Measurement of Alignment between Standards and Assessment

Gunilla Näsström

Department of Educational Measurement
Umeå University
No. 3

Department of Educational Measurement
Umeå University
Thesis 2008

Printed by Print & Media, Umeå University: 2005262
September 2008

ISSN 1652-9650
ISBN 978-91-7264-662-9

Abstract

Many educational systems of today are standards-based and aim at for alignment, i.e. consistency, among the components of the educational system: standards, teaching and assessment. To conclude whether the alignment is sufficiently high, analyses with a useful model are needed. This thesis investigates the usefulness of models for analyzing alignment between standards and assessments, with emphasis on one method: Bloom's revised taxonomy. The thesis comprises an introduction and five articles that empirically investigate the usefulness of methods for alignment analyses.

In the first article, the usefulness of different models for analyzing alignment between standards and assessment is theoretically and empirically compared based on a number of criteria. The results show that Bloom's revised taxonomy is the most useful model. The second article investigates the usefulness of Bloom's revised taxonomy for interpretation of standards in mathematics with two differently composed panels of judges. One panel consisted of teachers and the other panel of assessment experts. The results show that Bloom's revised taxonomy is useful for interpretation of standards, but that many standards are multi-categorized (placed in more than one category). The results also show higher levels of intra- and inter-judge consistency for assessment experts than for teachers. The third article further investigates the usefulness of Bloom's revised taxonomy for analyses of alignment between standards and assessment. The results show that Bloom's revised taxonomy is useful for analyses of both standards and assessments. The fourth article studies whether vague and general standards can explain the large proportion of multi-categorized standards in mathematics. The strategy was to divide a set of standards into smaller sub-standards and then compare the usefulness and inter-judge consistency for categorization with Bloom's revised taxonomy for undivided and divided standards. The results show that vague and general standards do not explain the large proportion of multi-categorized standards. Another explanation is related to the nature of mathematics that often intertwines conceptual and

procedural knowledge. This was also studied in the article and the results indicate that this is a probable explanation. The fifth article focuses on another aspect of alignment between standards and assessment, namely the alignment between performance standards and cut-scores for a specific assessment. The validity of two standard-setting methods, the Angoff method and the borderline-group method, was investigated. The results show that both methods derived reasonable and trustworthy cut-scores, but also that there are potential problems with these methods.

In the introductory part of the thesis, the empirical studies are summarized, contextualized and discussed. The discussion relates alignment to validity issues for assessments and relates the obtained empirical results to theoretical assumptions and applied implications. One conclusion of the thesis is that Bloom's revised taxonomy is useful for analyses of alignment between standards and assessments. Another conclusion is that the two standard setting methods derive reasonable and trustworthy results. It is preferable if an alignment model can be used both for alignment analyses and in ongoing practice for increasing alignment. Bloom's revised taxonomy has the potential for being such an alignment model. This thesis has found this taxonomy useful for alignment analyses, but its' usefulness for increasing alignment in ongoing practice has to be investigated.

Key-words: alignment, standards, assessment, Bloom's revised taxonomy, the Angoff method, the borderline-group method, usefulness, validity

Acknowledgements

With the completion of this thesis, I made a long journey and faced a great challenge from being a farmer's daughter in the south of Lapland to complete this academic thesis. Many persons have helped and inspired me on this journey and I want to thank them all. However to be brief, I will restrict my acknowledgements to a few persons.

To begin with, I want to express my gratitude to my main supervisor Widar Henriksson, who has helped me through the whole process patiently and has always given me good advice. I also want to thank all my assistant supervisors who have supported me in one part each of my doctoral studies. Jan-Olof Lindström, my first assistant supervisor, has helped me to find my track in the jungle of research with his enormous knowledge about education. Kjell Lundgren, my second assistant supervisor, has encouraged me to go on. Peter Nyström, my third assistant supervisor, has through his questions and skills in formulating texts helped me to complete my thesis.

I want to thank everyone at the Department of Educational Measurement for creating a very pleasant atmosphere to work in. I also want to thank all the members of the project group for National tests for allowing me to still be a member and for their interest in my work. Many thanks also go to Susanne and Dave Alger as well as Gunnar Persson for proofreading all my manuscripts. I am especially grateful to Tova Stenlund, my previous roommate, for all our discussions about alignment, reliability, Bloom's revised taxonomy, children and animals and also for her proofreading of my Swedish summary. I also want to thank my other previous roommate Hanna Eklöf to me for creating a nice atmosphere to work in and her patience with everyone asking me about books in our library. Many thanks are also due to Lotta Jarl and Björn Sigurdsson for their help with all practical issues related to my dissertation.

No articles would have been written without any data. Therefore I want to thank all the teachers who have voluntarily participated in my studies. Many

thanks also go to Göran Bergqvist, Ingela Eriksson, Timo Hellström, Carl-Magnus Häggström, Anna Lind Pantzare, and Gunnar Wästle in the project group for National tests for their participation in my studies.

Finally there are some people outside the academic world who also have supported and inspired me on my journey toward completion of this thesis. My parents have raised me to believe in myself and to argue for my opinion. These abilities are important in the academic world, but I received my training more often on my knees in the potato field than by sitting in a seminar room. Therefore I want to thank my parents for being there for me and for being curious about my work. I also want to thank my two daughters, Helen and Sofie, for keeping me in reality and for all their hugs when I needed consolation. I also want to thank Göran for all his support and encouragement.

Measurement of alignment between standards and assessment

This thesis is based on the following articles:

I. Näsström, G., & Henriksson, W. (in press). Alignment of standards and assessment: A theoretical and empirical study of methods for alignment. Accepted for publication in Electronic Journal of Research in Educational Psychology, 6(3), xx-xx.

II. Näsström, G. (2008). Interpretation of standards with Bloom's revised taxonomy: A comparison of teachers and assessment experts. Submitted for publication.

III. Näsström, G. (2008). Alignment between standards and assessment: An evaluation of the usefulness of Bloom's revised taxonomy. Submitted for publication.

IV. Näsström, G. (2008). Interpretation of standards with Bloom's revised taxonomy: Does a division influence its usefulness? Submitted for publication.

V. Näsström, G., & Nyström, P. (in press). A comparison of two different methods for setting performance standards for a test with constructed-response items. Accepted for publication in Practical Assessment, Research & Evaluation.

All referencing to these articles will follow the enumeration used above.

Table of contents

1. Introduction

An educational system normally consists of three components: standards, teaching and assessment. Standards are in this respect descriptions in policy documents, defining what the students are expected to know and be able to do as well as how well the students are expected to attain this knowledge and these skills (Popham, 2003). Teaching is supposed to give all students an opportunity to attain such knowledge and skills (Fuhrman, 2001). Assessment is supposed to give information about how well the students have attained the expected knowledge and skills. When all three components work together, i.e. are aligned, education is expected to be efficient and students are expected to get an opportunity to learn what is expected (Biggs, 2003). Therefore, alignment is a fundamental condition for a functional standards-based educational system (Smith & O'Day, 1990).

There are many participants in an educational system: those who formulate the standards, all the teachers who work with and assess the students, and those who construct the large-scale assessments, and therefore there is a risk that the components are not function well together. To conclude if and to what degree the components are aligned, analyses are needed. Based on results from such analyses, changes can take place to increase the alignment among the components. Commonly, two components are compared to each other in such analyses, i.e. standards with assessment, standards with teaching or teaching with assessment (Roach, Niebling & Kurz, 2008).

Each of the components in an educational system consists commonly of several smaller parts. The standards are commonly more than one standard, an assessment consists most often of more than one item and teaching consists of a number of teaching activities. The first step in an alignment analysis is to categorize all the smaller parts of the compared components of an educational system with a model (Bhola, Impara & Buckendahl, 2003). Alignment models can be used for such a categorization, but these models differ in number of and definitions of criteria and categories in each

criterion. Commonly two criteria are used, one for content and one for cognitive complexity. A model that can categorize the components of an educational system with similar criteria as the ones in alignment models may also be useful in alignment analyses. Examples of such models are Guilford's taxonomy (1967), the framework in TIMSS (Mullis et al, 2001) and Bloom's revised taxonomy (Anderson & Krathwohl, 2001).

The result of this first step in an alignment analysis is a matching between the small parts of one component with the small parts of another component in an educational system. This matching is a question of whether each small part of one component, for example all assessment items, are aligned with at least one small part of the other component, for example with at least one standard. Such a comparison is important in the development of an assessment. However, even if all assessment items are aligned with standards, the assessment as a total may assess only a small proportion of the standards. Therefore, the degree of alignment between the total of one component with the total of another component is more interesting than comparing each small part of the compared components. Also Webb, as a pioneer in developing models for analyzing alignment, agrees that the degree of alignment is important in his definition of alignment: "… the degree to which expectations and assessments are in agreement and serve in conjunction with another to guide the system toward students learning what they are expected to know and do. " (Webb, 1997, p. 4). Expectations in this definition correspond to standards. The analysis of the degree of alignment between the totals of two components is therefore an important second step in an alignment analysis.

The more recently developed alignment models focus mainly on analyzing alignment of standards and assessments and consider alignment between standards and teaching only to a small proportion. There are several reasons for this circumstance. One reason is that standards are often defined without considering specific teaching methods and learning material (Smith & O'Day, 1990; Jongsma, 2007; SOU 2007:28) and therefore a variation in teaching is expected. A second reason is that assessments, especially large-

scale, standardized assessments, are expected to assess the standards because the results from assessments are often used to give feed-back to students (Guskey, 2007), for accountability decisions (Daggett, 2000), for evaluations of educational reforms (Herman, Webb & Zuniga, 2007), and are of great interest for the public. A third reason is that large-scale standardized assessments are expected to influence teaching (Linn, 2006). A high degree of alignment between standards and assessments is therefore sufficient and indicates with trustworthiness that the educational system is functioning.

This thesis deals with measurement of alignment between standards and assessments, because of the reasons mentioned above. The thesis focuses on evaluations of the usefulness of models for analyzing alignment between standards and assessments, with main emphasis on the usefulness of Bloom's revised taxonomy. Bloom's revised taxonomy is found to be the most useful model for categorization of standards and assessments, considering a number of criteria, in an investigation of possible models (see article I). Further analyses of the usefulness of Bloom's revised taxonomy are presented in articles II-IV.

Analyses of alignment between assessments and standards usually focus on standards that define what students are expected to know and be able to do. However, alignment between the standards that define how well students are expected to attain such knowledge and skills and the cut-scores set on a specific assessment is also important. If the cut-score for each stipulated performance level are well aligned with the standards, then students attaining a specific performance level have the knowledge and skills expected in these performance standards. To arrange this kind of alignment, a so-called standard-setting method is used. In this thesis two standard-setting methods, the Angoff method and the borderline-group method, are evaluated in article V.

1.1 Disposition of the thesis

This thesis consists of a summary and five empirical articles (article I-V). Following this introductory chapter, the contextual and the theoretical framework of alignment is presented, alignment is related to validity, and presents models for analyses of alignment are presented in chapter 2. In chapter 3 the most carefully evaluated model in this thesis for analyses of alignment, Bloom's revised taxonomy, is presented. The five articles are thereafter summarized in chapter 4 and issues regarding reliability and validity in these articles are addressed in chapter 5. In chapter 6 the main findings of this thesis are discussed and advantages and risks of striving for a high degree of alignment with all means are analyzed. Suggestions for further studies are given in chapter 7. In chapter 8, the thesis is summarized in Swedish. In the Appendix that then follows, the use of Bloom's revised taxonomy in alignment analyses is exemplified. Thereafter, the articles follow in numerical order.

2. Alignment

This chapter first defines alignment in educational settings, then relates alignment to validity and at the end presents procedures for measuring the degree of alignment between standards and assessments.

In educational settings, alignment is commonly referred to when two or all three components of an educational system, i.e. standards, teaching and assessment, are consistent (e.g. Biggs, 2003), in agreement (e.g. Bhola el al., 2003), matched (e.g. La Marca, 2001) or work together (Ananda, 2003). In its simplest form alignment is a match between one small unit of one component in an educational system and one or a couple units of another component. For example, one assessment item is matched to one or two standards. The result of such matching will just report how many of the assessment items are aligned with at least one standard. In the construction of an assessment it is important to only have aligned items, but it tells

nothing about how well the assessment as a whole assess the students' attainment of all standards. In an extreme case, all items in an assessment can be aligned to only one standard and the rest of the standards are left unassessed. Therefore, the degree of alignment between two components of an educational system is and should be more emphasized, because it expresses, for example, how well an assessment assesses all standards. The importance of the degree of alignment is emphasized by Webb (1997), Porter (2002) and Resnick, Rothman, Slattery and Vranek (2003-2004).

Before a more thorough presentation of alignment, the three components of an educational system, i.e. standards, teaching and assessment, are defined. Standards in education are a concept with different meanings in different settings, in different countries and at different periods in history. Sometimes the meaning of standards is implicit and sometimes explicit. Three meanings of standards are common in education: 1) standards as quality indicators; 2) standards as descriptions; and 3) standards in terms of performance standards.

In the first meaning, standards are quality indicators for an educational system (English, 2000). These types of standards are like the ones found in Standards for Educational and Psychological testing (AERA, APA & NCME, 1999). Standards as quality indicators deal with control of resources, personnel and business in order to give all students an opportunity to attain the expectations and also deal with the productivity of schools and the whole educational system (English, 2000). One commonly used quality indicator of productivity is the acceptable percentage distribution of students at different performance levels in schools and in the educational system as a whole (e.g. Newton, 2000; Wiliam, 2000). In norm-referenced educational systems such a quality indicator is logical, because in such systems a student's performance is compared to other students' performance for the purpose of ranking the students (Wiliam, 2000) and politicians may have an interest in prescribing the distributions they accept (Wood & Power, 1984). In standards-based educational systems a prescriptive percentage distribution of students on performance levels is illogical, because the students should, in

such systems, be distributed on the different performance levels based on how well they attain the expected knowledge and skills, not in comparison with other students. Common phrases like “high standards”, “the highest standard in the world” and “raising standards” are commonly related to standards as quality indicators.

The second meaning of standards, i.e. standards as descriptions of what students are expected to know and be able to do, is the most recent meaning of standards in education and is a result of the educational reform started in the US in the 1980s (Fuhrman, 2001). This educational reform is usually called the standards-based reform. In 1989, The National Council of Teachers of Mathematics was the first organization in the US to formulate these kinds of standards (Ravitch, 1992). However, descriptions of what students are expected to know and be able to do are not a new invention, they have existed for about one century but used to be called goals, aims, and objectives previously (Popham, 2000). Next a brief historical exposé of standards as descriptions of expected knowledge and skills is presented.

Policy descriptions of what students are expected to know and be able to do have been an important component of educational systems in almost a century. These descriptions have changed levels of specification and names in the course of history, but have often been called objectives. The development of objectives, to what have become standards today, started in beginning of the 20th century and has changed direction several times. In 1918 Bobbitt stated the need of formulating precise objectives to define the content of education. This approach resulted in a large number of objectives (Sosniak, 1994a). In the 1950s Tyler changed the direction toward a small number of consistent and important objectives. Tyler (1969) preferred more generally stated objectives defining both content and student behaviour. The next stage in the development of objectives came in the 1960s and stayed during the first half of 1970s. This stage was initiated by Mager and was influenced by programming teaching developed by Skinner (Popham, 1993). Objectives in this stage were commonly called behaviour objectives and specified both behaviour and performance in observable terms (Eraut, 1991). These objectives were

supposed to be precisely and clearly defined to avoid ambiguous interpretations (Popham, 1993). However, these specifications resulted in long lists of objectives. In the next stage in the development, the objectives became fewer, broader and more general again. In the early 1990s objectives were renamed to content standards in the US because of the implementation of the standards-based reform (Fuhrman, 2001). Sweden was also implemented a standards-based education in the middle of 1990s with curricula and syllabi defining the standards (Kjellström & Pettersson, 2005). In this educational reform, the so called performance standards were also included into policy documents, at least in Sweden (e.g. Kjellström & Pettersson, 2005). Content standards are descriptions of what students are supposed to know and be able to do, while the performance standards are descriptions of how well the students should attain this knowledge and these skills (Popham, 2003) often for a number of different performance levels. These kinds of performance standards can be both explicit descriptions in policy documents and concrete examples of performance at different levels (Linn, 1994).

There are different kinds of standards in the US today, ranging from long lists of narrowed facts or skills to broad and vague descriptions (Luft, Brown, & Slutherin, 2007). Popham (1997) concluded that standards developed in the 1990s were suffering from the same problems as the objectives had had: the standards were too many and too vague and general to provide sufficient clarity for teachers. During the last decade there has been a a tendency to camouflage a large number of narrowed standards under one broad defined standard (Popham, 2000), but there are no attempts to define standards more precisely. Therefore, Popham's conclusions from 1997 still hold good.

The third meaning of standards is connected to standard-setting and is most often called performance standards. In this meaning, performance standards are both quality definitions of how well students should attain specific knowledge and skills for different performance levels and a cut-score for the respective performance level on a specific assessment (Kane, 2001). These cut-sores are points on the score scale for the specific assessment and are assumed to be an operationalization of the definitions (Kane, 2001). The

procedure of transforming the definitions to cut-scores is called standard-setting and this is actually an aligning of the definitions and the assessment. The goal is to set cut-scores that are consistent with the standards.

In this thesis, standards are defined as descriptions of what students are expected to know and be able to do as well as descriptions of how well students are expected to attain such knowledge and skills. This kind of standards are presented in policy documents, like curricula and syllabi. A syllabus is simply a document organizing a subject, i.e. an area of teaching (Eash, 1991), which naturally will include subject specific standards. In contrast, a curriculum is a concept with different meanings in different settings (Connelly & Lantz, 1991). The definitions range from "'everything that goes on in school'" (Windh & Gingell, 1999, p. 52) to documents with standards as descriptions (Donn, 1994). A curriculum is distinguished from the syllabus, in that a syllabus concerns a specific subject while a curriculum concerns an educational system as a whole (Windh & Gingell, 1999). In Sweden, a curriculum is defined by a policy document for each educational level (e.g. Utbildningsdepartementet, 1994), defining standards for the education as a whole. A syllabus in Sweden is a policy document for one subject at a specific educational level, containing two types of standards, namely goals and grading criteria (Kjellström & Pettersson, 2005). Goals are descriptions of what students are expected to know and be able to do. Grading criteria are mainly descriptions of how well students at each performance level are expected to attain these goals, but some grading criteria are also describing the kinds of knowledge and skills students are expected to have. In this thesis, syllabi as well as curricula are defined as policy documents including standards as descriptions.

In article I, performance standards are used as descriptions of how well students are expected to attain knowledge and skills, while in article V performance standards are defined as the third meaning of standards.

Teaching is another component of an educational system and is here defined as the process of supporting the students to acquire expected knowledge and skills by means of instruction (Safritz, Koppe & Soper, 1988). Thereby

teaching is all kinds of instruction that is offered to the students to give them an opportunity to attain all standards.

Assessment as the third component of an educational system is in this thesis defined as a procedure of collecting, synthesizing, and interpreting information about students' achievement to support teachers in their decision making, to evaluate teaching (Popham, 2003), and to evaluate schools (Linn, 2006) and educational reforms (Herman et al., 2007). Results from an assessment can also be used in accountability decisions and to support the student in their further learning. Assessment includes all kind of judgements about students' achievement, i.e. written, oral, performance and all other kinds of items. Paper-and-pencil tests are only one form of assessment. In this thesis, except in article I, item is used for the smallest part of an assessment. In article I, question is used equivalent to item.

2.1 Alignment as chain links

Alignment can be compared to links in a chain and the components of an educational system will then represent the jewellery between the links (see Figure 1). In a chain, the strength of each link decides how strong the chain will be and how strongly the jewellery will be held together. If one of the links is easy to break, then the jewellery may easy fall apart and even get lost. If, on the other hand, all links are strong then the chain will resist a lot of violence and still keep the jewellery close together.

If the links are strong in an educational system, i.e. if there is a high degree of alignment, then the components of an education system will hold together and give the students a good opportunity to attain the standards. It is expected that a high degree of alignment between the components will improve students' learning (La Marca, Redfield, Winter, Bailey & Hansche, 2000; Anderson, 2002; Farenga, Joyce & Ness, 2002; Biggs, 2003), make an educational system effective (Webb, 1997), be important in evaluations of educational reforms (Herman et al., 2007), and give students, parents, the public and the politicians proper information (Herman et al., 2007).

Therefore, alignment is assumed to be a fundament in standards-based education (Smith & O'Day, 1990; Fuhrman, 2001). These advantages of alignment are also assumed when the link between only standards and assessments is considered. Besides these advantages, a high degree of alignment between standards and an assessment is also important in the validation of interpretation of assessment results (La Marca, 2001; Rothman, 2003) and for accountability decisions (La Marca, 2001; Koretz & Hamilton, 2006; Haertel & Herman, 2005).

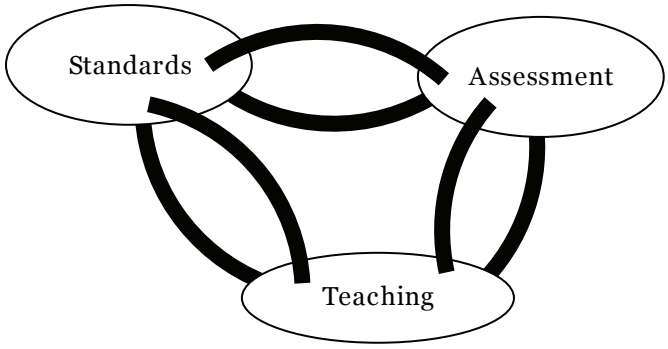

Figure 1. Alignment as chain links between the components of an educational system

A high degree of alignment between standards and an assessment indicates that most of the standards are assessed and that there is a balance between the standards and the assessment items, i.e. a strong link. This may influence teachers to teach all standards, i.e. the teaching will be more closely aligned to standards and assessment. With a high degree of alignment, an assessment also assesses the students' attainment of most of the expected knowledge and skills. The students are offered an opportunity to attain all standards and their learning is expected to be high (Linn, 1994). Hence the education will be efficient and the results from the assessment gives information of how well students attain the expected knowledge and skills, a good basis for accountability decisions and for information to students, parents and the public.

If, on the other hand, the links in the chain are weak, i.e. there is a low degree of alignment, then the components will easily drift apart and give different messages to the students about what they are supposed to know and be able to do. This may have negative consequences for the whole educational system. When only a small sample of the standards is assessed in an assessment, i.e. the link is weak, then there is risk that teachers and schools will focus only on the assessed standards and leave the unassessed standards out (Linn & Baker, 1996; Resnick et al., 2003-2004), i.e. weaken the link between standards and teaching. Thereby the students will get a smaller opportunity to attain all standards. Even if teachers teach all standards, many students have a strategy to focus their learning on what is assessed (Miller & Parlett, 1974) and therefore give themselves a smaller opportunity to learn all standards. In such case the education system will be less efficient and the basis for accountability decisions will be weaker.

If the links are too weak, then they will easily break and the two components they are holding together will split. If, for example, the link between teaching and assessments will break, then these two components will be isolated from each other and may emphasize different knowledge and skills.

A low degree of alignment between standards and an assessment is problematic for an educational system. However, when an assessment has items that are not matched to any standards this can damage the whole educational system and give students disadvantages. In such case the students will have a smaller opportunity to perform well on the assessment, and especially on the un-aligned items, if they have received teaching that is aligned with the standards. There is also a risk that, after the appearance of un-aligned items in standardized assessments, the teachers will change their teaching to also include such knowledge and skills that are assessed but not defined by the standards and may also exclude un-assessed standards. The teachers' intentions will then be to give their students a good opportunity to perform well on such an assessment, but thereby they weaken the link between standards and teaching. Even if it is easy to exclude un-aligned

items in the construction of an assessment, such items have appeared in standardized assessments (Resnick et al., 2003-2004).

The most efficient way to achieve a high degree of alignment is to start with an analysis of the standards and then develop teaching and assessments to match these standards (Resnick et al., 2003-2004; Baker, 2004; Martineau, Peak, Keene & Hirsch, 2007). To use the chain metaphor: It is most efficient to start with the standards and then make the links to teaching and assessment strong. However, it is more common to have a conviction that there is a high degree of alignment among the components and after the teaching or after an assessment has been designed verify this conviction with an analysis with a specific model.

2.2. Validity and alignment

Alignment between standards and an assessment is assumed to be important for validation of assessment scores (La Marca, 2001; Rothman, 2003), but alignment and validity are two distinct concepts with a relationship between each others (Webb, 1997). Alignment refers to how consistent two or all three components of an educational system, i.e. standards, teaching and assessments, are (Webb, 1997). According to Standards for Educational and Psychological testing (AERA et al., 1999, p. 9): "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by purposed uses of tests." By this definition of validity, an assessment itself is not a validity issue, it is the interpretations of assessment results that are validated.

In the following section the discussion about the relationship between alignment and validity is situated in standards-based educational systems and is based on three prerequisites for standards-based education: Firstly, the standards are the centre around which teaching and assessments should be arranged (Baker, 2004). Secondly, teaching is supposed to give all students an opportunity to attain all standards (Fuhrman, 2001). Thirdly, results from assessments, especially from large-scale and standardized

assessments, are used for assessing students' learning, for grading students, for evaluating educations, schools and maybe even teachers, for evaluating educational reforms (Herman, et al., 2007), for accountability decisions (Koretz & Hamilton, 2006), and in information to students, parents, the public and politicians (Herman et al., 2007). All these uses assume that the interpretations of assessments results state how well students have attained all standards. Therefore, alignment is a fundament of standards-based education and is also important for validation of interpretation of assessment results.

The degree of alignment between standards and an assessment can both be a characteristic of the assessment itself and evidence in the validation of interpretations of assessment results. It is possible to measure the degree of alignment before the assessment is administrated and thereby alignment is a characteristic of the assessment itself and not a validity issue according to the definition in Standards for Educational and Psychological testing (AERA et al., 1999). However, if there are several assessments to choose among, then the degree of alignment can be a strong argument for choosing one specific assessment. Hence, alignment influences interpretations of assessment results indirectly for all the uses of the chosen assessment and influences the consequential validity of the assessment.

The degree of alignment can also be evidence that may support interpretations of assessment results and therefore is an issue of validity. In the design of an assessment, especially a standardized and large scale assessment, general specifications for the assessment are formulated before the actual assessment is constructed and these specifications are commonly based on an analysis of the existing standards. The specifications for an assessment can either include all or a sample of the standards, depending on the limitations in time and resources or on directions from those who order the assessment. These specifications form the domain for the specific assessment. The type of evidence that alignment represents differs depending on whether the assessment specifications include all standards, just a sample of the standards or other knowledge and skills than those in the standards. Depending on what

is included in the assessment specifications, four different cases can appear and alignment has different roles and is related to validity in different ways in each case. Next, the four cases will be presented.

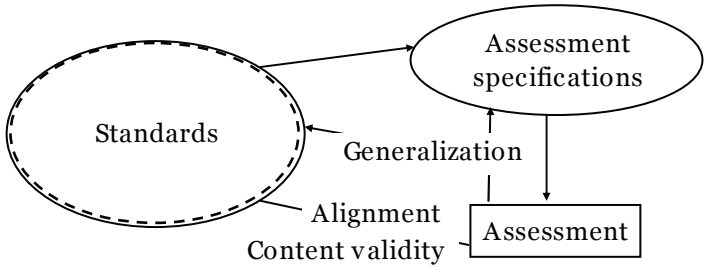

Figure 2. Case 1: The relationship among standards, assessment specifications and assessment regarding validity and alignment issues when all standards are included in the assessment specifications.

In the first case, all standards are included in the assessment specifications and this case is presented in Figure 2. In this case the items in the assessment should be a representative sample of all possible items assessing all standards. Alignment between all standards and the assessment is a measure of how representative the sample of items is and it is thereby the same as content validity. There is agreement among researchers that alignment is comparable to content validity (e.g. Webb, 1997; Anderson & Krathwohl, 2001; La Marca, 2001; Ananda, 2003; Bhola et al., 2003), but whether content validity is a validity issue is a matter of discussion. According to Messick (1989) content validity is a characteristic of a specific assessment and important for the construction of the assessment, but it does not directly influence the interpretations of assessment results and therefore is not a validity issue. On the other hand, Lissitz and Samuelsen (2007) claim that only internal characteristics of an assessment, with strong emphasis on content validity, are a matter of

validity. They also claim that external factors, especially considering interpretations of assessment results, are important but not a part of validity. Irrespectively of whether content validity is a validity issue or not, alignment, in this case, is the same as content validity.

Results from an assessment are commonly generalized to make an interpretation of the students' performance in the domain of the assessment (Kane, 2002), i.e. in this case interpretation about the students' attainment of all standards (see Figure 2). Alignment can therefore be used as one type of evidence to support the generalization of results to interpretations of performance of all standards. Thereby alignment is a validity issue.

The second case appears when only a sample of the standards is included in the assessment specifications. This case is visualized in Figure 3.

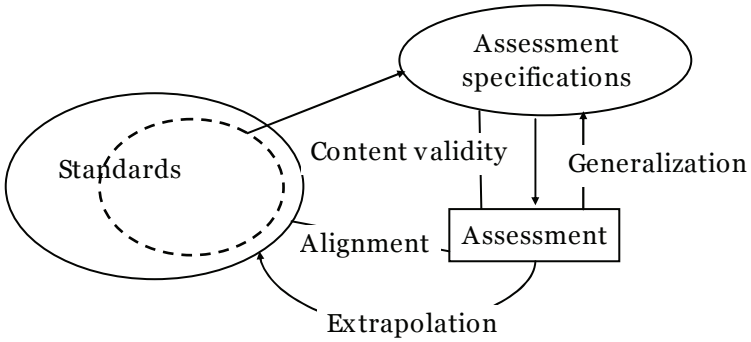

Figure 3. Case 2: The relationship among standards, assessment specifications and assessment regarding validity and alignment issues when only a sample of all standards are included in the assessment specifications.

In this second case (see Figure 3), the items in the assessment should be a representative sample of all possible items assessing only the standards in the assessment specifications. Content validity in this case is an evaluation of how representative the items in the assessment are in relation to the

universe of items assessing the standards in the assessment specifications. The analysis of alignment is still in relation to all standards. Therefore, in this case alignment and content validity are not the same thing.

Assessment results in this case will be generalized to the domain of standards in the assessment specifications and need to be extrapolated to make interpretations of the students' attainment of all standards (Kane, 2006). To be able to extrapolate appropriately, the standards in the assessment specifications should be representative of all standards. Alignment can hence be evidence supporting an extrapolation. Of course, if a larger proportion of all standards is included in the assessment specifications, then the probability of appropriate extrapolations is higher. In this respect alignment is a part of construct validity.

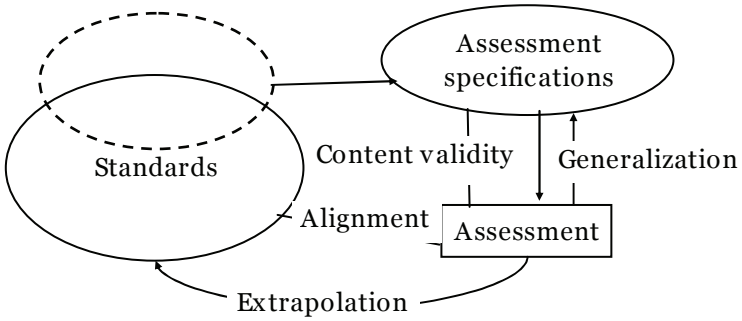

Figure 4. Case 3: The relationship among standards, assessment specifications and assessment regarding validity and alignment issues when a sample of all standards are included in the assessment specifications as well as other knowledge and skills.

In the third case, only a sample of the standards is included in the assessment specification together with knowledge and skills that are not defined by any standard (see Figure 4). In this case only a low degree of alignment between standards and assessment is expected. Content validity is in this case an evaluation of how representative the items are in

comparison to the assessment specifications, i.e. all knowledge and skills in the specifications, not only those in the standards. Alignment is even in this case a comparison between the assessment and all standards. Therefore, content validity and alignment are indicators of two different domains and are not comparable. Alignment can give only weak evidence for extrapolations of assessment results to make interpretations about students' attainment of all standards. The proportion of assessed standards will be small and has to be representative of all standards to be able to use alignment as evidence for extrapolation. The other assessed knowledge and skills can influence the extrapolation.

In this case, the assessment can influence how the teachers will teach their future students and therefore alignment can influence the consequential validity.

In a fourth case, the assessment specifications do not include any of the standards. In this case, alignment can only be used as an indicator of how similar the domain defined by the assessment specifications and the domain of the standards are.

Alignment can also influence the consequences of using an assessment, regardless of whether all standards or just a sample of them are included in the assessment specifications. If there is a low degree of alignment between all standards and an assessment, then grading of students, evaluation of teaching and educational reforms, accountability decisions and information will have a weak base as regards students' attainment of all standards. There is also a risk that teachers will teach only what is assessed and the students will thereby have less opportunity to learn all standards (e.g. Linn & Baker, 1996). Therefore alignment influences the consequential validity of an assessment.

In summary, alignment can be both a characteristic of the assessment itself and validity evidence.

2.3. Procedures for measuring alignment

For analyses of alignment several specific models have been developed, but models developed for other purposes have also been used for alignment analyses. The most frequently used models for analyzing alignment are Webb's (1997) model, the Achieve model (see e.g. Resnick et al., 2003-2004) and a model developed by Porter and Smithson (2001). This last mentioned model is sometimes called Surveys of Enacted Curriculum, but is in this thesis hereafter called Porter's model. The framework developed for TIMSS studies (e.g. Robitaille et al, 1993) has also been used in analyses of alignment between the framework and curricula in participating countries (e.g. Schmidt, McKnight & Raizen, 1997) and between the framework and textbooks (e.g. Howson, 1995).

Most reported alignment studies have analyzed alignment between standards and assessments (e.g. Bhola et al., 2003; Herman et al., 2007), but some studies have also analyzed alignment between standards and teaching as well as between teaching and assessment (e.g. Porter, 2002). Most of the commonly used alignment models today, except for Porter's model, have been developed exclusively for analyses of alignment between standards and assessments. Porter's model was developed for comparison between any pair of the components in an educational system, i.e. between standards and teaching, between standards and assessments and between teaching and assessments (e.g. Porter, 2002). Next in this chapter, common characteristics of models for analyses of alignment between standards and assessments, including Porter's model, are presented.

In almost all alignment models, the assessment items are categorized by at least two criteria and are related to the standards (Bhola et al., 2003). The two most commonly used criteria in alignment models are content and cognitive complexity (see article I). However, the definitions of these criteria and the categories in each criterion are at least partly different among the models. The number of categories in each criterion also varies among the models. In particular, the content criterion differs among these models. In

Webb's model and in the Achieve model the content categories are just the standards, while topics and subtopics are common content categories in other models (e.g. Porter & Smithson, 2001; Mullis et al, 2001). The criterion of cognitive complexity also varies partly among the alignment models, both in number of categories and in names and definitions of the categories. For example Webb (2007) calls this criterion depth of knowledge consistency and has four categories, while in Porter's model this criterion is called cognitive demand and the number of categories has varied between four and nine in different versions of the model (Porter & Smithson, 2001). By defining this criterion as cognitive complexity, the criterion is assumed to lie on a scale from low to high cognitive complexity.

Assessment items are always categorized individually by a panel of judges, but the different models have different approaches to the categorization of standards. The standards are either categorized, at least regarding the criterion of cognitive complexity, by the same criteria as for the assessment items individually by the judges (e.g. Porter, Smithson, Blank & Zeidner, 2007), through consensus discussion between all the participating judges (Webb, 2007) or by examples of assessment items given by an experienced reviewer (Rothman, Slattery, Vranek & Resnick, 2002). Most of these models require that the judges are familiar with the standards, either by their profession, by reading the standards in the training session (Rothman et al, 2002) or by consensus discussions (Webb, 2007). In Webb's (2007) model the judges, even though they are familiar with the standards, are expected to interpret the standards differently and therefore this model has included consensus discussions about the categorization of the standards in the training of the judges. In the other models, there seems to be an assumption that familiarity with standards is a sufficient condition for agreement in interpretation of standards.

Other models than the specifically developed alignment models may also be used in analyses of alignment between standards and assessments, if they offer some kind of tool for categorization regarding content and cognitive complexity. Such a model can be a framework or a taxonomy, like Bloom's

revised taxonomy. Because of the importance alignment has in standards-based educational systems, it may be expected that in the future new alignment models will be developed, the present models will be revised and models developed for other purposes will be used in alignment analyses. If a future alignment model is useful both for analyses of alignment and for encouraging alignment among the components of an educational system in the ongoing practices it will be extra valuable according to Roach et al. (2008).

The categorization of standards and assessment items indicates mainly whether each assessment item is matched to one or a couple of the standards, it says nothing about how well the whole assessment assesses all standards. Therefore, a measure of the degree of alignment between an assessment as a whole and all standards is more appropriate to report. The two most commonly used measures of degree of alignment are range and balance (e.g. Porter, 2002; Webb, 2002; Resnick et al., 2003-2004). Range can be defined as how large the proportion of the standards is assessed by at least one assessment item. Balance is often defined as how well the assessment as a total emphasizes the same categories as all the standards. Both Porter (2002) and Webb (2002) offer an index for calculating the degree of balance. These two indices are very similar and Porter's index is used in articles II-IV. This kind of measures is exemplified in the Appendix.

When the degree of alignment has been measured, the question whether the degree of alignment is acceptable has to be answered. There is no obvious answer to that question and to set an acceptable level is a policy decision that has to be taken by the governing body. However, different alignment models approach the question of the acceptable level of alignment in different ways. Some models offer cut-scores for an acceptable level for each criterion, while some models neglect this issue. In for example Webb's model, acceptable levels for each analyzed criterion and for the assessment as a whole are specified (see for example Webb, 2007). To give an example from Webb's model, an acceptable level of categorical concurrence, i.e. his content criterion, for one standard is to have at least six items assessing that standard (Webb, 2007). For the balance criterion, Webb has specified the

acceptable level at 0.7 (Webb, 2002). The Achieve model also offers acceptable levels for range and balance. In this model, the acceptable level for range is specified to be higher than 0.50 (Resnick et al., 2003-2004). In contrast, in Porter's model (e.g. Porter & Smithson, 2001) no acceptable levels are specified.

The methods for aligning performance standards and an assessment, also called standards-setting methods, have another approach regarding the appropriateness of the derived cut-scores. These methods use reliability and validity evidence to confirm the appropriateness of the derived cut-scores (e.g. Kane, 1994).

To summarize this chapter, alignment is an issue of how consistent the components of an educational system are. A high degree of alignment is expected to improve the students' learning, to evaluate and improve the efficiency of an educational reform and to be valuable for the appropriateness of accountability decisions. The degree of alignment between standards and an assessment can be both a characteristic of the assessment and validity evidence. Depending on if all or just a sample of the standards are included in the assessment specifications, results from alignment analyses can serve as different types of validity evidence. Several models for alignment analyses have been developed, but models developed for other purposes may also be useful if they can categorize both content and cognitive complexity. A categorization results in a matching between individually assessment items and individually standards, but to be able to determine the degree of alignment, the range and balance between an assessment as a whole and all standards are commonly measured.

3. Bloom's revised taxonomy

The main part of this thesis has evaluated the usefulness of one model for analysis of alignment between standards and an assessment and this model is Bloom's revised taxonomy (Anderson & Krathwohl, 2001). The choice of Bloom's revised taxonomy is based on a theoretical and an empirical investigation of possible models for alignment analyses presented in article I. In the theoretical investigation in article I, three conditions of a useful model were stated. The model should 1) be able to categorize both content and cognitive complexity; 2) assume that cognitive complexity lies on a scale; 3) be applicable in chemistry. Two models fulfilled these conditions, namely Porter's model and Bloom's revised taxonomy and these two models were empirically investigated in article I. The two models were applied on a syllabus and an assessment in chemistry for upper secondary schools in Sweden and the usefulness was investigated considering Hauenstein's (1998) five rules and inter-judge consistency. According to Hauenstein's five rules a model should 1) be applicable; 2) be totally inclusive, i.e. all standards and assessment items can be categorized; 3) have mutually exclusive categories, i.e. unambiguously place one standard or one item into only one category; 4) follow a consistent principle of order; and 5) use terms in categories and sub-categories that are representative of those used in the field. Two aspects of applicability were considered, whether judges can use the model to categorize both standards and assessment items, and to what degree the categories in the model are utilized. In this empirical investigation Bloom's revised taxonomy was found to be more useful than Porter's model. Therefore, the articles II-IV studied the usefulness of Bloom's revised taxonomy further.

Before a more thoroughly definition of Bloom's revised taxonomy and its relation to Bloom's original taxonomy, there is a need to define the concept taxonomy and relate it to categorization. Categorization, and classification, is simply an ordering of entities into categories and classes respectively (Wood & Linsey, 2007). In a taxonomy, the categories are ordered by their

relationship to each others and can thereby function as a categorization tool (Seigerroth, 2003). A taxonomy is often but not always hierarchical (Baliey, 1994) and in education particular many recent taxonomies have excluded the assumption of strict hierarchy, e.g. Hauenstein (1998) and Anderson and Krathwohl (2001).

3.1. Bloom's original taxonomy versus Bloom's revised taxonomy

Bloom's revised taxonomy (Anderson & Krathwohl, 2001) is as the name indicates a revision and a development of Bloom's original taxonomy (Bloom, Engelhart, Furst, Hill & Krathwohl, 1956). Bloom's original taxonomy was published in 1956 and has been applied to education since then (Anderson & Sosniak, 1994) and is still in use (e.g. Manaris et al., 2007; Zheng, Lawhorn, Lumley & Freeman, 2008). That taxonomy has been applied at all levels of education, from pre-school education (e.g. Bogan & Porter, 2005) to higher education (e.g. Granello, 2001) and in all types of academic subjects (e.g. Granello, 2001; Squire, 2001; Holmes, 2002; Castle, 2003; Kastberg, 2003; Pugente & Badger, 2003; Paziotopoulus & Kroll, 2004; Williams, Honghong, Burgess, Chenghui, Yuxiu & Yao, 2006; Manaris et al., 2007; Zheng et al., 2008). The use of this taxonomy is not limited to the US, it has been used all over the world (e.g. Chung, 1994; Lewy & Báthory, 1994; Squire, 2001; Williams et al., 2006).

The influence of Bloom's original taxonomy has been mainly on teaching (e.g. Anderson, 1994; Paziotopoulus & Kroll, 2004; Williams et al., 2006; "Lesson planning", 2007) and assessment (e.g. Anderson, 1994; Bennett, 2001; Holmes, 2002; Oliver, Dobele, Greber & Robert, 2004), but influence on standards and syllabi may also have occurred (Sosniak, 1994b). The common categorization of learning and thinking in lower and higher ordered levels can be seen as a result of the use of Bloom's original taxonomy (Anderson, 1994) and was a reaction against the large emphasis on facts in teaching and assessments previously.

Bloom's original taxonomy has been criticized as regards both the use of and the construction of the taxonomy. One criticism of the use is that the taxonomy is analytic and breaks the teaching and assessments into small parts, while learning is assumed to be holistic (Chyung & Stepich, 2003). Booker (2007) also criticizes the extended use of the taxonomy, because it has resulted in an emphasis on higher-ordered processes in teaching and reducing the learning of facts. Facts are, according to Booker, important for the learning of higher-ordered processes. Bloom's original taxonomy has also been criticized for being out of date (Sugrue, 2002).

The construction of Bloom's original taxonomy has also been criticized regarding the cumulative hierarchy and the order of the categories in the taxonomy. Bloom's original taxonomy (Bloom et al., 1956) consists of six categories: 1) Knowledge; 2) Comprehension; 3) Application; 4) Analysis; 5) Synthesis; 6) Evaluation. These categories are ordered from simple to complex and constitute a cumulative hierarchy, which means that a more complex category makes use of and is built on the less complex categories (Bloom et al., 1956). This assumption of a cumulative hierarchy has been investigated and questioned (e.g. Kropp & Stoker, 1966; Kreitzer & Madaus, 1994). The order of the categories according to complexity has also been studied and not been proven (Kreitzer & Maduas, 1994). Especially the ordering of Evaluation, Synthesis, and Analysis has been questioned. The category that has been most problematic in Bloom's original taxonomy is the Knowledge category, because it mixes different types of knowledge with the simple process recall. In the other categories also knowledge is required but only the Knowledge category contains different types of knowledge.

From the publishing of Bloom's original taxonomy in 1956 and to the start of the revision of the taxonomy in the late 20th century, a lot of educational research was performed and there have been several changes in the educational systems. The use of Bloom's original taxonomy was still being extended in the late 20th century and therefore a revision of the taxonomy was considered to be worth the efforts. The revision resulted in Bloom's revised taxonomy (Anderson & Krathwohl, 2001). Eight main changes of the

original taxonomy were made in the revised version. Firstly, the intended audience was changed from educators at college level to teachers at all levels of the educational system, with special emphasis on teachers in elementary and secondary schools. Secondly, the purpose of the taxonomy was also changed from being a tool for categorization of standards and mainly assessment items to analyzing and developing standards, teaching, and assessment as well as to emphasizing alignment among these components. Thirdly, the single dimension in the original taxonomy became two dimensions in the revised taxonomy. This split into two dimensions is mainly due to the change in the formulation of the standards. In the 1950s the standards mostly have defined content while since the 1970s the standards have combined content with a process. The content is often expressed as a noun and the process as a verb, together indicating what the students should be able to do (verb) with a specific content (noun) (Anderson & Krathwohl, 2001). The Knowledge category in the original taxonomy makes up the knowledge dimension in the revised taxonomy and has been split into three categories: *factual knowledge*, *conceptual knowledge*, and *procedural knowledge*. The process in the Knowledge category in the original taxonomy, recall, constitutes a category of its own, *remember*, in the second dimension of the revised taxonomy, the cognitive process dimension. Fourthly, a new category, *metacognitive knowledge*, has been added to the knowledge dimension in the revised taxonomy. In research, there was a trend at the time for the revision to incorporate metacognitive knowledge in education (see for example Marzano, 2001) and therefore this new category was added in the revised taxonomy. Fifthly, the change in formulation of standards has also changed the labels of the categories, from being only nouns in all categories in the original taxonomy to becoming verbs in the corresponding categories in the cognitive process dimension in the revised taxonomy. Sixthly, the claim that the taxonomy is a cumulative hierarchy has been dropped in the revised taxonomy. The authors of the revised taxonomy claim that the corresponding dimension from the original taxonomy, the cognitive process dimension, is hierarchical but only in the sense that they are ordered in terms of increased complexity.

Seventhly, two categories in the original taxonomy has not only changed form from nouns to verbs, but also been renamed. These two categories are: 1) Comprehend which has changed its name to *understand*; and 2) Synthesis which has changed its name to *create*. Eighthly, two of the categories in the original taxonomy have changed places on the complexity scale, namely synthesis (*create*) and evaluation (*evaluate*).

Next, a more detailed presentation of Bloom's revised taxonomy is presented.

3.2. Bloom's revised taxonomy

Bloom's revised taxonomy (Anderson & Krathwohl, 2001), has been developed for a wider purpose than alignment analyses, but alignment is also emphasized in the purpose. The authors state that:

> …teachers need a framework to help them to make sense of objectives and organize them so that they are clearly understood and fairly easy to implement. This framework may help teachers plan and deliver appropriate instruction, design valid assessment tasks and strategies, and ensure that instruction and assessment are aligned with the objectives. (Anderson & Krathwohl, 2001, p. XXII).

Bloom's revised taxonomy has two dimensions: the knowledge dimension and the cognitive process dimension. The knowledge dimension corresponds to content in alignment analyses and is defined as different kinds of knowledge. This dimension consists of four main categories with two or three sub-categories each. The main categories are *factual knowledge, conceptual knowledge, procedural knowledge* and *metacognitive knowledge*. The sub-categories are presented in Figure 5 (see Appendix). The categories in the knowledge dimension are assumed, by the authors, to lie along a continuum from concrete as in *factual knowledge* to abstract as in *metacognitive knowledge*. However, the authors admit that there is no clear-cut border between the main categories *conceptual knowledge* and *procedural knowledge*.

The cognitive process dimension corresponds to cognitive complexity in alignment analyses and is focused on how the knowledge is used. This dimension consists of six main categories with two to seven sub-categories each, with a total of 19 sub-categories. The main categories are *remember, understand, apply, analyze, evaluate* and *create*. The sub-categories are presented in Figure 5 (see Appendix). The underlying continuum in the cognitive process dimension is cognitive complexity, ranging from low cognitive complexity in *remember* to the highest cognitive complexity in *create*.

The two dimensions form a two-dimensional taxonomy table with 24 cells (see Figure 5 in Appendix). The four rows in the taxonomy table represent the main categories in the knowledge dimension and the six columns the main categories in the cognitive process dimension. The sub-categories in each dimension define the categories and should not be used separately. Therefore the sub-categories are collected under each category and only the categories form cells in the taxonomy table. A sub-category can be used to place a standard or an assessment item more easily in one category. Each standard and assessment item is first categorized in each dimension and then placed in the corresponding cell in the taxonomy table.

In articles I-IV the judges were allowed to place each standard and assessment item in up to three cells. The choice to allow a single standard or assessment item to be placed in more than one cell, i.e. be multi-categorized, was based on the fact that types of knowledge are commonly treated as categories without any ordering (e.g. de Jong & Ferguson-Hessler, 1996; Kjellström & Pettersson, 2005) even though the authors of Bloom's revised taxonomy assume an ordering. Hence the categories in the knowledge dimension are treated as independent categories in articles I-IV and do not include other categories, with one exception. *Factual knowledge* and *conceptual knowledge* are commonly treated as the same type of knowledge, often called declarative or conceptual knowledge. *Factual knowledge* can be compared to the cornerstones of *conceptual knowledge*. Therefore, *conceptual knowledge* includes *factual knowledge* and the judges were allowed to choose only one of

these two categories. The categories in the cognitive process dimension are ordered in a similar way as in other models, i.e. along a scale of cognitive complexity, and therefore the judges were allowed to choose only the category with the highest degree of cognitive complexity.

In the Appendix, the use of the taxonomy for categorization of standards and assessment items as well for alignment analyses is exemplified.

4. Summary of the articles

In this chapter the five articles that together represent the core content of this thesis will be summarized. All these articles evaluate the usefulness of models for analyzing alignment between standards and assessment mainly empirically. The first article (I) was co-written with Widar Henriksson and the fifth article (V) was co-written with Peter Nyström. In the other three articles (II-IV), I am the sole author.

The first four articles (I-IV) consider the usefulness of mainly one model for measuring the degree of alignment between standards and an assessment – Bloom's revised taxonomy. The evaluation of the usefulness was based on Hauenstein's (1998) rules and on intra- and inter-judge consistency. In the fifth article (V) two so-called standard-setting methods are compared and validated.

4.1. Article I. Alignment of standards and assessment: A theoretical and empirical study of methods for alignment

In article I, different models were compared regarding their usefulness for alignment analyses. A model is assumed to be useful for analysis of alignment between standards and assessments, if it can categorize standards as well as assessment items regarding two criteria, namely content and cognitive complexity. Cognitive complexity is assumed to lie on scale from low to high complexity and this is also required for a useful model. In this

article, a useful model should also be applicable to chemistry. Two models, Bloom's revised taxonomy and Porter's model, were found to fulfil these requirements and their usefulness was empirically investigated on a syllabus and an assessment in chemistry for upper secondary schools in Sweden. Their usefulness was compared based on Hauenstein's five rules and on inter-judge reliability. The results show that Bloom's revised taxonomy empirically was more useful than Porter's model.

4.2. Article II. Interpretation of standards with Bloom's revised taxonomy: A comparison of teachers and assessment experts

Article II reports a study where the usefulness of Bloom's revised taxonomy was evaluated for two differently composed panels of judges categorizing the standards in a syllabus in mathematics for upper secondary schools in Sweden. One panel was composed of four teachers in mathematics and the other panel was composed of four assessment experts. The categorization took place on two occasions for each panel and thereby both inter- and intra-judge consistency was reported for each panel. The levels of both inter- and intra-judge consistency were higher for the assessment experts than for the teachers. Another conclusion from this study is that Bloom's revised taxonomy is, according to Hauenstein's first three rules, useful on the whole. However, the usefulness of the taxonomy could have been better if a smaller number of standards had been placed in more than one category, i.e. multi-categorized. Three possible explanations of the large proportion of multi-categorized standards were given in the article: 1) vague and general standards; 2) the definitions of the categories in Bloom's revised taxonomy; 3) the instructions to the judges.

4.3. Article III. Alignment between standards and assessments: An evaluation of the usefulness of Bloom's revised taxonomy

In article III the usefulness of Bloom's revised taxonomy for analysis of alignment between standards and an assessment was further investigated. Bloom's revised taxonomy was applied to a syllabus and an assessment in mathematics for upper secondary schools in Sweden and the usefulness was evaluated considering Hauenstein's first three rules and inter-judge consistency. Five assessment experts formed a panel that categorized both the standards and the assessment items with Bloom's revised taxonomy on one occasion. One result of this study was that Bloom's revised taxonomy as on the whole is useful for analyses of both standards and assessment items. However, this study also reported both multi-categorized standards and multi-categorized assessment items, even if the assessment items were less multi-categorized than the standards. The level of inter-judge consistency was a bit higher for the assessment items than for the standards.

4.4. Article IV. Interpretation of standards with Bloom's revised taxonomy: Does a division influence its usefulness?

Article IV studied whether the large proportion of multi-categorized standards could be explained by vague and general standards. In this study the standards in a syllabus in mathematics for upper secondary schools in Sweden were divided by principle in Bloom's revised taxonomy, which states that a standard is composed by a verb and a noun. One panel of five assessment experts categorized the undivided standards on one occasion and the divided standards on a second occasion. A hypothesis was formulated: the division of the standards will result in a lower number of multi-categorized standards and a higher level of inter-judge consistency. The results showed a lower proportion of multi-categorizations for the divided

standards compared to the undivided standards, but a large proportion of the divided standards were anyway multi-categorized. No increase in inter-judge consistency was found. Therefore, the conclusion in this article was that the hypothesis was not verified. Other explanations of the large proportion of multi-categorized standards, like the definitions of the categories in Bloom's revised taxonomy and the instruction to the judges, are more likely than vague and general standards. The dual character of mathematics, i.e. intertwining of conceptual and procedural knowledge, may also be an explanation of the large proportion of multi-categorized standards and this explanation was also investigated in this article. Most of the multi-categorized standards were categorized as both conceptual and procedural knowledge supporting the explanation of the dual character of mathematics.

4.5. Article V. A comparison of two different methods for setting performance standards for a test with constructed-response items

Article V focuses on another dimension of alignment between standards and an assessment. In this article the relationship between performance standards, defined as descriptions of how well students are expected to know and be able to do at different performance levels, and a specific cut-score for each performance level on a score-scale for an assessment is studied. Performance standards are aligned with the score scale through a standard-setting procedure. In this study, the validity of two different standard-setting methods, the Angoff method and the borderline-group method, was compared when they were applied to a national test in mathematics for upper secondary schools in Sweden. Both methods were found to derive valid cut-scores. This study also found for the borderline-group method, a positive relationship between the level of performance for the whole teaching group and the performance for the borderline-group students in each teaching group. This indicates the importance of having representative students groups in the borderline-group method.

5. Reliability and validity issues in the articles

Reliability and validity are important issues that must be considered at all times, also in evaluations of models for analyzing alignment between standards and assessments. Lack of reliability means that the instrument yields different results on different occasions even though the standards and the assessment are the same. All the evaluated alignment models, i.e. Bloom's revised taxonomy, Porter's model, the Angoff method and the borderline-group method, are based on human judgements and therefore both intra- and inter-judge consistency are an important aspects of reliability. A high level of intra-judge consistency indicates stability in the use of the model, while a high level of inter-judge consistency indicates that the specific model will give the same result regardless of judges (Stephens, Vos, Stevens & Moore, 2006).

In articles I-IV inter-judge consistency and in article II also intra-judge consistency was reported. By treating the categories in the knowledge dimension of Bloom's revised taxonomy as independent categories, the obtained data in articles I-IV were on nominal level. Two indices of intra- and inter-judge consistency are dominant for data on this scale level: percent agreement and kappa (e.g. Watkins & Pacheco, 2000). Kappa is considered superior to percent agreement because of its ability to account for chance (Watkins & Pacheco, 2000; Goodwin, 2001), but both indices have both advantages and disadvantages. Percent agreement is easy to calculate and to understand and can be used with all kinds of data (Goodwin, 2001), but it is unable to account for chance and results of different studies are difficult to compare. Kappa takes chance agreement into account and is comparable across studies (Watkins & Pacheco, 2000), but this index also has disadvantages. The calculation of kappa is complex, a sufficient number of categorizations are needed in each cell and only data on nominal level can be used (Watkins & Pacheco, 2000). Kappa values can vary between -1.0 and +1.0 and little attention has been given to negative values (Goodwin, 2001). A value of 0 indicates that the

agreement is less than expected by chance and positive values indicate that the agreement is higher than expected by chance alone. However, what indicates negative values?

In articles I-IV in this thesis the judges were allowed to place one standard or one assessment item in more than one cell (i.e. multi-categorize) and the judges utilized this possibility. Multi-categorized standards and assessment items caused problems in the calculation of kappa, but not in the calculation of percent agreement. One unexpressed condition of kappa is that one categorized unit should be placed in only one category. The data in articles II-IV was to a fairly large extent multi-categorized and therefore needed be modified to be able to calculate kappa values. The cell that most judges agreed on for each unit was chosen for the calculation of kappa. This modification may have affected the kappa values, but there is no comparable index for multi-categorized data on nominal scale level.

The reported levels of inter-judge consistency for categorization of standards were about the same in articles I-IV for the assessment experts, while the levels were much lower for the teachers in article II. The reported kappa coefficients (Fleiss's (1971) kappa) were between 0.30 and 0.47 for the assessment experts and between 0.15 and 0.26 for the teachers. The percentages of perfect agreement were also reported in articles I-IV. In article I, the two judges agreed on the categorization of 53% of the standards and of 48% of the assessment items. In article II there were four judges in each panel and in articles III and IV five judges. In these three articles the percentages of perfect agreement ranged from 12% to 26% of the standards. In article III all the five judges agreed on the categorization of 30% of the assessment items. These results show a weak tendency that the percentage of perfect agreement is lower with a larger number of judges, while this pattern is not shown in the reported kappa coefficients. Therefore, kappa seems to be a better measure for comparisons of levels of inter-judge consistency between panels of different sizes.

Inter-judge consistency has also been reported for categorization of assessment items in articles I and III. The kappa coefficients were the same

in both studies, 0.36, indicating fair agreement following the scale of Landis and Koch (1977). There were, however, differences in level of percent agreement in these two articles. In article I the judges agreed on 48% of the assessment items, compared to 30% of the assessment items in article III. These differences may be due to differences in the size of panels.

Intra-judge consistency was only reported in article II and in this article two different composed panels of judges were compared when they categorized the same standards on two occasions. One panel was composed of four teachers and one panel of four assessment experts. The assessment experts categorized the standards in the same way on both occasions to a larger extent than the teachers. Using the scale of Landis and Koch (1977), the kappa coefficients indicate moderate agreement for the assessment experts and only slight agreement for the teachers.

There is a lack of studies comparing the levels of intra- and inter-judge consistency for categorization of standards. On the other hand, studies have reported measures of inter-judge consistency for categorization of assessment items with Bloom's original taxonomy (e.g. Fairbrother, 1975; Seddon, 1978) and with other alignment models (e.g. Herman et al., 2007). The kappa coefficients in articles I and III are higher than for categorization with Bloom's original taxonomy (Fairbrother, 1975), but lower than in the study by Herman, Webb, and Zuniga (2007).

The choice of using assessment experts instead of teachers in articles III and IV was based on the much higher levels of inter- and intra-judge consistency for the assessment experts in article II. The teachers selected in article II are assumed to be more qualified for analyses of standards than teachers in general, because of their engagement in the developing of national tests. Therefore even lower levels of inter- and intra-judge consistency can be expected if teachers in general categorize standards. If the levels of inter- and intra-judge consistency are too low, then the result of an alignment analysis should be questioned.

Even if it is important to conclude whether a model yields the same result when applied on different occasions and with different judges, this is only one part of the validation of interpretation of the results obtained with the studied model. Other types of validity evidence are needed.

Kane (1994) has developed a framework for evaluating the validity of standard-setting procedures, but no comparable framework has been developed for models measuring the degree of alignment between standards and assessments. Both standard-setting methods and alignment models are similar in their construction: 1) they are based on human judgements; 2) the judges should be experienced and preferably familiar with the standards and the students; 3) training of the judges is an important part of each procedure; 4) a method/model can only be more or less appropriately applied; 5) they give no perfect result, only a more or less trustworthy result; and 6) the results are used for policy decisions, like decisions about appropriate distributions of students on the different performance levels, for accountability decisions and in evaluations of educational reforms. Therefore, Kane's framework is also useful for models analyzing the degree of alignment.

Kane's framework for evaluating the validity of standard-setting procedures has tree main categories of validity evidence: procedural, internal and external evidence. Procedural evidence deals with how reasonably, systematically and defensibly the model has been carried out and thereby with the appropriateness of the result. Internal evidence deals with data generated within the procedure and with a specific focus on consistency of the results. Inter- and intra-judge consistency is important internal evidence, because high levels of inter- and intra-judge consistency indicate high trustworthiness of the result (Hambleton & Pitoniak, 2006). External evidence is, according to Kane (1994), based on comparisons with external sources. For example a cut-score or a degree of alignment is more trustworthy if different models or different panels of judges give similar results (Hambleton & Pitoniak, 2006).

All types of validity evidence were represented in the articles. In articles I-IV the evaluations of the usefulness of mainly Bloom's revised taxonomy was based on at least three of Hauenstein's (1998) five rules. These five rules deal with how appropriate the model is for deriving reasonable, systematic and defensible analyses of alignment between standards and assessments and therefore are procedural evidence. The procedural evidence supports the usefulness of Bloom's revised taxonomy for analyses of alignment between standards and assessment, with a slight reservation for the rather large proportion of multi-categorizations in mathematics in articles II-IV. The reported levels of intra- and inter-judge consistency represent internal validity evidence. The higher levels of intra- and inter-judge consistency for the assessment experts compared to those of the teachers (see article II) indicate a higher trustworthiness for the categorizations made by the assessment experts. Another internal evidence is the comparison between the two distributions of standards in the cells in Bloom's revised taxonomy reported in articles II and IV. If one distribution of the standards in the taxonomy table for all judges in one panel is similar to the distribution for all judges in the other panel, then this indicates a high level of trustworthiness of the analysis on group-level. This kind of comparisons in articles II and IV shows a high level of similarity and therefore the results have high trustworthiness.

In article I external evidence is also presented, in the form of a comparison of models. Two models were empirically investigated for alignment analysis: Bloom's revised taxonomy and Porter's model. This external evidence supports the validity of using Bloom's revised taxonomy rather than Porter's model.

In article V two standard-setting methods, the Angoff method and the borderline-group method, were compared and validated by all types of validity evidence according to Kane's framework. Both the Angoff method and the borderline-group method were found to derive valid cut-scores.

6. Discussion

This thesis has focused on evaluations of the usefulness of models for analyzing alignment between standards and assessments. The standards are of two types: standards which define what students should know and be able to do and performance standards that define how well the students attain the knowledge and skills. Different models for analyzing alignment between an assessment and each type of standards have been evaluated.

In this chapter, methodological considerations and results in the articles are first discussed, followed by reflections about Bloom's revised taxonomy. The chapter ends with a discussion about the advantages and risks about striving for high degree of alignment with ever means.

6.1. The articles

In this thesis, the main emphasis has been on the evaluation of usefulness of one model for analyses of alignment between standards and assessments, namely Bloom's revised taxonomy, and results about the usefulness have been reported in articles I-IV. In article V two methods for setting performance standards were compared and validated. First, methodological considerations in articles I-IV are discussed, followed by discussions about results from all five articles.

In article I, Bloom's revised taxonomy was found to be the most useful model for analyses of alignment between standards and assessment in chemistry. In articles II-IV the usefulness of the taxonomy for analyses of alignment was further investigated but in mathematics instead of chemistry. The reason for changing subject is that there are official national assessments in mathematics but no official ones in chemistry. The assessment in chemistry used in article I is a commercial assessment that is used in many schools in Sweden, but it has no official status as a national assessment. By changing the subject, the number of assessment experts and teachers with experience of developing national assessments also increased. Bloom's revised

taxonomy is developed to be useful in all kind of academic subjects and should therefore be as useful in mathematics as well as in chemistry. The change of subject hence should not have influenced the usefulness.

The same four assessment experts participated as judges in articles II-IV and the same fifth assessment expert was included in articles III and IV. The reason for choosing four judges in each panel in article II was simply that four was the size of the available reference group of teachers. In articles II-IV, the assessment experts were the same and this may have influenced the results in the article, especially in the later articles III and IV. One risk of having the same judges in several studies is that they learn how to categorize with the taxonomy and also get a chance to discuss their latest categorization with each other off the record. However, it was at least one month between two occasions of categorization and it was most probably that the judges had forgotten how they categorized the last time. On two successive occasions the judges categorized either only standards, standards and assessment items or divided standards and this decreases the possibility for the assessment experts to remember how they last categorized. If the judges had, off the record, discussed their last categorizations and thereby reached a higher degree of consensus, this would have shown in a higher level of inter-judge consistency in articles III and IV. That was not the case. In article I, two other assessment experts were engaged and the kappa coefficients in articles I-IV were about the same size, indicating that the judges may not have been influenced by each other between the occasions. The repetition of categorization may also have influenced the judges' ability to use the taxonomy, but this cannot be verified or rejected by the results in this thesis.

By having the same assessment experts as judges in articles II-IV, the generalizability of the results may be shaky. However, the intention was to evaluate the usefulness of Bloom's revised taxonomy for alignment analyses in a realistic situation and not to generalize the results to the population of assessment experts. Such a realistic situation is to have a small group of qualified judges to categorize standards and assessment items with Bloom's revised taxonomy and train them in how to categorize. When alignment

analyses become a frequent business, it is preferable and economical to have the same judges on several occasions because the training session can then be shortened. The situation in articles III-IV is then representative for such a preferable future.

In this thesis, the main emphasis has been on the evaluation of Bloom's revised taxonomy. This taxonomy was found in article I to be a more useful model than other models for analyzing alignment between standards and an assessment. In articles II-IV, Bloom's revised taxonomy was further studied, based on the first three of Hauenstein's five rules: 1) applicability of the taxonomy, 2) the inclusiveness, and 3) the exclusivity of the categories. In these three articles, the taxonomy was assumed to be applicable if the judges could use the taxonomy for categorization and if most of the cells in the taxonomy table were utilized. To evaluate the two excluded rules (rules 4 and 5) other types of studies are needed. The fourth rule states that the categories should be ordered by a consistent principle and this rule may be studied with comparable studies as the one by Kropp and Stoker (1966) and the ones presented in Appendix C in Anderson and Krathwohl (2001). To investigate the fifth rule, i.e. whether the terms in the categories and the sub-categories are representative for the field, experts can be asked to judge this.

In articles II-IV the taxonomy was found to be applicable, because all judges could use the taxonomy for categorization and because a majority of the cells in the taxonomy table were utilized with one exception. Only 11 of the 24 cells were utilized when the assessment items were categorized in article III compared to between 16 and 21 of the cells when the standards were categorized in articles II-IV. Assessments often assess only a sample of the standards and therefore it may be expected that the assessment items do not assess all standards. Therefore, the low level of utilization of cells for the assessment items is more likely to be explained by the limited range of chosen items rather than by a less useful taxonomy. In all articles, the taxonomy was also found to be totally inclusive because all standards and assessment items were categorized by all judges.

Hauenstein's third rule states that a useful taxonomy should have mutually exclusive categories, i.e. each standard and assessment item should be placed in only one cell. The judges were, however, allowed to multi-categorize, i.e. place each standard and assessment item in up to three cells in the taxonomy table, and they utilized this possibility to a great extent in articles II-IV, but not in article I. In article I Bloom's revised taxonomy was applied to chemistry and only one standard, and no assessment item, was multi-categorized. In articles II-IV, when the taxonomy was applied to mathematics, the judges multi-categorized between 14% and 90 % of the standards. In article III, 37% of the assessment items were multi-categorized. In article II the teachers multi-categorized standards to a lesser extent than the assessment experts on both occasions of categorization of same standards. The teachers multi-categorized 14% and 37% of the standards on the respective occasion, compared to 89% and 97% for the assessment experts. In all these four articles (I-IV) the judges received the same instruction and introduction and therefore the results from the articles can be assumed to be comparable. Four explanations of the large proportion of multi-categorized standards and assessment items in mathematics were suggested in articles II-IV: 1) vague and general standards; 2) the instruction to the judges; 3) the dual character of mathematics; 4) the definitions of the categories in Bloom's revised taxonomy.

The first explanation, i.e. whether vague and general standards can explain the large proportion of multi-categorized standards, was investigated in article IV. In this article, the standards were divided into sub-standards, following the principle in Bloom's revised taxonomy which states that a standard consists of a noun and a verb. However, also the sub-standards in article IV were multi-categorized to a great extent and therefore this explanation is less probable.

The second explanation, i.e. whether the instruction to the judges may have encouraged the judges to multi-categorize by allowing this option, is not investigated in these studies. However, in article I the judges were given the same instruction, but only one standard was placed in two cells by one judge

compared to many in articles II-IV. This contradicts the notion that the permission to multi-categorize stimulates the judges to multi-categorize. Thereby, this second explanation is less probably, but is not sufficiently investigated. Further studies are needed to find out how the instruction influences the proportion of multi-categorized standards and assessment items. One such study may be to ask judges whether there are any problems with placing one standard in only one cell.

The third explanation, i.e. whether the dual character of mathematics can explain the large proportion of multi-categorized standards and assessment items, can be discussed by comparing the four articles (articles I-IV). According to Rittle-Johnson & Alibali (1999) mathematics is often assumed to intertwine *conceptual* and *procedural knowledge* and thereby a standard and an assessment item may naturally be multi-categorized. In article IV almost every multi-categorized standard was categorized as both *conceptual knowledge* and *procedural knowledge*, and this supports the idea of the dual character of mathematics. The results in article I, where Bloom's revised taxonomy was applied to chemistry, also support the notion that the nature of mathematics is a probable explanation of the large proportion of multi-categorizations in articles II-IV. In article I, only one standard and no assessment item was multi-categorized, compared to a large proportion in articles II-IV. Even though chemistry also is a subject with a lot of conceptual and procedural knowledge, the intertwining nature was not found. This explanation is worth to investigate further, by, as a suggestion, comparable studies in other subjects than chemistry and mathematics. If such studies give similar results as in article I, then this support this explanation. However, if studies in other subjects also result in large proportion of multi-categorized standards, then other explanations than the dual character of mathematics are more probably.

The fourth explanation, i.e. whether the structure of Bloom's revised taxonomy, and especially the definitions of the categories, caused the large proportion of multi-categorizations, is not investigated in the articles. If the categories are not exclusively defined, as Hauenstein (1998) demands for the

categories in a taxonomy, standards can be placed in more than one category. Therefore, it is possible that the categories in Bloom's revised taxonomy are at least partly not unambiguously defined to a sufficient extent. Further studies are needed to find out if the categories need to be more unambiguously defined. However, to revise the whole taxonomy again is a huge task and therefore it is more economic to investigate and exclude other possible explanations before a revision.

The low levels of inter- and intra-judge consistency found in articles I-IV question the assumption that familiarity with the standards is a sufficient condition for consistent interpretation of standards. All judges participating in these articles are assumed to be familiar with the standards, because of their professions, but these articles report at most moderate agreement among the judges. Even more troubling are the results in article II, which also reports low levels of intra-judge consistency. If one person interprets the same standards in different ways on two occasions, then this indicates problem with the stability of the results with the model. These worrying results indicate the importance of making sure that the judges in alignment studies are well familiar with the standards and that differences among the judges are considered during the procedure or when results are reported.

The rather low levels of inter-judge consistency found in the articles I-IV may raise questions beyond the usefulness of Bloom's revised taxonomy, especially when interpretation of standards is important. In the Education Act (Ministry of Education and Science in Sweden, 2000) the importance of equivalent teaching and equivalent grading of all students in all schools in whole Sweden is emphasized. With low levels of agreement among teachers about how to interpret standards, the equivalence in education and grading in Sweden has to be questioned. Even more worrying is the result for the teachers in article II, which showed very low levels of inter-judge consistency, because these teachers are more trained to analyze standards than teachers in general. Therefore, it can be expected that teachers in general agree about interpretation of standards to an even less extent than the teachers in article II. This can have negative consequences for the

equivalent education and grading in the whole country. To increase the agreement about interpretation of standards, one practicable way is to arrange consensus discussions among teachers and schools.

In alignment models, the importance of having the judges interpret the standards consistently is emphasized (Bhola et al., 2003). However, to base an alignment analysis on the assumption that familiarity with the standards is a sufficient condition for consistent interpretation of standards should be questioned, because of the low levels of inter- and intra-judge consistency reported in articles I-IV. Different alignment models deal with consistency in interpretation of standards in different ways: In Webb's model disagreement in interpretation of standards among the judges is expected and therefore this model includes consensus discussions about how to interpret standards in the training session. Based on the results in articles I-IV this seems to be a sound approach. In the Achieve model, the judges are expected to be experienced and knowledgeable educators and become familiar with the standards through reading them (Rothman et al., 2002). This model offers examples of items categorized as belonging to a specific standard as direction to the judges on how to categorize the items, and these guiding examples together with recent familiarity with the standards is, according to the model, sufficient to attain consistency in interpretations of the standards. In Porter's model (e.g. Porter & Smithson, 2001) standards, assessment items and teaching are categorized with the same categorization tool. The judges in this model are supposed to be experts (Porter et al., 2007), but differences in interpretations of standards are not considered during the training session or when the results are reported. Based on the results in articles I-IV, the approaches of both the Achieve model and Porter's model do not consider the differences in interpretation of standards among the judges sufficiently. It is important for the trustworthiness of results from alignment analyses that either have consensus discussions about interpretation of standards in the training session or report levels of inter-judge consistency for the reader to consider when results of alignment analyses are reported.

The knowledge dimension in Bloom's revised taxonomy, with its general types of knowledge as categories is unique compared to content categories in other alignment models. In Webb's (1997) model and in the Achieve model (Resnick et al, 2003-2004), the content categories are the present standards and the assessment items are just matched with one or two standards in the analyses. These two models hence provide no tool for analyzing content in the standards. In Porter's model (Porter & Smithson, 2001), and in the models used in TIMSS (e.g. Mullis et al, 2001) and PISA (OECD, 1999), subject-specific topics make up the categories in content criteria and therefore a set of categories has to be developed for each subject to be analyzed. To categorize content as types of knowledge has both advantages and disadvantages, compared to the other models. One advantage is that the categories are generally defined and can be used in all academic subjects, while most alignment models are subject-specific and can only be used in one subject. Another advantage is that the use of the same model in different subjects facilitates comparisons of the degrees of alignment in different subjects. One disadvantage may be that even if an analysis found a high degree of alignment between standards and an assessment, there may be some topics that are unassessed. However, analysis of content is important also in the construction and in the validation of an assessment and in such analyses topics are naturally in focus. Thereby, content validation and alignment analysis may be complementary. The broad and general standards today are also an argument for categorization of content in general knowledge categories instead of in topics. If broad and general standards are categorized by topic-based categories, most of the standards will be placed in several categories to cover the whole content and thereby the proportion of multi-categorized standards will be high. With general content categories, the proportion of multi-categorized standards will be low. This was the result in article I. The standards and assessment items in chemistry were to a large extent multi-categorized by Porter's topic-based model, but only one standard was multi-categorized with Bloom's revised taxonomy. Lower levels of multi-categorizations increase the trustworthiness of the model

and makes the comparisons between standards and assessments easier to visualize. This supports the view that a model with general content categories, like Bloom's revised taxonomy, is trustworthier than topic-based models.

By providing content categories in an alignment model standards can also be analyzed with the same tool as the assessment items. This has an advantage: The judges are forced to interpret the standards and place them in the same kind of categories as the assessment items. Such a categorization can be used in consensus discussions about interpretations of standards or for reporting the levels of inter- judge consistency.

When Bloom's revised taxonomy is applied to education (e.g. Anderson & Krathwohl, 2001; Su, Osisek & Starnes, 2004; articles II-IV in this thesis), some combinations of categories from each dimension are more likely. Anderson and Krathwohl (2001) found that *remember factual knowledge*, *understand conceptual knowledge* and *apply procedural knowledge* are probably the most common used cells in the taxonomy table. In articles II-IV in this thesis, the cells *apply procedural knowledge* and *understand conceptual knowledge* were also commonly used, especially *apply procedural knowledge*. Two differences were, however, found in the article compared with the result in Anderson and Krathwohl: 1) the cell *apply conceptual knowledge* was the second most frequently used cell, and 2) the cell *remember factual knowledge* was used only to a very small extent for the standards and not at all for the assessment items. One explanation of the extended use of the cell *apply conceptual knowledge* in articles II-IV is the large proportion of multi-categorizations, because many standards and assessment items were placed in both *apply procedural knowledge* and *apply conceptual knowledge*. There are two explanations of the small use of *remember factual knowledge*: Firstly, there is a policy in Sweden that states that standards should be generally formulated and emphasize higher order thinking instead of long lists of facts. Secondly, the national assessments in mathematics consist mainly of constructed-response items and very rarely of selected-response items assessing *factual knowledge*. Even though some

cells were more commonly used, all, but one, cells, were utilized in articles II-IV. This indicates that the categories in both the dimensions are representative of standards and assessment items.

The concentration of many standards into a few cells seems not to be a problem for the usefulness of Bloom's revised taxonomy as long as most of the cells in the taxonomy table are utilized to some extent. However, if many of the standards are concentrated to a few cells, then this can limit the teaching and therefore have negative consequences for what kind of knowledge and skills the students are expected to attain. This is, however, a didactical problem and therefore out of the scope of this thesis.

Article V dealt with alignment between performance standards and the score scale for an assessment. In Sweden the performance standards are officially defined in each syllabus and are called grading criteria. The alignment between these performance standards and an assessment is created by setting appropriate cut-scores, which are used for separating students on different performance levels based on their performance on the assessment. In article V, a comparison of two methods for setting cut-scores, the Angoff method and the borderline group method, was reported. Both these methods resulted in valid cut-scores and a naturally attendant question is which method to recommend? The two methods are both recommendable, but they have different advantages and disadvantages. The choice of method should be based on what advantages are more valuable and which disadvantages are more problematic in the specific situation. The Angoff method requires a smaller number of judges and is easier to administrate than the borderline-group method. The borderline-group method is on the other hand simpler to understand and is a more natural task for the judges than the Angoff method. In the borderline-group method the judges are dealing with familiar students, while in the Angoff method the judges have to conceptualize hypothetical students. The borderline-group also requires a large number of representative student groups to obtain a sufficient number of representative borderline-students at each performance level. The Angoff method can be implemented before

the administration of the assessment, while the borderline-group derives cut-scores after the administration. The choice is now up to those who will use the cut-scores.

6.2. Reflections on Bloom's revised taxonomy

Even if there are several advantages of the content criterion in Bloom's revised taxonomy mentioned before, I want to make some remarks the ordering of the categories. This criterion is called knowledge dimension by the authors and is composed of four categories: *factual knowledge, conceptual knowledge, procedural knowledge* and *metacognitive knowledge*. These content categories represent the most common knowledge categories (e.g. de Jong & Ferguson-Hessler, 1996; Reis-Jorge, 2005), but the authors of Bloom's revised taxonomy assume that these categories lie on a scale from concrete to abstract. The authors admit, however, some overlaps on the scale among the sub-categories in *conceptual knowledge* and *procedural knowledge*. By assuming the ordering of these categories, the authors also assume that higher ordered categories include lower ordered categories. However, knowledge is usually assumed to consist of different types of knowledge without any clear ordering (e.g. de Jong & Ferguson-Hessler, 1996), except for the relationship between *factual knowledge* and *conceptual knowledge*. In subjects with a lot of procedural knowledge, like mathematics, conceptual knowledge may be seen as less concrete than procedural knowledge. For example from my own experience, it was no problem to learn how to derive functions following algorithms (*procedural knowledge*) but understanding the actual concept of 'derivate' (*conceptual knowledge*) came later and felt more abstract. Therefore, actually the ordering of the categories on a scale, not the categories, in the knowledge dimension in Bloom's revised taxonomy can actually be questioned. Moreover, none of the dominant alignment models order the content categories on a scale. Therefore, from this point of view Bloom's revised taxonomy should also treat the categories in the knowledge dimension as independent categories.

The second alignment criterion, concerning cognitive complexity, is called the cognitive process dimension in Bloom's revised taxonomy and is assumed by the authors to lie on a continuum of cognitive complexity without any overlaps between the categories. Both Porter's model (Porter & Smithson, 2001) and Webb's model (1997) also order their corresponding criteria along the same continuum. Therefore the ordering of the categories in the cognitive process dimension is trustworthy. I want, however, to make two remarks about categories and sub-categories in this dimension in Bloom's revised taxonomy. Firstly, the label of the second category *understand* is problematic, because of the great variety of different meanings that concept has both theoretically and in everyday life. The teachers participating in article II made comments about the unsuitability of using that label for the category. My suggestion is that this category should change its label to something less ambiguous. My second remark is a question whether the sub-category *executing*, in the category *apply*, actually is more cognitively complex than the category *understand*. In Bloom's revised taxonomy, *executing* is defined as "Applying a procedure to a familiar task" (Anderson & Krathwohl, 2001, p. 67). Applying a familiar algorithm to a familiar task in subjects with a large proportion of *procedural knowledge*, for example in mathematics, is comparable to recalling relevant knowledge from long-term memory. In such subject this is no more than just recalling the specific algorithm and inserting the present numbers. Therefore, the sub-category *execute* is, in my opinion, as cognitively complex as the sub-category *recalling* in the category *remember* and therefore *execute* should logically be placed in that category.

Standards today express expected knowledge and skills as a combination of content and a cognitive process, i.e. they state both *what* students should know and *how* they should use this what. The content is commonly defined with one type of expressions and the cognitive process with another type of expressions. Nouns in standards are commonly connected to the content, for example "the concept of changing coefficients and derivates" or "the role of chemistry in society". A verb is the most common expression for a cognitive

process, for example “interpret”, “formulate” and “evaluate”. This indicates that there are two independent dimensions in standards. In Bloom’s revised taxonomy, the knowledge dimension is used for analyzing content, while the cognitive process dimension is used for analyzing, as the label implies, cognitive processes. Thereby, it can be assumed that the two dimensions in Bloom’s revised taxonomy are independent. There is, however, a relation between the two dimensions. Content without a connected cognitive process is as meaningless as a cognitive process without content. For example, a standard that is expressed as knowledge about World War II, tells nothing about how the students should be able to use this knowledge. Should the student understand, give examples, explain or analyze the war? On the other hand, a standard that states that students should be able to understand and analyze, tells nothing about what the students should understand and analyze. For this reason there is a relation between content and cognitive processes.

The use of Bloom’s revised taxonomy in alignment analyses is only one of four purposes of the taxonomy, while the most dominant alignment models (Webb’s model, Porter’s model and the Achieve model) are developed for just one purpose, i.e. to analyze alignment. Roach et al. (2008) emphasize the need to develop tools that can be used both for alignment analyses and for increasing alignment among the components of an educational system in the daily practices. With the many-sided purposes of Bloom’s revised taxonomy, this taxonomy has the potential to be such a tool Roach et al. ask for.

6.3. High degree of alignment in the first place – advantages and risks

In standards-based educational systems alignment among standards, teaching and assessments is accepted as a fundament (Smith & O’Day, 1990) and a high degree of alignment is expected to give several advantages. It may, for example, improve the students’ learning (Biggs, 2003), make the educational system more effective (Herman et al., 2007) and lead to more accurate accountability decisions (Haertel & Herman, 2005). In validation of

interpretations of assessment results, alignment between standards and assessment is also important. Therefore, a high degree of alignment between standards and assessment should be in focus while constructing an assessment and also when assessment results are reported and used for different purposes. However, to allow a high degree of alignment between standards and assessments to come in the first place in education and to use all means to reach as high degree as possible without considering the consequences for the whole educational system may be risky. Several factors, besides the individually items in the assessment, influence the degree of alignment between standards and assessments and can cause problems. These factors are either concern the standards or the assessment.

The standards define both the size and the substance of the domain of what the students should know and be able to do. Two factors influence the size of this domain, namely the number of standards and how general the standards are. If there are a small number of specific standards, then the domain will be small and a high degree of alignment can easily be achieved by one assessment. In that case, a small number of assessment items can cover the whole domain. On the other hand if there are a large number of general standards, then the domain will be large and a low degree of alignment is expected. A small number of very general standards or a very large number of specific standards can also give a large domain. The items in an assessment can then hardly cover such a large domain and thereby the items will only be a sample of all items assessing a sample of the standards. The level of alignment will then be low.

The assessment can also influence the level of alignment, by the specifications for the assessment and by limitations in time, resources and assessment formats. In the design of an assessment, specifications about what and how to assess are stated before the construction of items. These specifications specify which standards to assess, thereby defining the domain to be assessed. If all standards are included in the assessment specifications, then it is possible to attain perfect alignment between standards and an assessment. If, on the other hand, only a sample of all standards is included

in the specifications, then the assessment domain will only be a subset of the domain defined by all standards and thereby limit the highest possible degree of alignment between all standards and the assessment. The highest possible degree of alignment will then be the quota between the number of standards in the assessment specifications and the total number of all standards. Limitations in time, resources and assessment formats may also influence the highest possible degree of alignment for an assessment. In education, there are normally limitations in time and resources for assessments, because teaching also requires time and resources. Limitations in time and resources lower the number of possible occasions for assessments and the number of items in each assessment and thereby limit the degree of alignment. Different types of standards require different formats of assessments and items and some kinds of standards are difficult to assess at all. Such limitations in assessment format may also limit the proportion of assessed standards and thereby limit the highest possible degree of alignment.

The standards are as vague and general today as the ones Popham found in 1997 and such standards define a large domain of knowledge and skills. Thereby, only a fairly low degree of alignment between standards and assessments may be expected. One way to enhance the degree of alignment is to increase the domain that is assessed and another way is to limit the domain defined by the standards. There are, however, several risks associated with both of these approaches.

A large domain of expected knowledge and skills will cover a wide range of content and cognitive processes. One way to obtain a higher degree of alignment is to let a set of assessments cover the whole domain, either by having extended assessments or by letting one or a couple of items cover each standard. There are risks associated with both these approaches. To have extended assessments on a sound psychometrics basis will take a lot of time, which will reduce the time for teaching. There is then a risk that the alignment between standards and teaching will drop considerably and that the students will receive less time for getting an opportunity to attain the

standards and probably learn less and perform more poorly on the assessment. The second approach, i.e. limit the number of assessment items to only one or a few per standard, influences negatively the reliability and the validity of interpretations of assessment scores. Assessment results are commonly used for generalizing performance on a specific assessment to performance on the domain defined by the assessment specifications (Kane, 2006). However, to generalize from one or just a couple of items is problematic because the sampling error will be too large according to Kane and the interpretations of students' performance on the domain will be psychometrically weak. The trustworthiness of using results from such assessments will be low.

The second way to obtain a higher degree of alignment is to limit the standards. Popham, among others, has expressed a worry that the standards today are too vague and general and define too large a domain to be able to teach and assess all standards appropriately. Therefore a reduction of the domain defined by standards, i.e. by having fewer and/or less general standards, can be a sound reform. However, a limitation of the standards can also be harmful to an educational system, if the standards are limited too much. One risk is that assessment specifications will constitute the standards and thereby emphasize easily assessable knowledge and skills. Many assessment items today assess processes with lower cognitive complexity (Resnick et al, 2003-2004) and therefore such standards will be at the lower end of the scale of cognitive complexity. This would counteract the intention of the standards-based reform to have challenging standards (Fuhrman, 2001). Another risk associated with limitations of standards is that more specified standards also specify teaching models and learning materials which is also against the intentions of standards-based education.

In summary, a high degree of alignment is important in standards-based educational systems. However, with general and vague standards it is difficult to construct perfect aligned assessments on a sound psychometric basis. To strive for a high degree of alignment with all means may influence both standards and assessments negatively, as well as the students' learning.

7. Future studies

Besides the issues discussed above, this thesis also has limitations. Based on these limitations, further studies are suggested in this chapter.

The purpose of this thesis was to evaluate the usefulness of models for analyzing alignment between standards and assessment, with main emphasis on Bloom's revised taxonomy. The evaluated models, especially Bloom's revised taxonomy, the Angoff method and the borderline-group method, were all found to be useful. The studies reported in the five articles were limited to 1) the usefulness of the evaluated models, not the construction of each model; 2) chemistry and mathematics; 3) alignment between standards and assessments; and 4) already existing assessments. This chapter will start with these limitations and suggest further studies.

Bloom's revised taxonomy was found to be a useful model for analyses of alignment, but the construction of the model is also worth studying. The ordering of the categories in the taxonomy has to be empirically and theoretically verified, because the ordering is a human construct not given by nature. For example in the discussion, the placing of the sub-category, *execute*, in the category *apply* instead of the category *remember* in the cognitive process dimension was questioned. Bloom's revised taxonomy is a revision of Bloom's original taxonomy and the cumulative hierarchical structure of the original taxonomy has been investigated in numerous empirical studies (Kreitzer & Madaus, 1994). These studies found that the three cognitively most complex categories in Bloom's original taxonomy were not in an appropriate order and therefore the order of two of the comparable categories in Bloom's revised taxonomy was different from that in Bloom's original taxonomy. However, if this order of categories in cognitive complexity in Bloom's revised taxonomy really is appropriate has to be verified.

The most recommended index for inter- and intra-judge consistency with data on nominal level is kappa, because it takes chance into account (Watkins & Pacheco, 2000). However, this index is only useful if one unit is

placed in only one category, not when one unit can be placed in more than one category, as was the case in articles II-IV. There is a need to develop an index for situations when multi-categorizations are allowed.

In the validation of the borderline-group procedure in article V, a positive relation was found between the performance of the borderline-examinees and the performance of the whole teaching group that those borderline-examinees belonged to and this is in accordance with results presented by Livingstone and Zieky (1989). However, this result raises the question whether judges when deriving the cut-score in the Angoff method, and in other standard-setting methods in which conceptualization of borderline-examinees is important, are also influenced by the performance level of the latest group of students they have taught. For the trustworthiness of derived cut-scores, this is important to investigate.

The evaluation of the usefulness of Bloom's revised taxonomy was limited to mathematics and chemistry, while the taxonomy is developed for general use in educational systems. To be able to recommend the taxonomy for all kinds of subjects, comparable studies have to be performed in other subjects. A first step would be to investigate the usefulness in subjects that have large-scale, standardized assessments. This kind of studies may also verify whether the large proportion of multi-categorized standards and assessment items in mathematics can be explained by the dual character of mathematics or whether Bloom's revised taxonomy needs to be revised again.

This thesis is limited to analyses of alignment between standards and assessments, but alignment between standards and teaching is also important. According to the authors of Bloom's revised taxonomy (Anderson & Krathwohl, 2001), the taxonomy is also aimed at being an aid in planning and delivering teaching aligned with the standards, and hence might also be useful for analyses of alignment between standards and teaching. Further studies are needed to investigate the usefulness for this purpose.

Roach et al. (2008) expressed the need to develop an alignment model that can be useful for both analysis of alignment after teaching or after the design of an assessment and "to engage in ongoing practices that result in increased alignment among the components" (p. 173). Bloom's revised taxonomy is developed to be such a model (Anderson & Krathwohl, 2001) and was, in this thesis, found to be useful for analysis after the design of an assessment. However, to be able to conclude whether this taxonomy really can increase the alignment among the components in an educational system by using it in the ongoing practices, further investigations are needed.

8. Svensk sammanfattning

Ett utbildningssystem består vanligtvis av tre delar: läro- och kursplaner, undervisning samt bedömning. I läroplanen och i kursplaner finns mål och betygskriterier som definierar vad eleverna förväntas kunna efter en utbildning och dessa intar en central roll i vårt svenska mål- och kriterierelaterade utbildningssystem. Syftet med undervisningen är att ge alla elever möjlighet att uppnå målen, medan bedömningen ska utvärdera om och hur väl eleverna uppnått målen.

I målrelaterade utbildningssystem betonas vikten av att delarna är samstämmiga (t.ex. Fuhrman, 2001; Biggs, 2003) och detta innebär att undervisningen läggs upp så att eleverna ges möjlighet att uppnå målen och att bedömningen utvärderar om och hur väl eleverna har uppnått målen. En hög grad av samstämmighet mellan utbildningssystemets delar antas vara viktig för elevernas inlärning (Biggs, 2003), för utbildningssystemets effektivitet (Webb, 1997), för utvärderingar av utbildningssystemet (Herman, Webb, & Zuniga, 2007), för beslut om resursfördelning baserade på skolresultat (La Marca, 2001; Koretz & Hamilton, 2006), för validering av resultat från bedömningar (Rothman, 2003) och för information till elever, föräldrar, politiker och allmänheten (Herman m.fl., 2007). Samstämmighet mellan mål och bedömning i ett utbildningssystem antas ge samma fördelar som samstämmighet mellan alla tre delarna och dessutom kan fungera som

en form av bevis i valideringen av användningen av bedömningsresultat (La Marca, 2001; Rothman, 2003). En av de första definitionerna av samstämmighet inom utbildningsfältet gavs av Webb 1997, då han definierade samstämmighet som den grad som målen och bedömningen överensstämmer med varandra och samverkar för att utbildningssystemet ska få eleverna att lära sig det som de förväntas lära sig.

Det bästa sättet att uppnå så hög grad av samstämmighet som möjligt är att utgå från läroplanen och respektive kursplan i uppläggningen av undervisningen och i skapandet av bedömningstillfällen (Baker, 2004; Martineau, Paek, Keene & Hirsch, 2007). Det är dock vanligare att graden av samstämmighet analyseras efter att undervisningen är genomförd och bedömningarna är skapade. För samstämmighetsanalyser finns det ett antal specifikt utformade modeller, t.ex. har Webb (1997), Porter & Smithson (2001) och Rothman, Slattery, Vranek & Resnick (2002) utvecklat sådana modeller. De flesta av dessa modeller är utvecklade för att göra analyser av samstämmighet mellan mål och bedömningar (Bhola, Impara & Buckendahl, 2003), men den modell som Porter och Smithson (2001) (härefter kallad Porters modell) har utvecklat är även tänkt för att göra jämförelser med undervisning.

Den här avhandlingen fokuserar modellers användbarhet för analyser av samstämmighet mellan två av utbildningssystemets delar, nämligen kursplaner och bedömningar. Det finns två huvudsakliga anledningar till att utesluta undervisning från avhandlingen. För det första antas undervisningen variera mellan elevgrupper, lärare och skolor eftersom den ska läggas upp efter elevernas förutsättningar (Utbildningsdepartementet, 1994) och eftersom målen inte ska styra val av undervisningsmetod och material (SOU 2007:28). För det andra intar det nationella provsystemet en central roll i styrningen av skolor, eftersom provresultat kan användas för att utvärdera om och i vilken grad eleverna har uppnått målen. Att betygen ska vara likvärdigt satta över hela Sverige innebär att även bedömningen som sker i skolorna ska vara likvärdig oavsett var i Sverige bedömningen görs.

I en samstämmighetsanalys kategoriseras varje mål och bedömningsuppgift efter innehåll och vanligtvis även efter kognitiv komplexitet med en modell (förklaras närmare i artikel I). Att varje uppgift är samstämmig med minst ett mål är en förutsättning i konstruktionen av en bedömning, men att alla uppgifter är samstämmiga med mål innebär inte med självklarhet att alla mål bedöms. Därför är det intressantare att analysera till vilken grad alla uppgifter i en bedömning är samstämmiga med alla mål för just den kursen. Graden av samstämmighet bestäms av hur uppgifterna fördelas på kategorierna jämfört med hur målen fördelas. Jämförelsen mellan spridningen av uppgifterna och målen är intressant, men ännu intressantare är balansen mellan fördelningen av uppgifterna och fördelningen av målen i verktygets kategorier (Bhola m.fl., 2003).

Förutom de modeller som är utvecklade för samstämmighetsanalyser, t.ex. Webbs och Porters modeller, finns det ett antal kategoriseringsmodeller som också kan vara användbara för sådana analyser. Ett villkor för dessa modeller är att de kan kategorisera innehåll och kognitiv komplexitet i både uppgifter och mål. Exempel på sådana kategoriseringsmodeller är Guilfords taxonomi (1967), Blooms reviderade taxonomi (Anderson & Krathwohl, 2001), och Marzanos nya taxonomi (Marzano & Kendall, 2007).

Målen i dagens utbildningssystem anses ofta vara vaga och generella (Popham, 1997; SOU 2007:28) och därför kan tolkningar av målen i undervisningen, för utvecklandet av bedömningar och i samstämmighets-analyser spreta ganska mycket. En statlig utredning (SOU 2008:27) konstaterade att lärare i den svenska gymnasieskolan verkar tolka målen på olika sätt. I samstämmighetsanalyser är det dock viktigt att deltagarnas tolkningar av målen överensstämmer.

Artikel I i denna avhandling jämförde olika modeller som kan vara användbara för samstämmighetsanalyser. Mål och uppgifter beskriver både ett innehåll och en process som anger hur detta innehåll ska användas. Olika processer ställer olika kognitiva krav och är därmed mer eller mindre kognitivt komplexa. Därför skulle en användbar modell kunna kategorisera ett mål eller en uppgift utifrån både innehåll och kognitiv komplexitet.

Kravet var också att kategorierna för kognitiv komplexitet ligger på en skala. Eftersom de utvalda modellerna i denna artikel prövades empiriskt på en kursplan och ett prov i kemi krävdes det att en användbar modell kunde kategorisera det ämnet. Två modeller, Blooms reviderade taxonomi (Anderson & Krathwohl, 2001) och Porters modell (se t.ex. Porter, 2002), ansågs uppfylla dessa krav och deras användbarhet jämfördes empiriskt. Jämförelsen baserades dels på interbedömaröverensstämmelse, dvs. graden av överensstämmelse mellan bedömarna i kategoriseringen av mål och uppgifter, och dels på Hauensteins (1998) fem regler. Enligt dessa regler är en kategoriseringsmodell användbar om den: 1) är applicerbar; 2) är totalt inkluderande; 3) har entydigt uteslutande kategorier; 4) har kategorier som är konsekvent ordnade efter en princip; 5) använder termer i kategorier och underkategorier vilka är representativa för området. Resultatet av den empiriska jämförelsen blev att Blooms reviderade taxonomi, med sina generella innehållskategorier, var mer användbar än Porters modell, med sina ämnesspecifika innehållskategorier.

Blooms reviderade taxonomi (Anderson & Krathwohl, 2001) är en vidareutveckling av Blooms taxonomi från 1956 och dess syfte är att utveckla mål, undervisning och bedömning samt stimulera samstämmighet mellan dessa komponenter i ett utbildningssystem. Denna nya taxonomi har tillämpats på olika utbildningar (t.ex. Su, Osisek, & Starnes, 2004; Hanna, 2007; Pickard, 2007), men utvärderingar av dess användbarhet i samstämmighetsanalyser lyser med sin frånvaro.

Blooms reviderade taxonomi består av två dimensioner: en kunskapsdimension och en dimension för kognitiva processer. Kategorierna i kunskapsdimensionen definieras av olika sorters kunskaper: *faktakunskap, begreppskunskap, procedurkunskap* och *metakognitiv kunskap*, och dessa kategorier analyserar innehållet i mål, undervisning och bedömning. Enligt taxonomins författare ligger dessa kategorier på en skala från konkret i *faktakunskap* till abstrakt i *metakognitiv kunskap*, men de medger att det finns en viss överlappning mellan *begrepps-* och *procedurkunskap*. Dimensionen för kognitiva processer fokuserar hur kunskapen används och

består av sex kategorier: *minnas, förstå, tillämpa, analysera, värdera* och *skapa*. Dessa kategorier ligger på en skala från låg grad av kognitiv komplexitet i *minnas* till hög grad av kognitiv komplexitet i *skapa*. Varje kategori i de två dimensionerna har två till sju underkategorier (se den engelska versionen i figur 5 i bilagan) och dessa underkategorier definierar respektive kategori. I kategorisering av t.ex. mål används dessa underkategorier för att placera ett mål i en viss kategori.

Blooms reviderade taxonomi erbjuder en två-dimensionell taxonomitabell med 24 celler (se figur 5 i bilagan). Raderna i tabellen representerar de fyra kategorierna i kunskapsdimensionen och kolumnerna de sex kategorierna i dimensionen för kognitiva processer. När till exempel ett mål placeras in i taxonomitabellen, kategoriseras det först i respektive dimension och därefter placeras det i den korresponderande cellen i taxonomitabellen.

I artiklarna II-IV utvärderades användbarheten av Blooms reviderade taxonomi ytterligare, utifrån de tre första av Hauensteins regler (1998) och interbedömaröverensstämmelse, samt i artikel II även intrabedömar-överensstämmelse (graden av överensstämmelse i kategoriseringen av mål och uppgifter mellan två tillfällen för samma bedömare).

I artikel II redovisas en studie där två grupper med olika sammansättningar fick kategorisera mål i en kursplan i matematik för gymnasieskolan med Blooms reviderade taxonomi. Den ena gruppen bestod av fyra matematiklärare och den andra av fyra provansvariga för nationella prov i matematik. Dessa två grupper fick kategorisera samma kursplan vid två tillfällen var. I denna studie jämfördes användbarheten och inter- och intrabedömaröversstämmelsen mellan dessa två grupper. Resultatet visade att både inter- och intrabedömaröverensstämmelsen var högre för gruppen provansvariga än för gruppen lärare. En annan slutsats från studien var att Blooms reviderade taxonomi var i stort användbar. Enligt Hauensteins tredje regel ska en användbar kategoriseringsmodell ha entydigt uteslutande kategorier. I denna studie var det dock tillåtet att placera ett mål i flera kategorier samtidigt, dvs. multikategorisera, och denna möjlighet utnyttjades, vilket försämrade användbarheten till viss del. I denna artikel

lanserades tre möjliga förklaringar till multikategoriseringen: 1) målen var vaga och generella; 2) det fanns brister i definitionerna av kategorierna i Blooms reviderade taxonomi, eller 3) det fanns brister i instruktionerna till dem som kategoriserade.

I artikel III utvärderades användbarheten av Blooms reviderade taxonomi för både mål och provuppgifter i matematik utifrån samma kriterier som i artikel II men vid bara ett tillfälle. Denna artikel redovisar en studie där en grupp bestående av fem provansvariga kategoriserade en kursplan och ett nationellt kursprov i matematik för samma kurs för gymnasieskolan. Ett resultat av denna studie var att Blooms reviderade taxonomi i stort var användbar för både mål och provuppgifter. Andelen multikategoriseringar var lägre för uppgifter än för mål, men även provuppgifter placerades i flera kategorier samtidigt. Studien redovisade även interbedömaröverens-stämmelse, vilken var något högre för provuppgifterna än för målen.

I artikel IV redovisas en studie som försökte utröna om den stora andelen multikategoriserade mål i artikel II kunde bero på vaga och generella mål. I denna studie delades målen i en kursplan i matematik upp i delmål enligt principen i Blooms reviderade taxonomi, vilken säger att ett mål bör bestå av ett substantiv och ett verb. En grupp på fem provansvariga fick först kategorisera målen utan uppdelning och därefter kategorisera delmålen. En hypotes formulerades i artikeln: uppdelningen av målen skulle ge en lägre andel multikategoriserade mål och en högre interbedömaröverens-stämmelse. Resultaten visar att andelen multikategoriseringar sjönk med uppdelningen av målen, men att även en stor del av delmålen var multikategoriserade. Däremot fanns det ingen ökning i interbedömar-överensstämmelsen. Slutsatsen av studien blev således att vaga och generella mål inte är en trolig förklaring till den stora andelen multikategoriserade mål. Förklaringar som rör definitionerna av kategorierna i Blooms reviderade taxonomi och instruktionerna till dem som kategoriserade är alltså mer troliga. I denna studie undersöktes även om förklaringen att matematikens natur att sammanfläta begreppskunskap och procedur-kunskap kan ha lett till den stora andelen

multikategoriseringar. Studien visade att de flesta multikategoriserade mål var kategoriserade både som *begreppskunskap* och *procedurkunskap*, vilket ger stöd till den sistnämnda förklaringen.

Artikel V fokuserade en annan dimension av samstämmighet mellan en kursplan och en bedömning. I en kursplan finns det förutom mål även betygskriterier, vilka ska ligga till grund för att bedöma hur väl eleverna har uppnått målen. Dessa betygskriterier ska konkretiseras på, och därmed göras samstämmiga med, en specifik bedömning genom att fastställa vilken prestation på bedömningen som är precis på gränsen mellan två betygssteg. Detta sker genom en s.k. kravgränssättning. I denna artikel jämfördes validiteten för två kravgränssättningsmetoder, Angoffs kravgränssättnings-metod och den s.k. borderline-group metoden, när dessa tillämpades på ett nationellt kursprov i matematik för gymnasieskolan. Båda metoderna visade sig ge valida kravgränser. För borderline-group metoden indikerades ett behov av att ha representativa undervisningsgrupper för att få valida kravgränser.

När utfallet för studierna I-IV jämförs går det att konstatera att Blooms reviderade taxonomi var användbar för analyser av samstämmighet mellan mål och bedömningar, både i kemi och i matematik. Dock blev en stor andel av målen i matematik multikategoriserade och detta försämrade användbarheten av Blooms reviderade taxonomi i just matematik. I kemi blev endast ett mål placerat i fler än en kategori, trots att bedömarna i alla fyra studierna var instruerade att de fick multikategorisera mål och uppgifter. Detta indikerar att det kan vara ämnet matematik som gett upphov till den stora andelen multikategoriseringar, inte själva taxonomin eller instruktionen till bedömarna. Kemi såväl matematik innehåller mycket begrepps- och procedurkunskap, men dessa två typer av kunskaper verkar vara sammanflätade i matematik men lättare åtskiljbara i kemi. Det behövs fler studier i andra ämnen än kemi och matematik för att studera taxonomins användbarhet i allmänhet.

Enligt Roach, Niebling och Kurz (2008) finns det ett behov av att utveckla en modell som kan användas både för samstämmighetsanalyser och för att stimulera en hög grad av samstämmighet i den pågående verksamheten.

Blooms reviderade taxonomi har en potential att vara en sådan modell. Denna avhandling har i sina studier av modellers användbarhet för analyser av samstämmighet mellan kursplaner och bedömningar visat att Blooms reviderade taxonomi fungerar tillfredsställande. Vidare studier är dock nödvändiga för att studera om taxonomin även kan stimulera samstämmighet i en pågående verksamhet.

References

American Educational Research Association, American Psychological Association & National Council for Measurement in Education (AERA, APA & NCME). (1999). *Standards for educational and psychological testing*. Washington: AERA.

Ananda, S. (2003). Achieving alignment. *Leadership, 33*(1), 18-21, 37.

Anderson, L. W. (1994). Research on teaching and teacher education. In L. W. Anderson, & L. A. Sosniak (Eds.), *Bloom's taxonomy: A forty-year retrospective* (pp. 126-145). Ninety-third Yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press.

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

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing. A revision of Bloom's taxonomy of educational objectives*. New York: Addison Wesley Longman.

Anderson, L. W., & Sosniak, L. A. (Eds.) (1994). *Bloom's taxonomy: A forty-year retrospective*. Ninety-third Yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press.

Bailey, K. D. (1994). *Typologies and taxonomies. An introduction to classification techniques* (Sage university paper series on Quantitative applications in the social sciences, series no. 07-102). Thousand Oaks: Sage.

Baker, E. L. (2004). *Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform* (CSE report 645). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Bennett, J. (2001). Practical work of the upper high school level: The evaluation of a new model of assessment. *International Journal of Science Education, 23*(1), 97-110.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning test with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21-29.

Biggs, J. (2003). *Teaching for quality learning at university*. Glasgow: The Society for Research into Higher Education & Open University Press.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H. & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I: Cognitive domain*. New York: David McKay.

Bogan, Y. K. H., & Porter, R. C. (2005). On the ball with higher-order thinking. *Teaching Pre-K-8, 36*(3), 46-47.

Booker, M. (2007). A roof without walls: Benjamin Bloom's taxonomy and the misdirection of American education. *Academic Questions, 20*(4), 347-355.

Castle, A. (2003). Demonstrating critical evaluation skills using Bloom's taxonomy. *International Journal of Therapy & Rehabilitation, 10*(8), 369-373.

Chung, B. M. (1994). The taxonomy in the Republic of Korea. In L. W. Anderson, & L. A. Sosniak (Eds.), *Bloom's taxonomy: A forty-year retrospective* (pp. 164-173). Ninety-third Yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press.

Chyung, S.-Y., & Stepich, D. (2003). Applying the "congruence" principle of Bloom's taxonomy to design online instruction. *Quarterly Review of Distance Education, 4*(3), 317-330.

Connelly, F. M., & Lantz, O. C. (1991). Definitions of curriculum: An introduction. In A. Lewy (Ed.), *The international encyclopedia of curriculum* (pp. 15-18). Oxford: Pergamon.

Dagget, W. R. (2000). Mowing from standards to instructional practice. *NASSP Bulletin, 84*(66), 66-72.

de Jong, T., & Ferguson-Hessler, M. G. M. (1996). Types and qualities of knowledge. *Edcuational Psychologist, 31*(2), 105-113.

Donn, G. (1994). Feminist approaches and the curriculum. In T. Husén & T. N. Postlethwaite (Eds.), *The international encyclopedia of education. Volume 3*. (pp. 2287-2292). Oxford: Pergamon.

Eash, M. J. (1991). Syllabus. In A. Lewy (Ed.), *The international encyclopedia of curriculum* (pp. 71-73). Oxford: Pergamon.

English, F. W. (2000). *Deciding what to teach and test: Developing, aligning, and auditing the curriculum*. Thousand Oaks: Corwin Press.

Eraut, M. R. (1991). Defining educational objectives. In A. Lewey (Ed.), *The international encyclopedia of curriculum* (pp. 306-317). Oxford: Pergamon.

Farenga, S. J., Joyce, B. A., & Ness, D. (2002). Reaching the zone of optimal learning: The alignment of curriculum, instruction, and assessment, In R. W. Bybees (Ed.), *Learning science and the science of learning* (pp. 51-62). Arlington: NSTA press.

Fairbrother, R. W. (1975). The reliability of teachers' judgement of the abilities being tested by multiple choice items. *Educational Research, 17*(3), 202-210.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382.

Fuhrman, S. H. (Ed.) (2001). *From the capitol to the classroom: Standards-based reform in the States*. Yearbook of the National Society for the Study of Education. Part II. Chicago: The University of Chicago Press.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science, 5*(1), 13-34.

Granello, D. H. (2001). Promoting cognitive complexity in graduate written work: Using Bloom's taxonomy as a pedagogical tool to improve literature reviews. *Counselor Education & Supervision, 40*(4), 292-306.

Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Guskey, T. R. (2007). Closing achievement gaps: Revisiting Benjamin S. Bloom's "Learning for mastery". *Journal of Advance Academics, 19*(1), 8-31.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity argument for accountability testing. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (pp. 1-34). The 104th yearbook of the National Society for the Study of Education. Part 2. Malden: Blackwell Synergy.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 433-470). Westport: American Council on Education & Praeger Publishers.

Hauenstein, A. D. (1998). *A conceptual framework for educational objectives. A holistic approach to traditional taxonomies*. Lanham: University Press of America.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments: A case study. *Applied Measurement in Education, 20*(1), 101-126.

Holmes, P. (2002). Assessment: New ways of pupil evaluation using real data. *Teaching statistics, 24*(3), 87-89.

Howson, G. (1995). *Mathematics textbooks: A comparative study of grade 8 texts*. Vancouver: Pacific Educational Press.

Jongsma, K. S. (2007). Standards: Powerful tools of unnecessary provocations? *The Reading Teacher, 46*(4), 340-341.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah: Lawrence Erlbaum Associates.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17-64). Westport: American Council on Education & Praeger Publications.

Kastberg, S. E. (2003). Using Bloom's taxonomy as a framework for classroom assessment. *The Mathematics Teacher, 96*(6), 402-405.

Kjellström, K., & Pettersson, A. (2005). The curriculum's view of knowledge transferred to national test in mathematics in Sweden. *ZDM, 37*(4), 308-316.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (pp. 531-578). Westport: American Council on Education & Praeger Publishers.

Kreitzer, A. E., & Madaus, G. F. (1994). Empirical investigations of the hierarchical structure of the taxonomy, In L. A. Anderson & L. A. Sosniak (Eds.), *Bloom's taxonomy: A fourty-year retrospective*. Ninety-third yearbook of the National Society for the Study of Education part II (pp. 64-81). Chicago: The University of Chicago Press.

Kropp, R. P., & Stoker, H. W. (1966). *The construction and validation of tests of the cognitive processes as described in the taxonomy of educational objectives*. Tallahassee: Florida State University, Institute of Human Learning and Department of Educational Research and Testing.

La Marca, P. M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation, 7*(21).

La Marca, P. M., Redfield, D., Winter, P. C., Bailey, A., & Hansche, D. (2000). *State standards and state assessment systems: A guide to alignment*. Washington: Council of Chief State School Officers.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lesson planning in the classroom. (2007). *Techniques: Connecting Education & Careers, 82*(9), 8-9.

Lewy, A., & Báthory, Z. (1994). The taxonomy of educational objectives in Continental Europe, the Mediterranean, and the Middle East. In L. W. Anderson, & L. A. Sosniak (Eds.), *Bloom's taxonomy: A forty-year retrospective* (pp. 146-163). Ninety-third Yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press.

Linn, R. L. (1994). Performance assessment: Policy, promises and technical measurement standards. *Educational Researcher, 23*(9), 4-14.

Linn, R. L. (2006). Issues in the design of accountability systems. In J. L. Herman, & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (pp. 78-98). The 104th yearbook of the National Society for the Study of Education, part 2. Malden: Blackwell Publishing.

Linn, R. L., & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron, & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (pp. 84-103). Ninety-fifth yearbook of the National Society for the Study of Education. Chicago: The University of Chicago Press.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational researcher, 36*(8), 437-448.

Livingstone, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education, 2*(2), 121-141.

Luft, P., Brown, C. M., & Slutherin, L. J. (2007). Are you and your students bored with the benchmarks? Sinking under the standards? Then transform your teaching through transistion. *Teaching Exceptional Children, 39*(6), 39-46.

Manaris, B., Wainer, M., Kirkpatrick, A. E., Stavley, R. H., Shamon, C., Leventhal, L., Barnes, J., Wright, J., Schafer, B., & Sanders, D. (2007). Implementations of the CC'01 human-computer interaction guidelines using Bloom's taxonomy. *Computer Science Education, 17*(1), 21-57.

Martineau, J., Paek, P., Keene, J., & Hirsch, T. (2007). Integrated, comprehensive alignment as a foundation for measuring student progress. *Educational Measurement: Issues and Practice, 26*(1), 28-35.

Marzano, R. J. (2001). *Designing a new taxonomy of educational objectives*. Thousand Oaks: Corwin Press.

Marzano, R. J., & Kendall, J. S. (2007). *The new taxonomy of educational objectives*. Thousand Oaks, California: Corwin Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp.13-103). New York: Macmillan.

Miller, C. M. L., & Parlett, M. (1974). *Up to the mark. A study of the examination game*. London: Society for Research into Higher Education.

Ministry of Education and Science in Sweden. (2000). *Education Act*. Retrieved July 1 2008 from http://www.sweden.gov.se/content/1/c6/02/15/38/1532b277.pdf .

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzales, E. J., Chrostowski, S. J., & O'Connor, K. M. (2001). *TIMSS assessment frameworks and specifications 2003*. Chestnut Hill: International Association for the Evaluation of Educational Achievement.

Newton, L. D. (2000). *Meeting the standards in primary science: A guide to the ITT NC*. London: RoutledgeFalmer.

OECD. (1999). *Measuring student knowledge and skills: A new framework for assessment*. Paris: OECD.

Oliver, D., Dobele, T., Greber, M., & Roberts, T. (2004). Comparing course assessments: When lower is higher and higher, lower. *Computer Science Education, 14*(4), 321-341.

Paziotopoulus, A., & Kroll, M. (2004). Hooked on thinking. *Reading Teacher, 57*(7), 672-677.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into Practice, 41*(4), 219-225.

Popham, W. J. (1993). *Educational evaluation*. Boston: Allyn and Bacon.

Popham, W. J. (1997). The standards movement and the emperor's new clothes. *NASSP Bulletin, 81*(21), 21-25.

Popham, W. J. (2000). Assessing mastery of wish-list content standards. *NASSP Bulletin, 84*(30), 30-36.

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment*. Alexandria: Association for Supervision and Curriculum Development.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Porter, A. C., & Smithson, J. L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S. H. Fuhrman (Ed.), *From the Capitol to the classroom. Standards-based reform in the States* (pp. 60-80). Chicago: National Society for the Study of Education, University of Chicago press.

Porter, A. C., Smithson, J., Blank, R., & Zeidner, T. (2007). Alignment as a teacher variable. *Applied Measurement in Education, 20*(1), 27-51.

Pungente, M. D., & Badger, R. A. (2003). Teaching introductory organic chemistry: 'Blooming' beyond simple taxonomy. *Journal of Chemical Education, 80*(7), 779-784.

Ravitch, D. (1992). National standards and curriculum reform: A view from the Department of Education. *NASSP Bulletin, 76*(24), 24-29.

Reis-Jorge, J. M. (2005). Developing teachers' knowledge and skills as researchers: a conceptual framework. *Asia-Pacific Journal of Teacher Education, 33*(3), 303-319.

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003-2004). Benchmarking and alignment of standards and testing. *Educational Assessment, 9*(1-2), 1-27.

Rittler-Johnson, B., & Alibali, M. W, (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other one? *Journal of Educational Psychology, 91*(1), 175-189.

Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools, 45*(2), 158-176.

Robitaille, D. F., Schmidt, W. H., Raizen, S. McKnight, C., Britton, E., & Nicol, C. (1993). *TIMSS monograph no. 1: Curriculum frameworks for mathematics and science*. Vancouver: Pacific Educational Press.

Rothman, R. (2003). *Imperfect matches: The alignment of standards and tests*. Paper commissioned by the Committee on Test Design for K-12 Science Achievement, March 2003.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* (CSE Technical Report 566). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Dodrecht: Kluwer Academic Publishers.

Seddon, G. M. (1978). The properties of Bloom's taxonomy of educational objectives for the cognitive domain. *Review of Educational Research, 48*(2), 303-323.

Seigerroth, U. (2003). *Att förstå och förändra systemutvecklings-verksamheter. En taxonomi för metautveckling* [To understand and to change system development organisations: A taxonomy for meta-development]. Doctoral dissertation. Linköping university.

Smith, M. S., & O'Day, J. (1990). Systematic school reform. In S. H. Furhman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233-267). The 1990 yearbook of the Politics of Education Association. London: The Falmer Press.

Sosniak, L. A. (1994a). Educational objectives: Use in curriculum development. In T. Husén, & T. N. Postlethwaite (Eds.), *The international encyclopedia of education. Volume 3* (pp. 1800-1804). Oxford: Pergamon.

Sosniak, L. A. (1994b). The taxonomy, curriculum, and their relations. L. W. Anderson, & L. A. Sosniak (Eds.), *Bloom's taxonomy: A forty-year retrospective* (pp. 103-125). Ninety-third Yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press.

SOU 2007:28. *Tydliga mål och kunskapskrav i grundskolan. Förslag till nytt mål- och uppföljningssystem.* [Clear goals and required knowledge in compulsory school. Suggestion to a new system of goals and follow-up]. Stockholm: Utbildningsdepartementet.

SOU 2008:27. *Framtidsvägen – en reformerad gymnasieskola.* [The future road – a reformed upper secondary school]. Stockholm: Utbildnings-departementet.

Squire, P. J. (2001). Cognitive levels of testing agricultural science in senior secondary schools in Botswana. *Education, 121*(3), 597-603.

Stephens, J-P, Vos, G. A., Stevens, E. M., & Moore, J. S. (2006). Test-retest repeatability of the Strain index. *Applied Ergonomics, 37*(3), 275-281.

Su, W. M., Osisek, P. J., & Starnes, B. (2004). Applying the revised Bloom's taxonomy to a medical-surgical nursing lesson. *Nurse Educator, 29*(3), 116-120.

Sugrue, B. (2002). *Problems with Bloom's taxonomy*. Retrieved June 23, 2008, from: http://www.performancexpress.org/0212/

Tyler, R. W. (1969). *Basic principles of curriculum and instruction*. Chicago: The University of Chicago Press.

Utbildningsdepartementet. (1994). *Läroplaner för det obligatoriska skolväsendet och de frivilliga skolformerna.* [Curricula for the Compulsory School System and for the Non-compulsory School System]. Stockholm: Utbildningsdepartementet.

Watkins, M. W., & Pacheco, M. (2000). Interobserver agreement in behavioural research: Importance and calculation. *Journal of Behavioral Education, 10*(4), 205-212.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research monograph, No. 6). Madison: National Institute for Science Education.

Webb, N. L. (2002). *An analysis of the alignment between mathematics standards and assessments for three states*. Paper presented at the annual meeting of the American Educational Research Association, April 1-5, in New Orleans, USA.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7-25.

Wiliam, D. (2000). Standards: What are they, what do they do and where do they live? In B. Moon, M. Ben-Peretz, & S. A. Brown (Eds.), *Routledge international companion to education* (pp. 351-363). London: Routledge.

Williams, A. B., Honghong, W., Burges, J. Chenghui, W., Yuxiu, G., & Yaim L. (2006). Effectiveness of an HIV/AIDS educational programme for Chinese nurses. *Journal of Advanced Nursing, 53*(6), 710-720.

Windh, C., & Gingell, J. (1999). *Key concepts in the philosophy of education*. London: Routledge.

Wood, K. L., & Linsey, J. S. (2007). Understanding the art of design: Tools for the next Edisionian innovators. In A. B. Markman, & B. H. Ross (Eds.), *Categories in use* (pp. 65-122). Amsterdam: Elsevier.

Wood, R., & Power, C. (1984). Have national assessments made us any wiser about `standards´? *Comperative Education, 20*(3), 307-321.

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Application of Bloom's revised taxonomy debunks the "MCAT myth". *Science, 319*(5862), 414-415.

Appendix

In Appendix, the use of Bloom's revised taxonomy for categorization of standards and assessment items as well as for alignment analyses is exemplified.

Bloom's revised taxonomy offers a taxonomy table with 24 cells (see Figure 5). To be able to place a standard or an assessment item in the taxonomy table, the noun and the verb in the standard or assessment item have to be identified. The noun represents content and is categorized by the categories in the knowledge dimension (the rows in the taxonomy table). The verb represents how the content should be used and is categorised by categories in the process dimension (the columns in the taxonomy table). The standard or the assessment item is then placed in the corresponding cell in the taxonomy table.

To exemplify how the taxonomy table can be used, the categorization of two examples of standards and two examples of assessment items in mathematics are presented in Figure 5 and motivated in the text below. In Figure 5, eight extra standards and ten extra assessment items are also placed to be able to show how to analyze the alignment between standards and assessment items. The examples of standards are labelled with an 'S' and a number, while the examples of assessment items are labelled with an 'A' and a number.

The first example, a standard, is formulated as follows:

S1. Pupils should be able to use mathematical models of different kinds, including those which build on the sum of a geometric progression.

The verb in this standard is 'use' and this verb can be found in the definition of the category *apply* in the cognitive process dimension. Therefore, this standard is categorized as *apply*. The noun in this standard is 'mathematical models' and a sub-category in the category *conceptual knowledge* in the knowledge dimension deals with models. Therefore, this standard is categorized as *conceptual knowledge*. The corresponding cell in the taxonomy table for this standard is *apply conceptual knowledge* (see Figure 5).

The Knowledge Dimension	The Cognitive Process Dimension												
	Remember		Understand						Apply		Analyze			Evaluate			Create		
	Recognising	Recalling	Interpreting	Exemplifying	Classifying	Summarising	Inferring	Comparing	Explaining	Executing	Implementing	Differentiating	Organizing	Attributing	Checking	Critiquing	Generating	Planning	Producing
Factual knowledge – Knowledge of terminology																			
Factual knowledge – Knowledge of specific details and elements																			
Conceptual knowledge – Knowledge of classifications and classifications			**S2 S3** *A2 A3*							**S1 S4 S5 S6**	*A5*				**S8 S9**		**S10**		
Conceptual knowledge – Knowledge of principles and generalizations																			
Conceptual knowledge – Knowledge of theories, models, and structues																			
Procedural knowledge – Knowledge of subject-specific skills and algorithms			**S2** *A4*							**S5 S6 S7** *A1 A6 A7 A8 A9 A10 A11*	*A12*						**S10**		
Procedural knowledge – Knowledge of subject-specific techniques and methods																			
Procedural knowledge – Knowledge of criteria for determining when to use appropriate procedures																			
Metacognitive knowledge – Strategic knowledge																			
Metacognitive knowledge – Knowledge about cognitive tasks																			
Metacognitive knowledge – Self-knowledge																			

Figure 5. Examples of standards (S1-S10) and assessment items (A1-A12) categorized with Bloom's revised taxonomy

The next example of standards is more complicated and is placed in two cells, i.e. is multi-categorized:

S2. The school in its teaching of mathematics should aim to ensure that pupils develop their ability to interpret a problem situation and formulate this in mathematical terms and symbols, as well as choose methods and aids in order to solve problems.

There are three verbs to consider: 'interpret', 'formulate' and 'choose' and all these verbs can be placed in the category *understand* in the cognitive process dimension. Interpreting is a sub-category in the category *understand*. To formulate can be compared to explain in mathematical terms and therefore is placed in the same category. To choose is to draw conclusions, i.e. infer, from the situation and is therefore also placed in the same category, i.e. *understand*. There are also two expressions of nouns to consider in the standard: 'a problem situation' and 'methods and aids'. A problem situation in mathematics deals with principles, generalizations, theories, models and structures, i.e. *conceptual knowledge*. However, knowledge of methods and aids are *procedural knowledge*. Because *conceptual knowledge* and *procedural knowledge* are assumed to be two separate types of knowledge, the standard is placed in both these categories. Therefore this standard is placed in two corresponding cells, namely *understand conceptual knowledge* and *understand procedural knowledge* (see Figure 5).

The third example is an assessment item:

A1. Differentiate $f(x) = x^3 - 6x$.

The verb in this assessment item is 'differentiate' and, in this case, this involves executing a procedure well-known for the student. Therefore this assessment item is categorized as *apply*. The noun in this assessment item is the presented function, but the function is implicitly understood as knowledge about a specific algorithm to solve this item with, i.e. *procedural knowledge*. Hence, this assessment item is placed in the cell *apply procedural knowledge* in the taxonomy table (see Figure 5).

The fourth and last explained example is also an assessment item:

A4. A certain geometric sum can be calculated by $\frac{4000 \cdot (1.03^5 - 1)}{1.03 - 1}$

Write down the terms of the geometric sum that can be calculated by the above expression.

The verb in this assessment item is 'write' and this verb neither explains really the task in this item nor hints at a sub-category in the cognitive process dimension. The students are supposed to explain the terms in the expression with words and thereby this assessment item is categorized as *understand*. The noun in this assessment item is the particular geometric sum and a geometric sum is formed by a principle, i.e. *conceptual knowledge*. Therefore, this assessment item is placed in the cell *understand conceptual knowledge* in the taxonomy table (see Figure 5).

The taxonomy table can visualize the distribution of standards and the distribution of assessment items and thereby make it easier to analyze the alignment between the standards and the assessment. To exemplify the use of the taxonomy table in alignment analysis, eight extra standards and ten extra assessment items are also placed in Figure 5 (labelled S3-S10 and A3-A12).

Figure 5 presents an alignment situation that is problematic, because all standards are not assessed and there are assessment items that assess something else than the standards. There are seven cells with standards and in three of these cells assessment items are also placed. The assessment items are placed in five cells. Three cells, *understand conceptual knowledge, understand procedural knowledge* and *apply procedural knowledge*, collected both standards and assessment items. In *apply procedural knowledge* more than half of the items are placed, but only about one third of the standards. In subjects like mathematics and chemistry procedural knowledge is important and therefore this knowledge category will collect a lot of categorizations.

As a measure of the degree of alignment, range and balance are commonly used. In this example, the range for the assessment items is five cells compared to seven cells for the standards, with three cells in common. To calculate a measure of balance between the standards and the assessment items, Porter's (2002) balance index can be used:

$B = 1 - \frac{\sum |x - y|}{2}$, where x is the proportion of the total number of categorized standards in each cell in the taxonomy table and y is the corresponding proportion of assessment items. When B=1 the distributions of standards and assessment items are the same and emphasize the same cells in the taxonomy table. B=0 means that the distributions are completely different.

The example in Figure 5 gives a balance index of 0.43 and this is a low level of balance. According to Webb (2002), who uses a similar index in his alignment method, a level of at least 0.7 indicates an acceptable level of alignment and between 0.6 and 0.7 a weakly acceptable level. The computed balance index for Figure 5 is not acceptable and neither is the range. If this kind of results is found before an assessment is administrated, then the assessment has to be revised and other assessment items have to be included. One way is to include assessment items that measure the cells with standards in the taxonomy table, maybe with a reduction of the number of assessment items in the cell *apply procedural knowledge*. If, on the other hand, the assessment is already administrated to the students, this result may imply many negative consequences for the education (see Chapter 3).

Article III reports categorization of one set of standards and one assessment and thereby the results can be used for analysis of alignment between these standards and the assessment. The article reports a balance index of 0.72, indicating an acceptable level of alignment. In Figure 1 in article III, the distributions of standards and assessment items can be viewed. One cell, *apply procedural knowledge*, collected almost half of the categorizations of the assessment items, while only about 21% of the categorizations of the standards are collected in that cell. The standards were spread on more cells than the assessment items. To increase the

degree of alignment between standards and the assessment, assessment items measuring *factual knowledge* should be added and a larger proportion of the assessment items should measure *conceptual knowledge*. Very few of the assessment items are categorized as *metacognitive knowledge* and this is a sound result. The categorized assessment had a summative purpose and *metacognitive knowledge* is personal and difficult to evaluate in a summative assessment (Pintrich, 2002).