagogical content knowledge

Teachers encountered problems when assessing students' levels of performance. Evaluation with the teachers showed that insight into the subject-specific core concepts and geographical relationships seemed to be equally important to judge diverging responses from students. In other words, teachers needed to have an extended geographical conceptual framework and comprehension of what it means to reason geographically in order to judge students' geographical understanding or geographical reasoning.

Hence, teachers should not only possess educational assessment knowledge and skills, but also a good level of pedagogical content knowledge (PCK). A good level of PCK in geography education implies that teachers have an extended geographical conceptual framework and understand what it means to reason geographically. This not only helps teachers to design a curriculum by aligning goals with instruction, but also to use assessment information to interpret students' learning in the content area, particularly when students' responses diverge from the intended attainment targets. Attainment targets, as such, reflect the subject-specific objectives that include a knowledge dimension and a cognitive dimension. The knowledge dimension defines 'what' should be learned and the cognitive dimension defines 'how' this can be demonstrated. Criteria for judging and marking students' performances should be in line with these attainment targets; however, teachers should also be able to judge and score students' responses that diverge from these targets.

In recent decades, several attempts have been made to define a framework with essential geographical core knowledge and the relationship to

geographical understanding, geographical thinking or geographical reasoning (see also Chapter 4). One way to define geographical knowledge and the relationship to geographical thinking was the analogy with a language. Learning geography was compared to the learning of a language, with the geographical facts being the 'vocabulary' and the essential general concepts and theories the 'grammar' (Jackson, 2006; Lambert, 2011). According to Lambert (2011), students should be able to demonstrate their ability to use this 'geographical language' in order to make connections between places and scales. Therefore, in geography education, the focus must be on the grammar as well as on the vocabulary to stimulate 'thinking geographically'.

Other ways to describe 'thinking geographically' came from geography educators in the Netherlands, who referred to this process as geographic relational thinking or as geographical reasoning. Favier and Van der Schee (2014a, p. 156) defined geographic relational thinking as "a higher order kind of thinking about relations and effects in geographical systems" in which higher order kinds of thinking refer to cognitive processes such as interpreting, organising and evaluating. Two types of relationships are distinguished in this definition. Firstly, vertical relationships within regions are distinguished and typified as relationships among physical geographical features within a region, among human geographical features within a region or between physical geographical and human geographical features within a region. Secondly, horizontal relations are characterised as the interplay of changes among regions. Both types of relationships, vertical and horizontal, are part of geographical or spatial systems.

The importance for students to think about geographical relationships in and between regions has also been emphasised by others (Hooghuis et al., 2014). Hooghuis et al. (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". However, this definition of geographical reasoning, as was the case with previous ones, did not further specify levels of performance or attainment within geographical reasoning or understanding. This raises the question of what the distinct levels of geographical reasoning might be and how different levels of progression in geographical reasoning could be identified.

An attempt to relate geographical understanding to progression in understanding was made by Bennetts (2005b). According to Bennetts, “at the core of the concept of understanding is the notion of ‘making sense’ of something, or ‘giving meaning’ to something” (2005b, p. 113). This requires making connections involving prior experiences and knowledge, core ideas or concepts, and mental processes. Geographical understanding adds the subject-specific ideas, concepts, methods and perspectives to this more general notion of understanding. To define progression in geographical understanding, Bennetts identified eight dimensions of progression in geographical understanding, which included increasing breadth, complexity, abstraction, distance from experience and an association with cognitive skills and affective elements.

Taylor (2013) compared this framework to others, and concluded that all frameworks referred to a form of increasing breadth, increasing depth, a move from concrete to abstract and the use of a wider range of techniques. Although there seems to be commonality between the frameworks, no single framework fully satisfies the need to identify students’ progression in understanding (Weeden, 2013).

The need for a framework of progression in geography education and the need for large-scale, longitudinal research to develop such a framework were underpinned by Lane and Bourke (2017) in their review study about assessment and geography education. They emphasised that clarity about the nature of progression in the learning of geography is needed, as is insight into which assessment instruments will provide valid and reliable measures of this progress. According to Lane and Bourke (2017, p. 11), “there is a need for consensus regarding elements of geographical literacy and the development and validation of instruments for assessing such”.

One of the key issues in developing such a framework is to draw a distinction between the aggregation of knowledge, mainly reflected by increasing breadth, and progression in geographical thinking, referring to the use of higher order concepts and the use of geographical procedural knowledge. A comparable framework was proposed by the Geographical Association (2014) for geography education in England. Recently, a similar framework was proposed by the Royal Dutch Geographical Society in the Netherlands as part of a curricular reform in the Netherlands (KNAG, 2017). The framework

reflects the ideas of the geography community in the Netherlands regarding achievement in geography education. Achievement in geography education is defined in terms of progression and understanding at different stages. The framework could therefore become an important instrument to scaffold geography teachers in the Netherlands, not only to identify and interpret students' geographical understanding and students' levels of performance in terms of progression, but also to scaffold teachers when constructing and scoring test items that focus on higher order cognitive processes, such as evaluating and creating.

7.5.3 Students' personal understanding

As stated above, it is important that teachers are able to identify and interpret students' geographical understanding. Frameworks with descriptions of students' expected levels of performance or attainment targets can help teachers to grasp the level of understanding in students' learning. These levels and supporting criteria, however, have to be interpreted by teachers and students in the same way (Black et al., 2010, 2011). A shared understanding of these criteria between teachers and students is crucial for judging students' levels of performance.

However, understanding is often affected by different perceptions of expected levels of performance by teachers and students. Teachers base their perceptions of what is expected of students on formal requirements from external stimuli, such as a syllabus with subject-specific objectives. Teachers combine their interpretations of these formal requirements with their beliefs about learning and their subject-specific knowledge in their perceptions of understanding; in this way, teachers develop their own expectations regarding attainable levels of understanding. This interpretation by teachers leads to what Entwistle and Smith (2002) designated as 'target understanding'.

Students, on the other hand, develop their own kind of understanding, which is a 'personal understanding'. Personal understanding, according to Entwistle and Smith (2002), is affected by students' interpretations of the targets set by teachers through their own experiences, including their existing knowledge, their motivation, their study approach and their expectations and beliefs about educational learning. For teachers, the essential problem is "becoming one of how to meet the content requirements placed upon us by syllabus documents and policy…while finding ways to help pupils move into

explanatory forms of understanding…" (Smith, 2002, p. 170). To overcome a potential discrepancy in teachers' and students' perceptions of understanding, it is important to develop ways to stimulate a mutual awareness between teachers and students concerning formal requirements and ways of learning.

This underlines the importance of aligning summative assessment practices with formative assessment practices. An important aspect of formative assessment is the provision of feedback on the task, the learning process and the students' self-regulation (Hattie & Timperley, 2007). Teachers should therefore practice with students in classroom settings and give the students feedback regarding their responses to test items, particularly those focusing on more demanding cognitive processes, such as evaluating and creating, before applying these test items in summative assessments.

7.5.4 The emotional dimension

An important aspect of professional growth in assessment literacy is a change in conceptions of assessment. As stated previously, altering teachers' conceptions is not simply a question of introducing new knowledge and skills, but is the result of reflection on educational goals, professional experimentation and students' learning outcomes. Although this reflection is guided by questioning teachers' knowledge, skills, beliefs and practices, part of the conceptions is formed by an affective, emotional dimension pertaining to the nature and purposes of assessment. According to Xu and Brown (2016, p. 156),

> …teachers tend to adopt new knowledge, ideas, and strategies of assessment that are congruent with their conceptions of assessment, while rejecting those that are not… The emotional dimension of conceptions may make conceptual change difficult, leading to less effective learning about assessment and reduced effectiveness in implementing new assessment policies… To improve teacher AL (assessment literacy) inevitably involves a long process of attending to, and possibly changing, teachers' existing conceptions of assessment.

The emotional dimension of teachers' conceptions is influenced not only by teachers' prior experiences, but also by perceived opportunities or constraints from external factors. One of the factors that emerged in this research was the impact of external high-stakes tests, such as the external end-of-school

exit examination. These tests particularly influence teachers' conceptions when the results of these tests are used for purposes of accountability. This might lead to a loss in teachers' confidence when constructing and scoring tests (Weeden, 2013).

Another factor that appeared to influence the emotional dimension of teachers' conceptions was their perceived lack of time to construct test items. These outcomes are in line with the results of the investigation by the Inspectorate of Education in the Netherlands in 2013, which also mentioned that teachers experience their workload as an important constraint (Chapter 1). It seems to be important that teachers who are willing to enhance their assessment literacy will be scaffolded in time. In particular, those teachers who want to become excellent at Mastery Level 3 and function as role models for their colleagues within school departments should be scaffolded by the school administrators, as argued previously.

7.5.5 A conceptual framework for future teacher professional development programmes with regard to teacher assessment literacy

To accomplish higher mastery levels in assessment literacy, and taking the outcomes of this research, the design principles and implications as mentioned above into account, the following framework for a TPDP concerning assessment literacy is proposed. The framework consists of three consecutive iterative stages. In each stage, teachers attend three meetings. In these meetings, the teachers are supported via instruction materials and are scaffolded to construct their assessments by demonstration, collaborative practice and (peer) feedback. Between the meetings, the teachers experiment with their assessment practices in their classrooms.

Stage 1. Focus on reflection and enactment between the external domain and the personal domain

At this stage of the programme (Figure 7.1), the focus is on strengthening teachers' knowledge base at Mastery Level 1 and questioning their conceptions regarding the purpose of assessment in relation to educational goals. The knowledge base includes knowledge about the subject-specific content and concepts and the relationship to diverse cognitive processes, in addition to educational assessment knowledge and skills. Reflection in this phase should focus mainly on the educational goals and how to align these

goals with instruction and assessment, both formative and summative. Although the focus at this stage is on these elements, it is also recommended to stimulate teachers to start experimenting with the newly constructed test items, to discuss these test items with peers in the programme or with their colleagues in their departments, to give and receive feedback on these test items and to reflect on students' responses to these test items in order to create the opportunity for teachers to develop their own professional growth pathways.

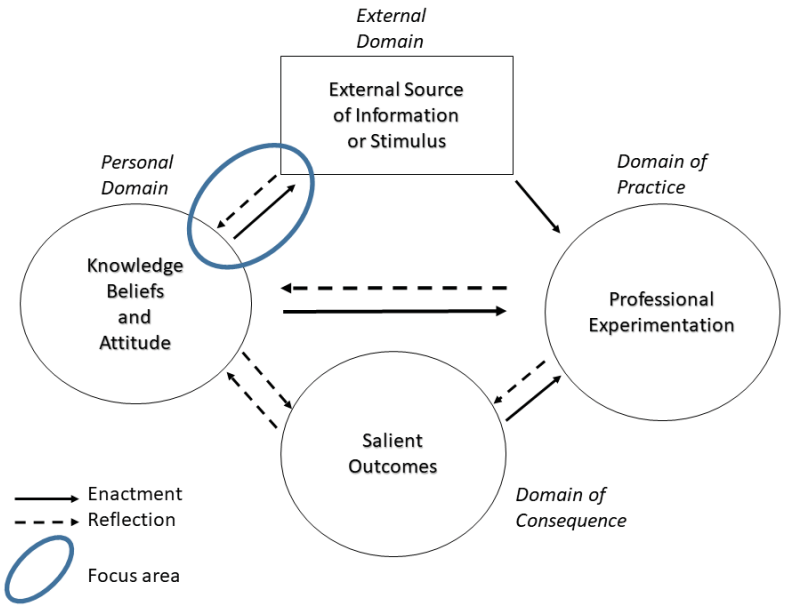


Figure 7.1 First stage in the framework for teacher professional growth in assessment literacy (model adapted from Clarke & Hollingsworth, 2002).

Stage 2. Focus on the enactment of extended knowledge and skills, professional experimentation and constructive alignment

The second stage (Figure 7.2) should focus on achieving Mastery Level 2, "an internalized set of understanding and skills of the interconnectedness of assessment, teaching and learning" (Xu & Brown, 2016, p. 159). At this stage of the programme, the focus is therefore on extending teachers' professional experimentation, enactment of new knowledge and skills, and reflection upon

these. To achieve Mastery Level 2, an emphasis on the constructive alignment of the goals of education, instruction and assessment is extremely important. Constructive alignment can be stimulated and scaffolded by using instruments, such as a model to construct and score test items, and a taxonomy table. Constructive alignment can also be stimulated by reflection upon teachers' assessment practices and the extent to which these are in line with the educational goals. In addition to the first stage, the reflection focuses more on the teachers' own assessment practices and the relationship with their educational goals, whereas the reflection in the first stage focuses more on the alignment of educational goals and assessments in general.

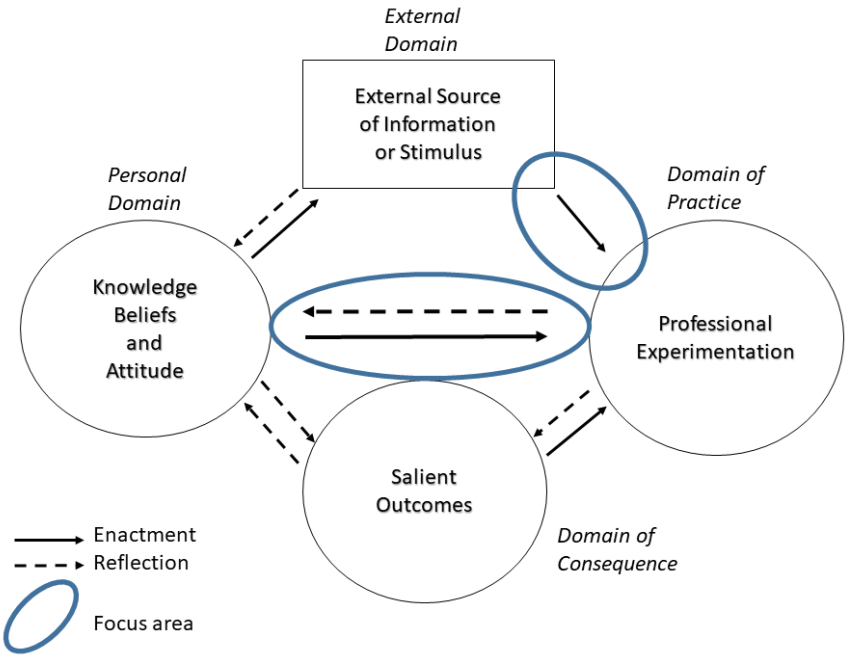


Figure 7.2 Second stage in the framework for teacher professional growth in assessment literacy (model adapted from Clarke & Hollingsworth, 2002).

Another important element at this stage is stimulating teachers' reflection on their conceptions. As argued previously, to achieve a higher mastery level in assessment literacy, teachers must be able to question and, if necessary, alter their conceptions. Again, it is recommended that the teachers be offered the opportunity to create their own growth pathways.

Stage 3. Focus on students' learning and their levels of performance

The third stage (Figure 7.3) should focus on teachers' reconstructions of their assessor identities and the self-directed awareness of assessment processes, thus achieving Mastery Level 3 by integrating the goals and practices with assessment policies and students' learning. The most important question at this stage is how students' learning can be stimulated by the teachers' own assessment practices. Mutual understanding between teachers and students regarding expected levels of performance can be prompted by the application of instruments, such as a flow chart with steps to construct, answer and score test items. At this stage, these instruments should scaffold students to become aware of the expected levels of performance, but could also be helpful in fostering students' peer and self-assessments. When students' peer and self-assessments are stimulated, teachers' summative assessment practices will be brought more in line with formative purposes. This is supposed to stimulate students' learning in a meaningful way.

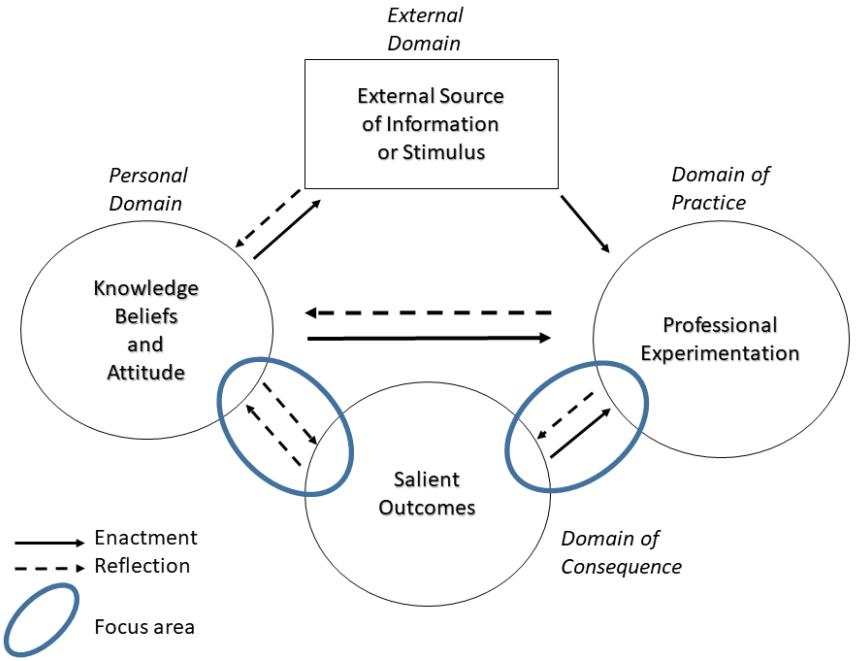


Figure 7.3 Third stage in the framework for teacher professional growth in assessment literacy (model adapted from Clarke & Hollingsworth, 2002).

Secondly, at this stage, teachers shape their conceptions of what assessment should entail, formed by policy and personal conceptions, constraints and their daily practices, through reflection. Reflection on what is expected from the teachers by external factors and how they want to contribute to students' learning by their assessment practices helps to reconstruct their assessor identity, and their achievement of Mastery Level 3.

Further research to evaluate this framework is recommended. A longitudinal study of teacher professional development concerning assessment literacy in three stages is needed to investigate the effect of teachers' professional growth on students' learning outcomes. This research should identify the extent to which students' performances and learning are enhanced by changing the assessment practices of teachers who attended the TPDP, and how teachers can be scaffolded to align students' performances and personal understanding with the teachers' target understanding.

7.6 Geography education, curriculum and assessment

Thus far, it has been argued that teacher professional growth with regard to assessment literacy can be evoked by a TPDP that takes the fact that teachers learn along different growth pathways to accomplish lasting changes in their knowledge, skills, beliefs and practices into account. Institutionalised and contextual influences have been left mostly unattended (see also Section 7.4.2), although some comments regarding the impact of the national exit examinations have been made. This final section will discuss the impact of contextual and institutionalised factors on geography teachers' assessment practices, and what is necessary in order to contribute to meaningful ways of teaching, learning and assessing geography.

7.6.1 The relationship to national exit examinations

Some comments have to be made concerning the relationship between the external examination (CE) and the internal examinations (SEs). As stated previously, respondents in this research reported on the impact they felt that the CE had on how they constructed their SEs. This might indicate that, for these teachers, the enacted curriculum and their SEs are based more on the expected content and format of the high-stakes CE than they are on the

content and objectives as presented in the syllabus of the examination programme.

The impact of the CE on the enacted curriculum and how teachers transfer this to their internal assessment practices can cause problems with regard to validity. The CE, which can be considered to be the most obvious example of a high-stakes test, can have, as Kuiper (2017) described, a 'pre-shadowing' effect on education, or, as Kuiper et al. (2017, p.86) put it, "what is tested makes beloved and what stays untested makes unbeloved". This affects how schools and teachers enact the curriculum in their classrooms. The enacted curriculum might reflect a selection of the content and objectives of the curriculum based on the expected content of the CE, rather than reflecting the entire subject-specific content and objectives of the examination programme as prescribed in the syllabus. In the worst case, this leads to the effect that has been reported in numerous circumstances, which can be described as 'teaching to the test'.

Consequently, the enacted curriculum might deviate from the intended curriculum when teachers select the objectives and content that guarantee the best chance for good results on the CE in the teachers' perceptions. 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.

In the case of the examination programme in pre-vocational geography education in the Netherlands, two main problems with regard to this issue can be distinguished. The first is the distinction between the content and objectives of the examination programme allocated to SEs on one hand and to the CE on the other. Both parts of the examination programme contain three areas of geography (see Chapter 2 and Appendix A for an explanation). The three areas belonging to the examination programme for SEs are supposed to

be assessed in these internal examinations. These examinations count for 50% for the overall result.

However, results from the two questionnaires in 2013 and 2015 that monitored the implementation of the new examination programme in pre-vocational geography education showed that three-quarters of the teachers in pre-vocational geography education (N=105 and N=106) assessed the CE examination programme in the SEs (Noordink, Oorschot, & Folmer, 2017). These results match the outcomes from the questionnaire that was used in this research. Results from this questionnaire pointed out that the percentage of the respondents (N=74) who assessed the CE examination programme in the SEs was even higher, at almost 90 per cent.

This can be explained by the fact that most geography teachers in pre-vocational geography education in the Netherlands teach three areas of geography for the SEs - Sources of Energy, Poverty and Wealth, and Boundaries and Identity - in the third grade. The three areas for the CE - Weather and Climate, Water and Population and Place - are taught in the fourth and final grade. Outcomes from the questionnaires provided by Noordink et al. (2017) confirmed this distinction. Teachers responded that the three specific areas of geography for the SE were taught in the third grade by 87% to 97% of the teachers. The three areas of geography in the fourth grade showed the same percentages.

Consequently, the content of the SEs in the fourth grade mainly reflect the examination programme of the CE. Furthermore, the outcomes of the questionnaires by Noordink et al. (2017) showed that the responding teachers estimated that the assessment of the CE areas of geography accounted for 54% (2013) to 64% (2015) of the SE results. Therefore, it can be argued that SE examination results, which count for 50% of the overall examination result, are dominated by CE content.

The second problem arising from the impact of the CE on the SEs is related to the type of test items in the CE. As reported in Chapter 3, almost three-quarters of the teachers in pre-vocational geography education believe that the test items in their SEs should reflect the formats used in the CE as far as possible. This can be considered problematic because the objectives of the SE deviate from the objectives in the CE to some extent, particularly in terms of achievement standards. In the examination programme for the SEs, specific

attention is paid to elementary (field) enquiries as part of the achievement standards. These elementary (field) enquiries seem to be less apparent in SEs than might be expected (Noordink et al., 2017).

Another reason that this can be considered to be problematic is because the CE mainly contains test items in formats that produce reliable test results (Harlen, 2005). Although there is nothing wrong in striving for the most reliable test results, the focus on reliability might be at the expense of attention to validity. Because the impact of the CE on how teachers construct their SEs appeared to be extremely strong, it is interesting to analyse recent CEs with regard to the issue of validity. Some comments on validity can be made by analysing the extent to which the test items in the CEs reflect geographical knowledge and cognitive processes.

An analysis of test items in the CEs from 2015, 2016 and 2017 with regard to the geographical knowledge and cognitive dimension scored exactly in the same way as in the test items of SEs as reported in Chapter 6, showing that the majority of test items focused on remembering and, to a lesser extent, on understanding, mainly with reference to conceptual knowledge (Table 7.2). Most of these test items were either in the format of multiple-choice questions or short, constructed response tasks. These types of test items are often considered to be items that can be marked readily and reliably. Test items demanding longer answers from students, focusing on higher order cognitive processes such as evaluating or creating, were missing. Therefore, the focus seemed to be more on the reliability of the examination results at the expense of the validity thereof.

The content analysis of SEs at the beginning of this research and at the end showed the same tendency with regard to the cognitive dimension and the formats for test items. This seems to confirm teachers' responses regarding the impact of the CE on their SEs. Compared to the content and objectives of the SE programme, this points to a lack of alignment. The achievement standards in the syllabus for the SEs also demand that students show higher order cognitive processes. Therefore, in terms of construct validity, the prevailing formats for test items in SEs and the CE seem to contribute to less valid examination results.

Table 7.2 Cumulative percentages of CE-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

It was argued previously in Chapter 3 that another approach to overcome these constraints between reliability and validity was necessary. More attention to the validity of CE and SEs is needed. This problem is not limited to geography education. Kuiper et al. (2017) underpinned the importance of a rebalance of reliability, validity and transparency in education in the Netherlands. A dependability approach, as suggested by Harlen (2005), might contribute to this solution (see also Chapter 3).

A dependability approach to the construction of CE and SEs in geography education in the Netherlands is not the only potential, or advocated, solution to contribute to more meaningful exams. A rethinking of the examination programme and the relationship between the content and purpose of the CE and SEs is also necessary. The distinction between a SE examination programme and a CE 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 programme must be kept to its full extent. Most teachers in pre-vocational geography education who responded to one of the questionnaires monitoring the implementation of the new examination programme in pre-vocational geography education responded that the examination programme

was “overloaded” (Noordink et al., 2017, p. 13). The perceived overloaded programme appears to have been induced by the range of subjects and regions within the examination programme.

It can be questioned whether this range of subjects and regions within the current examination programme ensures the enhancement of students’ progressions in geographical understanding in the best possible way. 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 (see also Section 7.5.2). Increasing breadth coincides with the aggregation of knowledge to a certain degree, whereas progression also focuses on the study of the distinct subjects in the examination programme in more depth. A strong focus on increasing breadth might be at the expense of increasing depth. In terms of the geography curriculum, more subjects and regions in the examination programme might induce a strong focus on learning concepts at more distinctive scales in more geographical contexts. Furthermore, as known from this research and others, it is not uncommon for this to be accompanied by the tendency to assess these concepts with a focus on the recall of knowledge.

To achieve more meaningful learning of the content in depth, a less overloaded examination programme might be required. Consequently, choices within the range of subjects and regions seem to be inevitable. This underlines the urgency for a rethinking of the examination programme, with a clear purpose for the SEs, to ensure that the examinations -both SEs and CE- contribute to meaningful ways of learning and assessing geography.

7.6.2 The impact of text books and taxonomies

Some of the external factors influencing teachers’ assessment practices are the conditions set by the text books. Teachers rely on the text books in constructing their SEs to a large extent (Chapter 3). The teachers who responded to the questionnaire indicated that the vast majority of them use test items accompanied by the text books. In the same questionnaire, teachers responded that more than half of them used the RTTI taxonomy when constructing their SEs (see Chapter 3). This taxonomy consists of four categories: remembering (R), executing a familiar task (T1), implementing an unfamiliar task (T2) and comprehension or understanding (I).

In numerous cases, the tests that are included in the text books are accompanied by a classification of test items based on this RTTI taxonomy. Because teachers seem to rely on these tests and classification of test items to a significant extent, this might also influence their assessment literacy. One of the standards for teachers' assessment literacy, standard IX according to Brookhart (2011, p.7), is that "teachers should be able to articulate their interpretations of assessment results and their reasoning about the educational decisions based on assessment results to the educational populations they serve (student and his/her family, class, school, community)".

When teachers use the RTTI classification of test items to interpret the assessment results and communicate these interpretations to students, parents or others within their schools, this classification must be undisputed. Unfortunately, this does not seem to be the case. In interviews, panel discussions or meetings with teachers and experts in this research, it became clear that this taxonomy was interpreted in multiple and diverse ways. For example, what was considered to be a test item focusing on I in the taxonomy by one teacher or expert was seen by others as a typical T2 or even T1 task. Given the importance of the interpretation of test results by teachers, this clearly requires a sound definition of the distinctive categories within this taxonomy. Without this clear definition, less emphasis must be placed on this taxonomy, which seems to be misleading when interpreting students' learning and their progress in learning.

7.6.3 Teacher assessment literacy and pre-service teacher education

The impact of pre-service teacher education on teachers' assessment literacy is a third important issue that needs to be addressed at this point. In teacher education for pre-vocational education, the focus in terms of assessment literacy is on the knowledge and skills that contribute to the teachers' assessment knowledge base. This knowledge base has several components, which strongly resemble the elementary knowledge and skills to construct, score and administer assessments in order to use students' results to make sound decisions.

In pre-service teacher education, apart from this elementary knowledge and skills, how students perceive the quality of modelling by the teacher educators is important. In a study by Levy-Vered and Nasser-Abu Alhija (2015), beginning

teachers who reported a higher quality of modelling on assessments by teacher educators showed a higher level of assessment literacy. In this regard, it cannot be left unnoticed that, since 2014, students who were trained to become teachers in pre-vocational geography education in the Netherlands have had to pass a test on basic subject-specific knowledge before graduating. The tests are supposed to measure students' elementary subject-specific knowledge. However, the construction of test items is based on a list of geographical concepts of lower and higher order. The test items are multiple choice items and, in the areas of geography that belong to the knowledge base, 125 overall. Although it cannot be argued that the test items focus exclusively on remembering geographical concepts, the idea that the test contains 125 multiple choice items pertaining to geographical concepts encourages the students to learn these concepts by heart just before the test is taken. The effect of these tests on students' beliefs regarding assessments with respect to modelling is unknown. However, it seems inevitable that this method of testing will affect students' conceptions of assessments. It also seems inevitable that this will affect their own assessment practices as beginning teachers.

7.6.4 The relationship between formative and summative assessment

A considerable amount of literature has developed around the theme of formative assessment (Gulikers & Baartman, 2017; Sluijsmans et al., 2013). Formative assessment, or Assessment for Learning, encompasses strategies to provide students with feedback concerning their learning progression (see Chapter 1). Formative assessment, in this sense, is supposed to contribute not only to students' learning, but also to their self-regulation. Although in studies by Gulikers and Baartman (2017) and by Sluijsmans et al. (2013) it was mentioned that, thus far, there is not much empirical evidence for the effects of formative assessment on students' learning, the potential for formative assessment practices to stimulate students' learning seems undisputed.

In their review, Gulikers and Baartman (2017) emphasised the importance of the role that teachers play in formative classroom assessment practices. They also stressed the teachers' competencies that are necessary for the successful enactment of formative classroom assessment practices and that which is needed for the professional development of teachers with regard to these practices. According to Gulikers and Baartman, teachers' competencies

include the knowledge of formative assessment practices, as well as sufficient pedagogical content knowledge and the competence to relate this knowledge to the understanding of students' learning progression. The latter is seen as one of the missing links in teacher professional development programmes. Teachers seem to have difficulty converting students' performances into new strategies, and the professional development programmes do not pay sufficient attention to all aspects of formative assessment, including this final step. These important themes of alignment among the distinctive phases in formative assessment and the need for teacher professional development are also mentioned in the review by Sluijsmans et al. (2013).

In contrast to the promotion of formative assessment, when discussing summative assessment, the limiting effects are usually emphasised. Although Sluijsmans et al. (2013) acknowledged the potential of summative assessments when these are used for formative purposes to contribute to students' learning, most studies have stressed the negative effects on students' motivation or learning (Harlen & Deakin Crick, 2002). Therefore, in their curriculum reflections, Kuiper et al. (2017) advocated a shift in education from an emphasis on summative assessment towards more formative assessment.

Notwithstanding the importance of focusing on formative evaluation practices, the relationship between the formative assessment practices and the content of summative assessments often remains unnoticed. When summative assessments are not brought into line with curricular goals and formative evaluations, the effect of these formative practices might be limited. In their study on the relationship between students' personal understanding and teachers' target understanding, Entwistle and Smith (2002) stressed the effect that teachers' choices in their pedagogy and assessments might have on students' perceptions of what is expected from them in terms of performance. They mentioned (2002, p. 330) "how differing conceptions (of teaching) affect the choice of both teaching methods and assessment procedures, and hence learning outcomes. Those choices then influence the approaches to studying adopted by students, and hence the levels of understanding they reach".

In other words, teachers' pedagogy and assessment procedures affect students' approaches to studying. This underlines that the pedagogy,

formative assessment practices and summative assessment practices should be brought in line with each other. When teachers' summative assessment practices differ from their pedagogy and formative assessment practices, this might influence students' conceptions and approaches to studying. For example, when the summative assessments focus on remembering geographical facts and concepts, students might adopt an approach to studying that is focused on the recall of knowledge rather than to reason geographically. This might become their preferred approach, regardless of the pedagogy and strategies during their classroom practice. Therefore, summative assessments should reflect the broader aims of geography education in terms of knowledge and cognition. When summative assessments focus more on cognitive processes that stimulate students' geographical reasoning, these summative assessments might function as a lever for instruction, pedagogy and formative evaluation in line with the aims and objectives of geography education. Subsequently, summative assessment will be brought more in line with formative assessment.

7.6.5 Assessment, the curriculum and the aims of geography education

The relationship between the curriculum and assessment is complex. The curriculum is often perceived as a collection of objectives regarding the content and the cognitive processes students are supposed to demonstrate. However, this perception might be limiting in terms of achieving the intended educational goals.

In this respect, Kuiper (2017) drew a distinction between *goals to strive for* and *achievement standards*. The former reflects the broader educational goals. Because these broader goals are often formulated in generic terms, schools and teachers have, within a certain range, the possibility of making choices with regard to the depth of subjects and objectives. The latter refers to the attainment targets students are supposed to demonstrate and, as such, are fundamental for the exam programme and the exams, both external and internal.

Ideally, the goals to strive for, achievement standards and exams are in line with each other. This does not mean that the exams reflect the content and objectives of the entire curriculum fully; rather, that 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

thesis, the educational goals for geography education. The exams are supposed to follow the content and objectives of the curriculum, and not vice versa (Kuiper, 2017).

This sequence becomes problematic when teachers do not teach the entire curriculum and when they 'teach to the test'. In this case, the exams no longer follow the curriculum, but regulate the enacted curriculum instead. This might widen the gap between the intended and enacted curricula.

To bridge the gap between the intended and enacted curricula, more emphasis on constructive alignment seems to be necessary. Constructive alignment focuses on the alignment of educational goals, instruction, pedagogy, assessment and achievement standards. These five aspects should be in line with each other. To bring the educational goals within geography education in line with instruction, pedagogy, assessment and achievement standards, Lambert (2011) emphasised the importance of powerful knowledge and a 'capabilities' approach. Powerful knowledge encompasses the meaningful knowledge that takes the students beyond their everyday experiences. Therefore, powerful knowledge is counterintuitive and should be taught (Mitchell & Lambert, 2015; Stoltman, Lidstone & Kidman, 2015). According to Lambert (2011), three domains are essential in powerful knowledge:

1. Deep descriptive and explanatory world knowledge
2. The development of relational thinking in geography
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. The capabilities approach invites teachers and curriculum leaders to reflect on how education contributes to human autonomy and potential (Geocapabilities, 2016). A capabilities approach allows teachers to connect the subject specific knowledge to the *goals to strive for* (Lambert, 2011). As stated by Lambert (2011, p. 258):

> 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 geography education, this approach is helpful in order to determine what the objectives should be, how these can be taught and which pedagogies are useful. This approach could, or perhaps should, also be helpful in order to align the ultimate goals in geography education with the geography exams, both external and internal. To which extent do the tasks in these exams contribute to young people's capabilities? For each exam or assessment task, this should be the ultimate question to be answered.

References

Airasian, P. W., & Miranda, H. (2002). The role of assessment in the revised taxonomy. *Theory into Practice, 41*(4), 249-254.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for teacher competence in educational assessment of students.* Washington, DC: National Council on Measurement in Education.

Anderson, L. W. (2002). Curricular alignment: A re-examination. *Theory into Practice, 41*(4), 255-260.

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.

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is - or might be - the role of curriculum materials in teacher learning and instructional reform? *Educational Researcher, 25*(9), 6-8,14.

Bandura, A. (1989). Regulation of cognitive processes through perceived self-efficacy. *Developmental Psychology*, *25*(5), 729-735.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology, 52*(1), 1-26.

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

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25.

Bennetts, T. (2005a). The links between understanding, progression and assessment in the secondary geography curriculum. *Geography, 90*(2), 152-171.

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

Biggs, J. B. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347-364.

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

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., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*(2), 215-232.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice, 18*(4), 451-469.

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. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education), 21*(1), 5-31.

Black, P., & Wiliam, D. (2012). Assessment for learning in the classroom. In J. Gardner (Ed.), *Assessment and learning* (pp. 11-32). London: SAGE.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*(8), 3-15.

Bourke, T., & Lane, R. (2017). The inclusion of geography in TIMSS: Can consensus be reached? *International Research in Geographical and Environmental Education, 26*(2), 166-176.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.) (1999). *How people learn: Brain, mind, experience, and school.* Washington DC: National Academy Press.

Brookhart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education: Principles, Policy & Practice, 8*(2), 153-169.

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom.* Alexandria, VA: Association for Supervision and Curriculum Development.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3-12.

Brookhart, S. M. (2014). *How to design questions and tasks to assess student thinking.* Alexandria, VA: Association for Supervision and Curriculum Development.

Brooks, C. (2008). Geographical knowledge and teaching geography. *International Research in Geographical and Environmental Education, 15*(4), 353-369.

Brooks, C. (2013). How do we understand conceptual development in school geography? In D. Lambert & M. Jones (Eds.), *Debates in geography education* (pp. 75-88). London: Routledge.

Butt, G., Weeden, P., Chubb, S., & Srokosz, A. (2006). The state of geography education in English secondary schools: An insight into practice and performance in assessment. *International Research in Geographical and Environmental Education, 15*(2), 134-148.

Cito. (2008). *Het schoolexamen in het voortgezet onderwijs. Verslag van een onderzoek naar de kwaliteit van het schoolexamen bij de vakken Engels, Nederlands, biologie en wiskunde* [School-based internal examinations in secondary education. Report of a research regarding the quality of school-based internal examinations for the subjects English, Dutch language, biology and mathematics]. Arnhem: Cito.

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. E. (1996). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment, 3*(2), 159-179.

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education, 18*(8), 947-967.

Cleary, T. J., & Zimmerman, B. J. (2004). Self-regulation empowerment program: A school-based program to enhance self-regulated and self-motivated cycles of student learning. *Psychology in the Schools, 41*(5), 537-550.

Davies, P. (2002). Levels of attainment in geography. *Assessment in Education: Principles, Policy & Practice, 9*(2), 185-204.

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability, 28*(3), 251-272.

Dirkx, K. J., Kester, L., & Kirschner, P. A. (2014). The testing effect for learning principles and procedures from texts. *The Journal of Educational Research, 107*(5), 357-364.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation, 14*(7), 1-11.

Ediger, M. (2001). Assessment, geography, and the student. *Journal of Instructional Psychology, 28*(3), 150.

Education Council of the Netherlands. (2011). *Excellente leraren als inspirerend voorbeeld* [Excellent teachers as inspiring models]. The Hague: Education Council of the Netherlands.

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.

Examenblad.nl. (2015). *De officiële website voor examens in het voortgezet onderwijs* [Official website for exams in secondary education]. Retrieved July 28, 2015, from https://www.examenblad.nl/examen/aardrijkskunde-gl-en-tl-vmbo-2/2015/vmbo-tl?topparent=vga6k854m5p9

Favier, T. T., & Van der Schee, J. A. (2014a). Evaluating progression in students' relational thinking while working on tasks with geospatial technologies. *Review of International Geographical Education Online, 4*(2), 155-181.

Favier, T. T., & Van der Schee, J. A. (2014b). The effects of geography lessons with geospatial technologies on the development of high school students' relational thinking. *Computers & Education, 76*(0), 225-236.

Firth, R. (2013). What constitutes knowledge in geography? In D. Lambert & M. Jones (Eds.), *Debates in geography education* (pp. 59-74). London: Routledge.

Fullan, M., & Langworthy, M. (2014). *A rich seam: How new pedagogies find deep learning*. London: Pearson.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915-945.

Geocapabilities. (2016). Retrieved October 19th, 2017 from http://www.geocapabilities.org/training-materials/module-1-the-capabilities-approach/aims/

Geographical Association. (2014). *An assessment and progression framework for geography*. Retrieved November 22nd, 2017 from http://www.geography.org.uk/news/2014nationalcurriculum/assessment/

Gulikers, J. T. M., & Baartman, L. T. J. (2017). *Doelgericht professionaliseren: formatieve toetspraktijken met effect! Wat DOET de docent in de klas?* [Targeted professional development: Effective formative classroom practices! What are teachers' classroom practices?] Retrieved October 19th, 2017 from https://www.nro.nl/kb/405-15-722-doelgericht-professionaliseren-formatieve-toetscompetenties-met-effect/

Guskey, T. R. (1986). Staff development and the process of teacher change. *Educational Researcher, 15*(5), 5-12.

Guskey, T. R. (2000). *Evaluating professional development*. Thousand Oaks, CA: Corwin Press.

Guskey, T. R. (2002). Professional development and teacher change. *Teachers and Teaching: Theory and Practice, 8*(3), 381-391.

Harland, J., & Kinder, K. (1997). Teachers' continuing professional development: Framing a model of outcomes. *British Journal of In-service Education, 23*(1), 71-84.

Harlen, W. (2004a). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Harlen, W. (2004b). 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*(2), 207-223.

Harlen, W., & Deakin Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review, version 1.1). In: *Research Evidence in Education Library*. Issue 1. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Harlen, W., & Deakin Crick, R. (2003). A systematic review of the impact on students and teachers of the use of ICT for assessment of creative and critical thinking skills. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education, 4*(3), 365-379.

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. New York: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Haubrich, H. (1992). *International charter on geographical education.* Nürnberg: Commission on Geographical Education of the International Geographical Union/HGD.

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.

Huizinga, T., Handelzalts, A., Nieveen, N., & Voogt, J. M. (2014). Teacher involvement in curriculum design: Need for support to enhance teachers' design expertise. *Journal of Curriculum Studies, 46*(1), 33-57.

Inspectie van het Onderwijs. (2013). *Professionalisering als gerichte opgave. Verkennend onderzoek naar het leren van leraren* [Professional development as intended mission. An investigative research on teacher learning]. Utrecht: Ministerie van Onderwijs, Cultuur en Wetenschap.

Jackson, P. (2006). Thinking geographically. *Geography, 91*, 199-204.

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

Jo, I., & Bednarz, S. W. (2014). Dispositions toward teaching spatial thinking through geography: Conceptualization and an exemplar assessment. *Journal of Geography, 113*(5), 198-207.

Joyce, B. R., & Showers, B. (2002). *Student achievement through staff development*. Alexandria, VA: Association for Supervision and Curriculum Development.

Karkdijk, J., Van der Schee, J., & Admiraal, W. (2013). Effects of teaching with mysteries on students' geographical thinking skills. *International Research in Geographical and Environmental Education, 22*(3), 183-190.

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.

KNAG. (2017). *Visie op het aardrijkskundeonderwijs* [Vision on geography education]. Retrieved November 22nd, 2017 from https://geografie.nl/artikel/conceptversie-visiedocument-aardrijkskundeonderwijs-denk-mee

Koloi-Keaikitse, S. (2016). Assessment training: A precondition for teachers' competencies and use of classroom assessment practices. *International Journal of Training and Development, 20*(2), 107-123.

Korthagen, F. A. J. (2004). In search of the essence of a good teacher: Towards a more holistic approach in teacher education. *Teaching and Teacher Education, 20*(1), 77-97.

Korthagen, F. A. J., & Vasalos, A. (2005). Levels in reflection: Core reflection as a means to enhance professional growth. *Teachers and Teaching: Theory and Practice, 11*(1), 47-71.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212-218.

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*(2), 243-264.

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

Leat, D. (1998). *Thinking through geography*. Cambridge, England: Chris Kingston Pub.

Leat, D., & McGrane, J. (2000). Diagnostic and formative assessment of students' thinking. *Teaching Geography, 25*, 4-7.

Leat, D., & Nichols, A. (2000). Brains on the table: Diagnostic and formative assessment through observation. *Assessment in Education: Principles, Policy & Practice, 7*(1), 103-121.

Leat, D., Van der Schee, J., & Vankan, L. (2005). New strategies for learning geography: A tool for teachers' professional development in England and the Netherlands. *European Journal of Teacher Education, 28*(3), 327-342.

Lee, P., & Shemilt, D. (2003). A scaffold, not a cage: Progression and progression models in history. *Teaching History* (113), 13-23.

Levy-Vered, A., & Nasser-Abu Alhija, F. (2015). Modelling beginning teachers' assessment literacy: The contribution of training, self-efficacy, and conceptions of assessment. *Educational Research and Evaluation, 21*(5-6), 378-406.

Maandag, D., Helms-Lorenz, M., Lugthart, E., Verkade, A., & Van Veen, K. (2017). *Features of effective professional development interventions in different stages of teacher's careers: NRO report*. Groningen: Teacher education department of the University of Groningen.

Mayer, R. E. (2002). Rote versus meaningful learning. *Theory into Practice, 41*(4), 226.

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

Mitchell, D., & Lambert, D. (2015). Subject knowledge and teacher preparation in English secondary schools: The case of geography. *Teacher Development, 19*(3), 365-380.

Ministry of Education, Culture and Science. (2013). *Key figures 2008-2012.* Retrieved July 28th, 2015, from www.government.nl/files/documents-and-publications/reports/2013/07/31/key-figures-2008-2012/keyfigures-lr-compleet.pdf

Munowenyu, E. (2007). Assessing the quality of essays using the SOLO taxonomy: Effects of field and classroom-based experiences by A level geography students. *International Research in Geographical and Environmental Education, 16*(1), 21-43.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Nieveen, N. M. (2010). Formative evaluation in educational design research. In T. J. Plomp & N. M. Nieveen (Eds.), *An introduction to educational design research* (pp. 89-103). Enschede: SLO.

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.

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

Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal, 44*(4), 921-958.

Plomp, T. J. (2010). Educational design research: An introduction. In T. J. Plomp & N. M. Nieveen (Eds.), *An introduction to educational design research* (pp. 9-50). Enschede: SLO.

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

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*(1), 4-15.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249-255.

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

Roberts, M. (2013). *Geography through enquiry: Approaches to teaching and learning in the secondary school*. Sheffield: Geographical Association.

Schunk, D. H. (2003). Self-efficacy for reading and writing: Influence of modeling, goal setting, and self-evaluation. *Reading &Writing Quarterly, 19*(2), 159-172.

Schussler, D. L. (2006). Defining dispositions: Wading through murky waters. *The Teacher Educator, 41*(4), 251-268.

Schussler, D. L., Stooksberry, L. M., & Bercaw, L. A. (2010). Understanding teacher candidate dispositions: Reflecting to build self-awareness. *Journal of Teacher Education, 61*(4), 350-363.

SLO. (2012). *Handreiking schoolexamen aardrijkskunde vmbo. Herziene versie voor het examenprogramma van 2015* [Guideline for internal school-based examinations in pre-vocational geography

education. Revised version for the examination program of 2015]. Enschede: SLO.

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.

Spencer, L., Ritchie, J., Ormston, R., O'Connor, W., & Barnard, M. (2014). Analysis: principles and processes. In J. Ritchie, J. Lewis, C. McNaughton Nichols, & R. Ormston (Eds.), *Qualitative Research Practice* (2nd ed., pp. 269–293). London: Sage Publications Ltd.

Spielman, A. (2017). *HMCI's commentary: October 2017*. Retrieved October 19th, 2017 from: https://www.gov.uk/government/speeches/hmcis-commentary-october-2017

Stimpson, P. (1992). Assessment in geography: An evaluation of the SOLO taxonomy. In H. Schrettenbrunner & J. Van Westrhenen (Eds.), *Empirical research and geography teaching* (pp. 157-177). Utrecht; Amsterdam: Koninklijk Nederlands Aardrijkskundig Genootschap; Centrum voor Educatieve Geografie Vrije Universiteit.

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

Stoltman, J., Lidstone, J., & Kidman, G. (2015). Powerful knowledge in geography: IRGEE editors interview Professor David Lambert, London Institute of Education, October 2014. *International Research in Geographical and Environmental Education, 24*(1), 1-5.

Taras, M. (2007). Assessment for learning: Understanding theory to improve practice. *Journal of Further and Higher Education, 31*(4), 363-371.

Taras, M. (2009). Summative assessment: The missing link for formative assessment. *Journal of Further and Higher Education, 33*(1), 57-69.

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.

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

Van Berkel, H., Bax, A., & Joosten-ten Brinke, D. (2014). (Eds.) *Toetsen in het hoger onderwijs* [Testing in higher education]. Houten: Bohn Stafleu Van Loghum.

Van den Akker, J. (2010). Curriculum design research. In T. Plomp & N. Nieveen (Eds.), *An introduction to educational design research. Proceedings of the seminar conducted at the East China normal university, Shanghai (PR China)*, (pp. 37-50). Enschede: SLO.

Van den Akker, J. J. H., Kuiper, W., & Hameyer, U. (2004). *Curriculum landscape and trends*. Dordrecht: Kluwer Academic.

Van der Schee, J. A., & Vankan, L. J. A. E. (2006). *Meer leren denken met aardrijkskunde* [More thinking through geography]. Nijmegen: Stichting Omgeving en Educatie.

Van der Schee, J., Vankan, L. J. A. E., & Leat, D. (2003). The international challenge of more thinking through geography. *International Research in Geographical and Environmental Education, 12*(4), 330-343.

Van der Vaart, R. (2001). *Kiezen en delen: beschouwingen over de inhoud van het schoolvak aardrijkskunde* [Divide and choose: reflections on the content of geography as a subject in secondary education] (Inaugural lecture). Utrecht: Faculteit Ruimtelijke Wetenschappen, Universiteit Utrecht.

Van Veen, K., Zwart, R., Meirink, J., & Verloop, N. (2010). *Professionele ontwikkeling van leraren. Een reviewstudie naar effectieve kenmerken van professionaliseringsinterventies van leraren* [Professional development of teachers. A review study towards effective characteristics of teacher professional development interventions]. Leiden: ICLON.

Vankan, L. J. A. E., & Van der Schee, J. A. (2004). *Leren denken met aardrijkskunde* [Thinking through geography]. Nijmegen: Stichting Omgeving en Educatie.

Voogt, J., Westbroek, H., Handelzalts, A., Walraven, A., McKenney, S., Pieters, J., & De Vries, B. (2011). Teacher learning in collaborative curriculum design. *Teaching and Teacher Education, 27*(8), 1235-1244.

Weeden, P. (2013). How do we link assessment to making progress in geography? In D. Lambert & M. Jones (Eds.), *Debates in geography education* (pp. 143-154). 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*(1), 15-21.

Whitcomb, J., Borko, H., & Liston, D. (2009). Growing talent: Promising professional development models and practices. *Journal of Teacher Education, 60*(3), 207-212.

Wiliam, D. (1993). Validity, dependability and reliability in National Curriculum assessment. *Curriculum Journal, 4*(3), 335-350.

Wood, P. (2013). How is the learning of skills articulated in the geography curriculum? In D. Lambert & M. Jones (Eds.), *Debates in geography education* (pp. 169-179). London: Routledge.

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58*, 149-162.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice, 41*(2), 64-70.

Summary

Teachers' assessment practices tend to focus more on rote learning than on types of meaningful learning. This widely recognised finding from the literature has tremendous implications for students' learning. Students' learning will be limited when teachers mainly assess students' recall of knowledge, particularly when the tests are used for summative purposes.

There is hardly any evidence of or information about geography teachers' assessment practices. In addition, the level of assessment literacy geography teachers in the Netherlands have is unknown. Part of teachers' assessment literacy is their knowledge about assessments, their skills to construct assessments and their conceptions regarding the purpose of assessments.

The first phase of this educational design research investigated the extent to which teachers' assessment practices reflected the insights from the literature concerning teachers' assessment practices. A content analysis of school-based internal examinations in pre-vocational geography education in the Netherlands was conducted to identify the type of geographical knowledge and which cognitive processes were assessed. The analysis of 1108 test items from 49 examinations of 13 different schools showed that approximately 60 per cent of the test items assessed conceptual knowledge and, with regard to the cognitive dimension, 62 per cent of all test items assessed a type of remembering. The examinations contained very few complex test items focusing on evaluating and creating.

In the panel interviews, which were conducted after the content analysis, participants acknowledged these findings. However, participating teachers mentioned that these outcomes were not in line with their educational goals. Their practices seemed to be influenced strongly by external factors, such as the exit examinations.

The influence of high-stakes tests, such as the exit examination, was also one of the outcomes of a questionnaire completed by 74 teachers in pre-vocational geography education in the Netherlands. In the questionnaire, teachers responded that preparation for the exit examinations was one of the purposes of the internal examinations, despite the fact that these internal

examinations had distinct objectives in terms of the geographical content, as well as methods and skills. As a result, a considerable majority of the teachers responded that the content of the exit examination was also assessed in their internal examinations and that, as far as possible, the teachers applied the same format for test items in their internal examinations as that used in the exit examination.

With regard to selecting the test items for the internal examinations, the teachers who responded to the questionnaire hardly constructed test items themselves. On the contrary, the teachers estimated that only 17 per cent of the test items was self-constructed. The other test items were mainly taken from the text books and older exams. Remarkably, the estimation of self-constructed test items by the elder and more experienced teachers was lower than was the estimation of teachers with some years of experience. In addition, this group of teachers, who had teaching experience of five to 14 years, perceived the percentage of test items in their internal examinations focusing on higher order cognitive processes to be higher than did the more experienced teachers.

The outcomes of the content analysis and questionnaire, together with a review of the literature, were used to design an intervention. The aim of the designed intervention was to find a solution to the problem that teachers rarely construct test items themselves, and that a majority of test items in the internal examinations focus on the recall of knowledge. Furthermore, the evaluation of the designed intervention should provide insight into how teachers can be scaffolded in their professional development regarding assessment literacy. An important aim of the intervention, therefore, was to identify how and why a designed teacher professional development programme (TPDP) could evoke a change in teachers' knowledge, skills, beliefs and practices regarding summative assessment.

Thus, the designed intervention consisted of two main components: 1) a toolkit containing examples of test items, strategies and instruments to scaffold teachers to construct and score test items that focused on the cognitive processes that were identified as meaningful learning and 2) a TPDP to evoke lasting changes in teachers' knowledge, skills, beliefs and practices. The intervention was based on tentative design principles reflecting the criterion of meaningful learning. Consequently, all the materials in the toolkit

were aligned and intended to contribute to meaningful learning via a focus on cognitive processes transcending rote learning, the integration of new information and prior knowledge, and the principle of divergent assessment.

The outline of the TPDP was based on Clarke and Hollingsworth's interconnected model of professional growth. The tentative design principles guiding the design of the programme reflected the aim to change teachers' knowledge, skills, beliefs and practices through enactment and reflection. Therefore, the TPDP had the following characteristics:

- In order to stimulate the enactment of new knowledge and skills, the meetings in the TPDP were based on new theories, demonstrations, collaborative practice and peer feedback.
- To stimulate teachers' core reflection, the TPDP contained instruments in order to alter teachers' beliefs.
- To change teachers' practices, the TPDP incorporated collaborative practice during the meetings, and instruments and strategies that could be used in the teachers' classroom practice.

The first prototype of the intervention was evaluated with experts to gather information about the soundness of the intervention. The intervention was then re-designed, and was tested and evaluated with a group of six teachers. At this stage of the research process, the evaluation was mainly formative, aiming to test the practicality and feasibility of the intervention.

The evaluation of the intervention with the six teachers revealed that the teachers were positive about some of the materials, instruments and strategies, and were less positive about others. One of the elements about which the teachers were positive was the provision of examples of 'good practices' and collaborative discussions of these examples. The teachers also considered the examples of pre-structured test-items to be more feasible than were open test items focusing on evaluating or creating. The teachers were also positive about a flow chart as an instrument to scaffold students. Students confirmed these findings regarding the flow chart in mini-interviews following the observed classroom practices. The teachers were less positive about a model to score students' responses to test items. The participating teachers mentioned that they experienced problems when scoring and marking students' responses that diverged from those that they expected. The

teachers felt it was difficult to judge the extent to which these diverging responses reflected the students' levels of performance.

The participating teachers valued the programme they attended highly. In the questionnaire and interviews at the end of the programme, the teachers' responses indicated that they had experienced some professional growth regarding their assessment knowledge, skills, beliefs and practices. This growth seemed to be evoked especially by reflection on the students' responses and performances. However, alternative growth pathways also appeared to exist, indicating the diversity of possible growth pathways within the programme.

Based on these outcomes, the intervention was re-designed and conducted with a different group of eight geography teachers. The programme they attended was extended via the addition of a second stage to facilitate the possibility for the teachers to enact new knowledge and skills in their classroom practices and to reflect on these practices in relation to their extended knowledge and skills on one hand, and to the students' responses on the other.

The evaluation of this programme showed that the teachers had changed their practices. Content analyses of their internal school-based examinations - one before the start of the programme, one after the first loop and one after the second loop - revealed that the teachers included fewer items that focused on the recall of knowledge. The median rating of the percentage of test items focusing on the recall of knowledge dropped from 71 per cent at the beginning to 55 per cent after the first loop, and to 42 per cent after the second. The latter examinations contained a higher percentage of test items focusing on meaningful learning, particularly understanding and applying. Only a few test items focused on evaluating or creating. Therefore, it can be argued that the designed intervention contributed to solving the problem, namely that the majority of test items in school-based examinations in pre-vocational geography education focus on the recall of knowledge.

As part of this research design, it is also important to identify how the intervention worked and why. Statements from the participating teachers in questionnaires during the programme and in the subsequent interviews indicated that the participating teachers' knowledge, skills and beliefs had changed. An analysis of the teachers' statements indicated that the teachers

showed professional growth along multiple and varied pathways. For some teachers, the extension of their knowledge and skills and professional experimentation with the new knowledge and skills led to changes. For other teachers, reflection upon students' responses seemed to be more prevalent in their professional growth. In each situation, it seemed to be important that teachers had the opportunity to experiment in their classrooms based on their enactment of new knowledge and skills, to reflect upon them and to reflect on the perceived impact of a change in their practice on students' outcomes.

In conclusion, the results of this research support the idea that professional growth with regard to teachers' assessment knowledge, skills, beliefs and practices can be fostered through a TPDP concerning the relationship between summative assessment and meaningful learning via diverse growth pathways. Secondly, the outcomes of this research support the idea that teachers can achieve higher mastery levels in assessments through an in-service teachers' professional development programme that includes

1) a focus on reflection on teachers' conceptions and educational goals,

2) a focus on the constructive alignment of goals, instruction and assessment practices, and

3) collaborative practices based on extended knowledge and skills.

To accomplish these higher levels, it seems to be important that such a programme contains multiple integrated consecutive cycles, and is related to real-life settings.

It is highly recommended that future teacher professional development programmes concerning assessment literacy take these characteristics into account. Furthermore, it is important to acknowledge that teachers must be scaffolded over time in order to achieve higher mastery levels in assessment literacy, and have the opportunity to become role models for their colleagues within their school departments. To become such models, teachers should be supported with instruments to extend their pedagogical content knowledge and to scaffold students in their learning processes. An important instrument to extend teachers' pedagogical content knowledge could be a framework for learning progression in geography education. A promising instrument to scaffold students is a flow chart that assists students to reason geographically.

Equally important seems to be a model to judge and score students' performances. Such a model could be helpful for the teachers when scoring test items that are more open and focus on higher order cognitive processes, such as evaluating or creating. It could also be helpful for students to assess their own performances and those of their peers, thus contributing to bringing summative assessment practices more in line with formative practices and the purposes of assessment.

Finally, to change geography teachers' assessment practices, structural changes with regard to institutionalised and contextual factors concerning the geography curriculum, the exam programme and summative assessment practices also seem to be necessary. Firstly, a revision of the exam programme and the purposes of the internal school-based examinations (SEs) and exit examination (CE) is strongly recommended. As became apparent from this research, teachers' beliefs and practices are strongly influenced by the content and format of the CE, although the content of the exam programme of the CE differs from the content of the exam programme for the SEs. This raises serious questions with regard to the validity of the SE results in terms of content validity as well as construct validity. Teachers tend to use the same formats for test items in their SEs as those in the CE, thus copying test items that are mainly short answer, constructed response tasks and multiple choice items, with a focus on the recall of knowledge and the understanding of elementary conceptual knowledge. This same tendency can be found in tests accompanying the text books, and even in the final test student teachers in geography education must pass before graduating.

A rethinking of constructive alignment in pre-vocational geography education in the Netherlands is advocated in this thesis. To align the ultimate goals of geography education with instruction and assessment, reflection on these goals and on the purpose and composition of the exams as a lever for geography education is essential. Ultimately, these goals should direct the content of the curriculum and the exams, and not vice versa.

Samenvatting

Uit de literatuur blijkt dat docenten zich bij toetsing vaker richten op reproductie van kennis dan op andere cognitieve processen. Dit gegeven heeft sterke implicaties voor het leerproces bij leerlingen. Het leerproces bij de leerlingen wordt beperkt wanneer de leerlingen vooral uit het hoofd leren. Deze vorm van leren wordt sterker gestimuleerd bij toetsing met een summatief karakter.

Over de wijze waarop docenten aardrijkskunde toetsen, is maar weinig bekend, net als over de toetscompetenties van docenten aardrijkskunde. Deze toetscompetenties omvatten kennis over toetsing, de vaardigheden om toetsen te maken en overtuigingen over de functie en doelen van toetsen.

In de eerste fase van dit educatief ontwerponderzoek is onderzocht in hoeverre de toetspraktijk van docenten aardrijkskunde overeenkomt met de bevindingen uit de literatuur. Een inhoudsanalyse van schoolexamens aardrijkskunde in het vmbo moest uitwijzen op welk type aardrijkskundige kennis en op welke cognitieve processen een beroep gedaan wordt. Uit deze analyse van 1108 verschillende toetsitems in 49 schoolexamens van 13 verschillende scholen komt naar voren dat ruim 60 procent van alle toetsitems een beroep doet op conceptuele kennis, vooral het leggen van geografische relaties en verbanden. Met betrekking tot de cognitieve dimensie blijkt dat 62 procent van alle toetsitems een beroep doet op het reproduceren van kennis. De schoolexamens bevatten nauwelijks toetsitems die een beroep doen op cognitieve processen als evalueren of creëren.

De deelnemers van twee panelinterviews herkennen deze uitkomsten. Tijdens deze panelinterviews komt ook naar voren dat de deelnemende docenten de uitkomsten van de inhoudsanalyse geen goede afspiegeling vinden van de doelen die zij stellen in hun aardrijkskundeonderwijs. Zij geven aan dat de wijze waarop zij hun schoolexamens construeren sterk beïnvloed wordt door externe factoren, zoals het centraal examen.

De invloed van het centraal examen, een toets waar veel van afhangt, komt ook naar voren als een van de uitkomsten uit een enquête ingevuld door 74 aardrijkskundedocenten in het vmbo. Docenten zien de voorbereiding op het

centraal examen als een belangrijk doel van het schoolexamen, ondanks het feit dat de schoolexamenstof andere eindtermen kent naar inhoud en vaardigheden. Hierdoor toetst een groot deel van de respondenten in hun schoolexamens ook de stof van het centraal examen. De respondenten gebruiken zo veel mogelijk de formats voor toetsitems uit het centraal examen bij de constructie van de schoolexamens.

De docenten geven aan dat zij relatief weinig toetsitems (17 procent) voor de schoolexamens zelf maken. De overige toetsitems zijn afkomstig uit de lesmethode en oudere (school)examens. Opvallend is dat deze schatting lager uitvalt bij de groep oudere docenten en de groep met meer werkervaring dan bij de groep respondenten met enige werkervaring. Deze groep respondenten met werkervaring tussen de 5 en 14 jaar schatten ook het percentage toetsitems in hun schoolexamens dat een beroep doet op hogere cognitieve processen hoger in dan de respondenten met meer werkervaring.

De uitkomsten van deze inhoudsanalyse en enquête, tezamen met het literatuuronderzoek, zijn gebruikt om een professionaliseringsprogramma voor docenten als interventie te ontwerpen. Het doel van deze interventie is om een oplossing te vinden voor het probleem dat docenten weinig toetsitems zelf construeren en dat een meerderheid van de toetsitems in de schoolexamens een beroep doet op reproductie van kennis. Een belangrijk tweede doel van de interventie is om te identificeren hoe en waardoor het professionaliseringsprogramma bij docenten een verandering kan bewerkstellingen in hun kennis, vaardigheden, overtuigingen en hun toetspraktijk in relatie tot de schoolexamens. Evaluatie van de interventie kan inzicht bieden in hoe docenten ondersteund kunnen worden in hun professionele ontwikkeling op dit punt.

De ontworpen interventie bestaat uit twee hoofddelen:

1) een handleiding voor docenten met voorbeelden van toetsitems en met strategieën en instrumenten om docenten te ondersteunen bij de constructie van toetsitems en de beoordeling van de antwoorden van leerlingen op toetsitems die een beroep doen op hogere cognitieve processen;

2) de opzet van het professionaliseringsprogramma, met het oogmerk een verandering te bewerkstelligen in kennis, vaardigheden, overtuigingen en de toetspraktijk van docenten.

De handleiding is gebaseerd op voorlopige ontwerpprincipes die het criterium van ‘betekenisvol leren’ weerspiegelen. Betekenisvol leren is daarbij gedefinieerd als: alle cognitieve processen die meer zijn dan reproduceren. Alle materialen in de handleiding zijn op elkaar afgestemd en bedoeld om bij te dragen aan het stimuleren van toetsitems die een beroep doen op deze hogere cognitieve processen. De toetsitems en ondersteunende materialen hebben als kenmerk de integratie van nieuwe informatie met bestaande kennis te stimuleren. Een ander kenmerk is het principe van divergent toetsen, ofwel het toetsen wat een leerling weet of kan doen in plaats van of een leerling iets weet of kan doen.

De opzet van het professionaliseringsprogramma is gebaseerd op het model van professionele groei van Clarke en Hollingsworth. De voorlopige ontwerpprincipes die dit programma richting geven weerspiegelen het doel om bij docenten een verandering te bewerkstellingen door middel van reflectie en bewust handelen. Het professionaliseringsprogramma heeft de volgende kenmerken:

- Om het bewust handelen en toepassen van nieuwe kennis en vaardigheden door docenten te bevorderen is de inhoud van de gezamenlijke bijeenkomsten in het professionaliseringsprogramma gebaseerd op theorie, demonstratie, gezamenlijke oefening en peer feedback.

- Om de overtuigingen bij docenten ter discussie te stellen bevat het professionaliseringsprogramma een aantal instrumenten om kernreflectie bij de docenten te stimuleren.

- Om de toetspraktijk van docenten te veranderen, bevat het professionaliseringsprogramma gezamenlijke oefening tijdens de bijeenkomsten en een aantal instrumenten en strategieën die docenten kunnen inzetten in hun dagelijkse praktijk.

Het eerste prototype van de interventie is geëvalueerd met een aantal experts. Deze evaluatie was gericht op het verzamelen van informatie of de interventie solide en consistent was. De interventie is daarna herzien en vervolgens beproefd en geëvalueerd met een groep van zes docenten. In deze fase van het onderzoek was de evaluatie vooral formatief, gericht op de bruikbaarheid en haalbaarheid van de interventie.

De evaluatie met de zes docenten onthult dat de docenten positief zijn over sommige materialen, instrumenten en strategieën uit het professionaliseringsprogramma en minder over andere. Eén van de onderdelen uit het programma waar de docenten positief over zijn, is het aanreiken van goede voorbeelden van toetsitems en de gezamenlijke discussie over deze toetsitems. Goede toetsitems zijn volgens de docenten vooral de toetsitems met meer structuur. Minder overtuigd van de bruikbaarheid zijn ze over de meer open toetsitems die een beroep doen op evalueren of creëren. De docenten zijn ook positief over een stappenplan voor leerlingen. In mini-interviews na afloop van geobserveerde lessen bevestigen leerlingen de bruikbaarheid van dit stappenplan. De docenten zijn minder positief over de bruikbaarheid van een model om de antwoorden van leerlingen te beoordelen. De docenten geven aan dat ze vooral moeite hebben de antwoorden van leerlingen te beoordelen zodra deze afwijken van het beoogde antwoord. Het is voor de docenten in deze situaties vooral moeilijk te beoordelen in hoeverre afwijkende antwoorden het gewenste niveau bij de leerling weerspiegelen.

De docenten zijn positief over de opzet en inhoud van het professionaliseringsprogramma als geheel. In de vragenlijst en interviews die aan het eind van het professionaliseringsprogramma zijn afgenomen, geven de docenten aan dat zij door het programma enige professionele groei in kennis, vaardigheden, overtuigingen en in hun toetspraktijk hebben doorgemaakt. Deze professionele groei lijkt vooral te zijn veroorzaakt door reflectie op de reacties en prestaties bij leerlingen. Professionele groei wordt evenwel ook op andere wijzen geïnitieerd. Dit duidt erop dat dit professionaliseringsprogramma diverse patronen voor professionele ontwikkeling mogelijk maakt.

Op basis van deze uitkomsten is de interventie opnieuw aangepast en vervolgens uitgevoerd met een andere groep van acht docenten. Het professionaliseringsprogramma dat deze docenten volgden is uitgebreid met een tweede cyclus om de mogelijkheden om bewust nieuwe kennis en vaardigheden toe te passen in de dagelijkse praktijk te vergroten. Dit zou ook moeten leiden tot het stimuleren van de reflectie bij docenten op deze praktijk en op de reacties en prestaties bij leerlingen.

De evaluatie van dit professionaliseringsprogramma laat zien dat de docenten hun toetspraktijk met de schoolexamens hebben veranderd. Een inhoudsanalyse van de schoolexamens, een aan het begin van het programma, een halverwege en een aan het eind, laat een daling zien van het aantal toetsitems gericht op de reproductie van kennis. De mediaan in het percentage van toetsitems dat een beroep doet op het reproduceren van kennis daalde van 71 aan het begin van het programma naar 55 halverwege en 42 aan het einde van het programma. De schoolexamens aan het eind van het programma bevatten vooral meer toetsitems die vragen naar een vorm van begrijpen of toepassen, minder naar een vorm van evalueren of creëren. Op basis van deze uitkomsten kan geconcludeerd worden dat de interventie heeft bijgedragen aan het zoeken naar een oplossing voor het probleem dat de schoolexamens aardrijkskunde in het vmbo sterk gericht zijn op het reproduceren van kennis.

Een ander belangrijk doel van de interventie is om te identificeren hoe en waardoor het professionaliseringsprogramma werkt. Uit de reacties op vragenlijsten en interviews tijdens en na afloop van het programma komt naar voren dat docenten denken dat hun kennis, vaardigheden en overtuigingen veranderd zijn. De docenten komen via verschillende patronen tot deze veranderingen. Bij sommige docenten wordt dit vooral veroorzaakt door het aanbod van nieuwe kennis en vaardigheden in combinatie met het bewust toepassen ervan in hun dagelijkse praktijk. Net als bij de groep van 6 docenten waarbij de interventie als eerste is getest, is voor andere docenten uit deze tweede groep van 8 reflectie op de reacties en prestaties bij leerlingen belangrijker om professionele groei door te maken. In iedere situatie is het van groot belang dat docenten de kans krijgen om te experimenteren in hun dagelijkse praktijk met de nieuwe kennis en vaardigheden, en om te reflecteren op deze praktijk en te reflecteren op de gepercipieerde invloed van de toetspraktijk op de resultaten bij leerlingen.

De uitkomsten van dit onderzoek ondersteunen het idee dat professionele groei van docenten in hun kennis, vaardigheden, overtuigingen en in hun toetspraktijk in relatie tot de schoolexamens bewerkstelligd kan worden via een professionaliseringsprogramma dat ruimte biedt voor verschillende patronen om tot deze groei te komen. Bovendien ondersteunt dit onderzoek het idee dat docenten hun toetscompetenties op een hoger niveau kunnen brengen via een in-service professionaliseringsprogramma, gericht op:

1) reflectie op de overtuigingen van docenten met betrekking tot de doelen van het onderwijs,

2) afstemming van deze doelen op instructie en toetsing, en

3) gezamenlijke oefening op basis van nieuwe kennis en vaardigheden.

Om deze professionele groei te bewerkstelligen moet het professionaliseringsprogramma bij voorkeur meerdere integratieve cycli omvatten en ingebed zijn in de dagelijkse praktijk van docenten.

Daarnaast is het belangrijk dat docenten voldoende tijd krijgen om hun toetscompetenties op een hoger niveau te brengen en daarmee een rolmodel kunnen worden voor hun collega's binnen de school. Hiervoor moeten docenten ondersteund worden met materialen om hun vakdidactische kennis uit te breiden en om leerlingen te kunnen ondersteunen in hun leerproces. Een belangrijk instrument om de vakdidactische kennis uit te breiden zou een raamwerk voor progressie in het aardrijkskundeonderwijs kunnen zijn. Een veelbelovend instrument voor het ondersteunen van de leerlingen en ze te helpen geografisch te redeneren is het stappenplan uit dit onderzoek. Een ander belangrijk instrument voor het ondersteunen van docenten en leerlingen zou een model kunnen zijn om de prestaties van leerlingen te beoordelen. Dit model kan vooral helpen om de meer open toetsitems die een beroep doen op hogere orde denkvaardigheden, zoals evalueren en creëren, te beoordelen. Dit model kan ook de leerlingen helpen zichzelf en elkaar te beoordelen en daarmee de summatieve toetsing dichterbij formatieve toetspraktijken te brengen.

Ten slotte, om de toetspraktijk van aardrijkskundedocenten in het vmbo te kunnen veranderen, lijken structurele veranderingen in het aardrijkskundecurriculum, het examenprogramma en de summatieve toetsing van belang. Een herziening van het examenprogramma en de doelen van het schoolexamen en centraal examen is aan te bevelen. Uit dit onderzoek blijkt dat docenten in hun overtuigingen en toetspraktijk sterk beïnvloed zijn door de inhoud en format van het centraal examen, hoewel er verschillen zijn in de inhoud en doelen van het schoolexamenprogramma en het programma voor het centraal examen. Docenten gebruiken in de schoolexamens formats voor toetsitems die sterk lijken op die in het centraal examen, meestal toetsitems die een kort gesloten of open antwoord van leerlingen vragen of multiple

choice items, gericht op het reproduceren van elementaire conceptuele kennis. Ook in de lesmethode heeft dit soort toetsitems de overhand. Hetzelfde lijkt overigens ook te gelden voor de afsluitende kennisbasistoets voor docenten in opleiding. Dit roept belangrijke vragen op over de validiteit van de schoolexamenresultaten in termen van inhouds- en constructvaliditeit.

Een betere afstemming tussen doelen, inhoud en toetsing in het aardrijkskundeonderwijs op het vmbo wordt in dit proefschrift bepleit. Om tot een goede afstemming van de doelen van het aardrijkskundeonderwijs met de inhoud en toetsing te komen, is reflectie op deze doelen, op de inhoud van het aardrijkskundeonderwijs en op de examens belangrijk. Uiteindelijk zullen de doelen sturend moeten zijn voor de inhoud van het curriculum en de toetsing, in plaats van andersom. Examens die zijn afgestemd op deze doelen kunnen daarbij wel als een hefboom gaan fungeren voor goed aardrijkskundeonderwijs.

Dankwoord

In 2013 werd mij de mogelijkheid geboden onderzoek te gaan doen. Ik kreeg een nieuwe functie bij de lerarenopleidingen van Hogeschool Windesheim en van mij werd verwacht dat ik zou gaan promoveren. ‘Geen probleem’, zei ik, want dat past goed bij mijn eigen wens om onderzoek te gaan doen.

Ik had ook al een onderwerp bedacht. Al een aantal jaren liep ik met de gedachte rond om onderzoek te gaan doen naar de wijze waarop in het aardrijkskundeonderwijs getoetst wordt. Mijn eigen ervaringen als docent aardrijkskunde in het voortgezet onderwijs en op de lerarenopleiding, maar ook als ouder, hadden me gesterkt in het vermoeden dat dit een goed onderwerp zou kunnen zijn. Goed, omdat er nog relatief weinig aandacht aan besteed was vanuit de onderwijsgeografie. Goed ook, omdat er volgens mij ook wel enige winst te behalen zou zijn in de wijze waarop we toetsen. Ik wilde hieraan een bijdrage leveren door vanuit een onderzoeksperspectief nog eens goed te doordenken wat en hoe we toetsen.

Aan twee belangrijke voorwaarden, tijd en een onderwerp, was daarmee voldaan. Na het vinden van begeleiders kon het onderzoek van start gaan. Ik kwam er al snel achter dat het hebben van tijd, een onderwerp en begeleiders misschien wel noodzakelijke voorwaarden zijn om succesvol onderzoek te gaan doen, maar nog geen voldoende voorwaarden. Voor succesvol onderzoek zijn zelfregulerende vaardigheden veel belangrijker. Drie factoren zijn van invloed op deze zelfregulatie: kennis, motivatie en doorzettingsvermogen. Deze factoren speelden ook in mijn eigen onderzoek in toenemende mate een grote rol.

Gelukkig heb ik hierbij veel hulp gehad. Want het blijft natuurlijk een interessant fenomeen; bedenken hoe je het leerproces bij anderen kunt stimuleren maar tegelijkertijd je eigen leerproces reguleren. Het eerste lijkt soms gemakkelijker af te gaan dan het laatste. Ik ben daarom vooral mijn begeleiders, Joop van der Schee, Wilmad Kuiper en Tine Béneker veel dank verschuldigd.

Joop, vanaf het allereerste begin ben jij degene geweest die me gestimuleerd heeft. Tijdens de begeleidingsgesprekken, maar ook over de mail gaf je niet alleen goede feedback op de inhoud, maar eigenlijk ook altijd positieve

feedback op het proces. Voor mijn motivatie en doorzettingsvermogen waren dat belangrijke aspecten.

Wilmad, vanaf het begin heb jij me erop geattendeerd hoe belangrijk het is om me niet alleen te richten op de onderwijsgeografische kant, maar vooral ook te zoeken naar mogelijkheden om docenten te professionaliseren. Met jouw scherpe blik en oog voor detail wist je me voortdurend bij de les te houden en tegelijkertijd te stimuleren nieuwe inhoudelijke domeinen te verkennen en daarmee mijn kennis te vergroten.

Als laatste, maar niet in het minst, wil ik Tine bedanken voor alle feedback en steun. Jouw inbreng was van grote waarde voor de onderwijsgeografische kant van het onderzoek. Met je inbreng heb je ook een zeer positieve rol gespeeld in het bevorderen van mijn motivatie en doorzettingsvermogen.

Zoals hierboven gesteld was het voor mij ook belangrijk om tijd te krijgen voor mijn onderzoek. Ik ben daarvoor de Hogeschool Windesheim, en in het bijzonder mijn leidinggevenden in de afgelopen vijf jaar, zeer erkentelijk. Zonder de steun vanuit mijn werkgever was het voor mij niet mogelijk geweest het onderzoek uit te voeren en af te ronden.

Vanaf het begin van mijn onderzoek zijn er meerdere collega's geweest die mij geholpen hebben of feedback hebben gegeven op tussenproducten. Als eerste wil ik daarvoor Sietze van der Vinne hartelijk danken. Sietze, ook met een bont en blauw lichaam heb je me nog geholpen bij mijn onderzoek. Dank daarvoor! Marijke Metz, dank voor je hulp en feedback bij de statistische verwerking van gegevens in de eindfase. Meijken Engbers, Theo Peenstra, Han Noordink, Frederik van Oorschot, Peter Borgman, Roeland Breukelman, Thea Hooyman, Fer Hooghuis, Hans Palings, Tim Favier, Geert van den Berg, Martijn Willemse, Adwin Bosschaart en Joke Voogt, hartelijk dank voor jullie deelname aan panelgesprekken, interviews of de andere manieren waarop jullie mij van zeer waardevolle informatie en feedback hebben voorzien en daarmee geholpen hebben mijn kennis te verdiepen. Ook de collega's die misschien niet direct een inhoudelijke bijdrage hebben geleverd, maar wel regelmatig informeerden naar de voortgang van mijn onderzoek wil ik hartelijk danken. Voor mijn motivatie en doorzettingsvermogen waren deze gesprekken minstens zo belangrijk!

Mijn onderzoek zou niet mogelijk zijn geweest zonder de inbreng en deelname van de docenten. Alle docenten die hun schoolexamen hebben toegestuurd, de enquête hebben ingevuld of hebben deelgenomen aan panelgesprekken wil ik daarvoor zeer hartelijk danken. Mijn dank gaat hierbij vooral uit naar de docenten die hebben deelgenomen aan de twee rondes van het professionaliseringsprogramma. Zonder jullie namen te noemen wil ik hier mijn grote dank betuigen aan jullie deelname! Ik kijk op de bijeenkomsten van het programma met zeer veel plezier terug. Hopelijk hebben de bijeenkomsten jullie net zo veel energie gegeven als ze mij hebben gegeven. Hoewel de groepen niet heel groot waren en de bijeenkomsten plaatsvonden aan het eind van de middag en begin van de avond, zijn we er toch maar mooi in geslaagd om een paar keer het verzoek van anderen te krijgen of het niet wat rustiger kon. Ik zie dit als een mooie vorm van erkenning voor onze gezamenlijke bijeenkomsten.

Ik wil mijn familie en vrienden danken voor hun steun in de afgelopen jaren. Jullie vragen en interesse zijn belangrijk voor mij geweest. Het hielp om door te zetten op de momenten dat het onderzoek wat minder ging. Soms hielp het om het onderzoek wat te relativeren. Ook dat kon bij mij geen kwaad. In het bijzonder dank hiervoor aan Pa, Pjotr, Sven en Evi!

Als laatste wil ik Yvonne bedanken, mijn lieve vrouw. We hebben samen veel gewandeld. Tijdens deze wandelingen kon ik mijn ideeën als eerste toetsen bij jou. Daarmee heb ook jij bijgedragen aan het vergroten van mijn kennis over het onderwerp van mijn onderzoek. Maar nog veel belangrijker waren de andere momenten, de momenten waarop ik stoom moest afblazen. Voor mijn motivatie en doorzettingsvermogen was jouw steun onmisbaar!

Curriculum Vitae

Erik Bijsterbosch is a geography teacher educator at Windesheim University of Applied Sciences in the Netherlands. His work focuses on curriculum, pedagogy and assessment in geography education.

From 1985 to 1990, Erik Bijsterbosch studied human geography at Utrecht University. After graduation, he took the postdoctoral course on teaching geography in secondary education. In 1992, Erik became a geography teacher in secondary education at Goois Lyceum in Bussum. He held this position until 1999.

For one year, from 1998 until 1999, Erik combined his work as a geography teacher with the position of geography teacher educator at Windesheim. Since 1999, being a teacher educator at Windesheim has become his main appointment. From 2010 until 2013, this appointment was combined with a position as geography lecturer and teacher educator at Utrecht University. In 2013, Erik became curriculum leader for the undergraduate initial teacher training programme in general secondary education at Windesheim. As part of this appointment, Erik started his PhD research on professional development of geography teachers regarding summative assessment practices.

During his career, Erik became a professional member of a commission with the task to design a new examination programme for the general education track and the pre-university track (2001–2003), was chair of the commission on geographical education of the Royal Dutch Geographical Society (KNAG) in the Netherlands (2003–2006), in 2017, he became a member of the international 'Trends in International Geography Assessment Study' workgroup, which develops a Geography Assessment Framework as part of TIMSS (the Trends in International Mathematics and Science Study), and he became chair of the commission with the task to update the syllabus for the examination programme pertaining to school-based examinations in pre-vocational education (2018).

Appendices

Appendix A: The educational system in the Netherlands

The Dutch education system

In the Netherlands pupils start with primary education at the age of 4 years and attend secondary education at the age of 12 years. Secondary education in the Netherlands comprises three different types of education; a four year pre-vocational education track (VMBO), a five year general education track (HAVO) and a six year pre-university education track (VWO). The choice between these types of education after primary education is based on a judgement by the primary school and an external exam.
Roughly 53% of all pupils in secondary education attend pre-vocational education, which is subdivided in four learning pathways: the basic vocational programme (bl), the middle-management vocational programme (kl), the combined programme (gl) and the theoretical programme (tl). These pathways are geared to subsequent pathways in vocational education.

After pre-vocational secondary education, at an average age of 16, pupils can attend a college for vocational education. Pupils having completed the general education track (HAVO) can attend a university of applied sciences which leads to a Bachelor degree. Pupils that have completed the pre-university track (VWO) can attend academic higher education that leads to a three year Bachelor's degree programme and subsequently an one or two year voluntary Master's degree programme (Ministry of Education, 2013, p. 8).

This research is conducted in the theoretical programme (tl) of pre-vocational education. The examination programme in pre-vocational education differs from the examination programme in the general education track and the pre-university education track. In pre-vocational education the examination programme contains six areas of geography; (1) Sources of Energy, (2) Poverty and Wealth and (3) Boundaries and Identity are the three areas for the internal school-based examinations and (4) Weather and Climate, (5) Water and (6) Population and Place for the external end-of-school (exit) examination. Besides these three areas a separate area with specifications for geographical skills and methods is included in the examination programme.

The examination programme for the general education track and the pre-university education track roughly comprises four different areas of geography besides the area with specifications for geographical skills and methods; (1) a human geographical area about global patterns and processes , (2) a physical geographical area about (geomorphic) processes and change, (3) an area with patterns, processes and interaction between people and environment in a specific realm or developing country and (4) geographical issues on a national or regional scale.

Besides these differences in content, the examination programmes also differ in the complexity of the knowledge and the cognitive processes. The objectives in the pre-university track are more demanding than those in the general education track and these are more demanding than those in the pre-vocational education track. Even within pre-vocational education distinction is being made between the diverse pathways. In the combined and theoretical pathway (gl and tl) pupils, for instance, have to study a case about the Amazone within the area Sources of Energy, whilst the pupils in the other pathways don't have to study this case. Furthermore, pupils in the combined and theoretical pathway are frequently asked to describe and explain certain patterns or processes, where the pupils in the other pathways only have to describe these. In this way, the examination programme in pre-vocational education distinguishes both in the knowledge dimension as in the cognitive dimension between the several pathways.

Appendix B: Taxonomy table, based on the original taxonomy table of the Revised Taxonomy of Bloom

(Anderson, Krathwohl, et al., 2001)

Knowledge Dimension		***Cognitive Process Dimension***				
		Remember	***Understand***	***Apply***	***Evaluate***	***Create***
		Recognizing *Recalling*	*Interpreting* *Exemplifying* *Summarizing* *Inferring* *Comparing* *Classifying* *Explaining* *Differentiating*	*Executing* *Problem solving*	*Attributing* *Critiquing*	*Predicting* *Organizing*
Factual Knowledge						
(a)	*Knowledge of specific details and elements*					
(b)	*Knowledge of simple concepts and terminology*					
Conceptual Knowledge						
(c)	*Knowledge of classifications and categories*					
(d)	*Knowledge of geographical principles or relationships between concepts*					
(e)	*Knowledge of geographical models and theories*					
Procedural Knowledge						
(f)	*Geographical skills*					
(g)	*Geographical methods*					
(h)	*Knowledge of criteria concerning geographical skills and methods*					
Metacognitive Knowledge						
(i)	*Strategic knowledge*					

Appendix C: Examples of test items from analyzed internal school-based examinations

1) Example of test item assessing remembering factual knowledge.

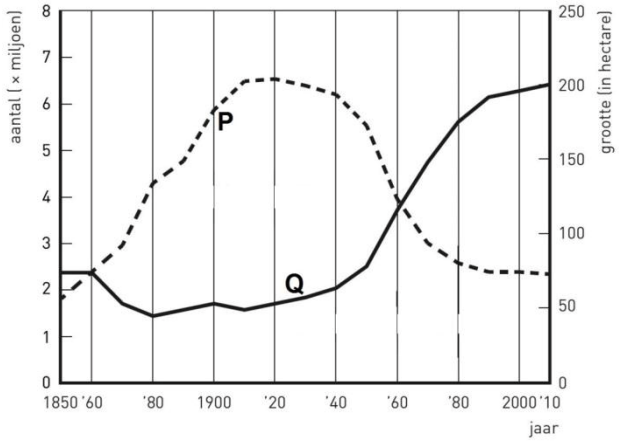
- In welk jaar werd Nigeria onafhankelijk?
- (*In which year Nigeria became independent?*)

2) Example of test item assessing remembering conceptual knowledge.

- Behalve saneren wil de gemeente ook iets doen aan de sociale cohesie in de wijk. Wat wordt er bedoeld met sociale cohesie?
- (*Except by remediation, the municipality wants to improve the social cohesion in the local district/neighborhood. What is meant by 'social cohesion'?*)

3) Example of test item assessing understanding conceptual knowledge of geographical principles or relationships between concepts.

Bron 6 Aantal boerderijen en hun gemiddelde omvang, in de VS (1850-2010). (*Figure 6. Number of farms and their average size in the US (1850-2010)*).



Gebruik bron 6.
(*Use figure 6*).

- a Neem de letters P en Q uit bron 7 over en schrijf erachter wat de lijn bij de letter weergeeft.
- (*a Write the letters P and Q on your paper en write behind it what the line for each letter indicates*).

- b Geef de verklaring voor de ontwikkeling van lijn P na 1910.
  (*b Explain the evolution of line P after 1910.*)

4) Example of test item assessing procedural knowledge.

- Gebruik kaartblad GB 181. Noem drie steden in het zuiden die het dichtstbevolkt zijn.
- (*Use atlas map 181. Mention three cities in the South with the highest population density.)*

5) Example of test item assessing evaluating.

Lees onderstaande nieuwsbericht:

Nederland trekt knip tegen sociale uitsluiting (21/11/13)
Nederland geeft relatief veel geld uit aan de bestrijding van sociale uitsluiting. Van alle Eu landen geeft alleen Cyprus een groter deel van haar budget hier aan uit. Nederland geeft wel veel geld aan de bestrijding van sociale uitsluiting. Hierbij wordt dan bijvoorbeeld geprobeerd om discriminatie terug te dringen.

- Vind jij dat Nederland minder geld moet uitgeven aan bestrijding van sociale uitsluiting? Leg uit waarom je dat vindt. Gebruik hierbij het begrip: sociale samenhang.

Read the news item below:

***The Netherlands invest against social exclusion** (21/11/13)*
The Netherlands spend a lot of money on the combat against social exclusion compared to other countries. Of all EU countries, only Cyprus spends a larger share of its budget on this combat. The Netherlands spend a lot of money to combat social exclusion. This is, for example, to reduce discrimination.

- *(Do you think the Netherlands should spend less money on the combat against social exclusion? Explain your answer. Use the concept of 'social cohesion' in your answer.)*

6) Example of test item assessing creating.

De regering is druk bezig om de achterstandsbuurten in de grote steden leefbaarder te maken. Het moeten weer krachtwijken worden. Ze hebben ook geld hiervoor vrijgemaakt. Jij mag een dag hierbij advies geven.

- Noem 3 verbeteringen/veranderingen die jij voorstelt en leg ze ook uit.
  Gebruik 100 woorden voor je antwoord. (tellen en het aantal woorden erbij zetten).

The government is working hard to make the poor neighborhoods more livable in the big cities. These neighborhoods should be revitalized. The government made money available for this revitalization. You're allowed to advice the government for one day.

- *Mention three improvements / changes you would propose and explain them.*
  ***Use 100 words for your answer**. (Count and put the number of words on your paper).*

Appendix D: Taxonomy table with numbers of test items in the first section of the toolkit and examples in Appendix E

<table>
<tr><td rowspan="4">Knowledge Dimension</td><td rowspan="4"></td><td colspan="5">Cognitive Process Dimension</td></tr>
<tr><td>Remember</td><td>Understand</td><td>Apply</td><td>Evaluate</td><td>Create</td></tr>
<tr><td>Recognizing
Recalling</td><td>Interpreting
Exemplifying
Summarizing
Inferring
Comparing
Classifying
Explaining
Differentiating</td><td>Executing
Problem solving</td><td>Attributing
Critiquing</td><td>Predicting
Organizing</td></tr>
<tr><td></td></tr>
<tr><td>Factual Knowledge</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>(a)</td><td>Knowledge of specific details and elements</td><td rowspan="2">2 test items, including example C.3b</td><td></td><td></td><td rowspan="6">2 test items, including examples C.3d and E</td><td></td></tr>
<tr><td>(b)</td><td>Knowledge of simple concepts and terminology</td><td></td><td></td><td rowspan="5">2 test items, including example D</td></tr>
<tr><td>Conceptual Knowledge</td><td></td><td></td><td></td><td></td></tr>
<tr><td>(c)</td><td>Knowledge of classifications and categories</td><td></td><td>1 test item</td><td></td></tr>
<tr><td>(d)</td><td>Knowledge of geographical principles or relationships between concepts</td><td>1 test item</td><td>7 test items, including examples A and C.3c</td><td></td></tr>
<tr><td>(e)</td><td>Knowledge of geographical models and theories</td><td></td><td></td><td></td></tr>
<tr><td>Procedural Knowledge</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>(f)</td><td>Geographical skills</td><td></td><td></td><td>6 test items, including examples B and C.3a</td><td></td><td></td></tr>
<tr><td>(g)</td><td>Geographical methods</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td rowspan="2">Knowledge of criteria concerning geographical skills and methods</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>(h)</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Metacognitive Knowledge</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>(i)</td><td>Strategic knowledge</td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

Appendix E: Examples of test items in first section of the toolkit

A. Understanding

Study figure 1.

Figure 1

1. 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. (3)

B. Applying

Study figure 2, which shows population statistics for Mali between 2000 and 2005.

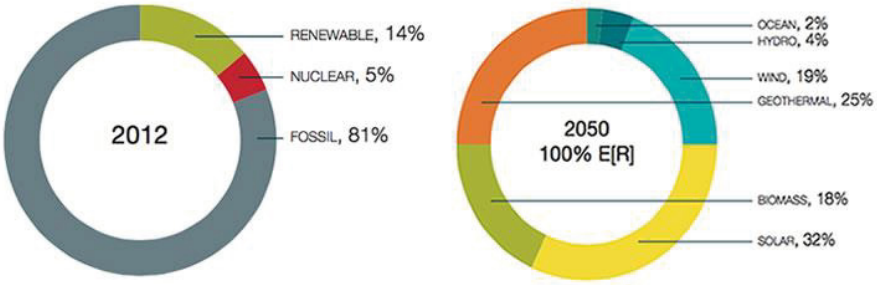
year	birth rate	death rate	net migration	life expectancy
2000	49.23	19.10	– 0.37	46.66
2001	48.79	18.71	– 0.36	47.02
2002	48.37	18.32	– 0.35	47.39
2003	47.79	19.21	– 0.34	45.43
2004	47.29	19.12	– 0.33	45.28
2005	46.77	19.05	– 0.33	45.09

Fig. 2

2. Calculate the population growth of Mali in 2005. You must show how you worked out your answer. [3]

C. Differential item

Figure 3: Development of global energy investments under the 100% energy (R)evolution case



Source: Energydesk, September 21, 2015 derived from https://energydesk.greenpeace.org/2015/09/21/heres-how-the-world-can-get-to-100-renewable-energy/

Study figure 3.

3a. Describe the changes in the global mix of energy resources until 2050.

3b. France, Brazil and the Netherlands must change their mix of energy resources to accomplish the predicted energy mix in 2050.

Which energy resource is still very important for these countries in 2012? Choose from the following resources: hydro, fossil and nuclear. You can use each resource only once!

Brazil: ……

France:……

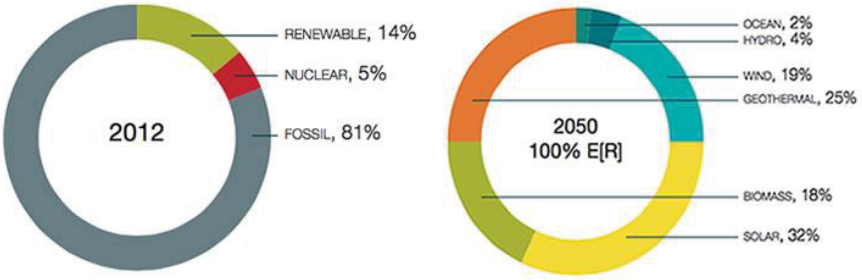
The Netherlands:……

3c. Figure 3 displays global development until 2050. Describe and explain how this development should look in the Netherlands to achieve 100% renewable energy resources by 2050.

3d. For which country, Brazil, France or the Netherlands, will the displayed development in figure 3 be most radical? Explain why you think that this development will be most radical for this country.

D. Predicting

Figure 3: Development of global energy investments under the 100% energy (R)evolution case



Source: Energydesk, September 21, 2015 derived from https://energydesk.greenpeace.org/2015/09/21/heres-how-the-world-can-get-to-100-renewable-energy/

Study figure 3.

4. Describe and explain what must be done in Brazil, France and the Netherlands to accomplish the mix of energy resources in 2050.

E. Evaluating

Study the text below about the Frisian identity.

It seems obvious: Frisian is spoken in the province of Friesland, and Limburgs is spoken in the province of Limburg. This obvious link, in this case, between a language and a province raises the question of whether such thing as a provincial identity exists and to what extent such identity coincides with local or regional identity.

Source: Cornips, L. & Stengs, I. (2010). Regionale identiteit; Lokale beleving van wie we zijn. Idee 31, (5), 10-13.

5. The authors raise the question whether a provincial identity exists in the Netherlands. What do you think – does a provincial identity exists? Write a short essay of approximately 100 words about this issue. Use the following guidelines:

- Demonstrate that you know how 'identity' can be defined.
- Evaluate to what extent a provincial identity coincides with a regional identity.
- Use examples to illustrate your opinion on this issue.
- Can you think of a counter argument?
- What is the relationship in the province you live in?