EDUCATION & PSYCHOLOGY I+D+i
Research, Innovation and Solutions on-line
Editorial EOS

La alineación de estándares y evaluación. Un estudio teórico y empírico de métodos para la alineación.

Gunilla Näsström, Widar Henriksson

Dept. of Educational Measurement, Umeå University, Umeå

Suecia

gunilla.nasstrom@edmeas.umu.se

Contacto: Department of Educational Measurement, Umeå University, SE-901 87 Umeå, Sweden

Resumen

Introducción. En un sistema escolar condicionado al cumplimento de objetivos, la concordancia entre los documentos normativos con los objetivos y la evaluación es importante. Para poder evaluar si los colegios y los alumnos han alcanzado los objetivos, la evaluación tiene que converger con los objetivos. Diferentes modelos y métodos pueden ser usados para medir la concordancia entre los objetivos y la evaluación. Basado en la suposición de que un modelo tiene que poder incluir contenido y complejidad cognitiva, fueron identificados nueve diferentes modelos y luego estos modelos han sido detenidamente examinados con la referencia a criterios teóricamente definidos. La conclusión es que la taxonomía revisada de Bloom y la taxonomía de Porter son los modelos más adecuados.

Método. La taxonomía revisada de Bloom y la taxonomía de Porter son comparadas sobre la base de datos empíricos provenidos de los objetivos y la evaluación de un curso de bachillerato de química en Suecia. La comparación es basada en cinco reglas y en reliabilidad inter-evaluativa.

Resultado. La taxonomía revisada de Bloom es más inclusiva y exclusiva que la taxonomía de Porter. La fiabilidad inter-evaluativa para la clasificación de objetivos fue significativamente mejor para la taxonomía revisada de Bloom que para la taxonomía de Porter.

Conclusión. Basada en las cinco reglas es la conclusión de que la taxonomía revisada de Bloom es el mejor modelo.

Palabras clave: *alineación, estandards, evaluación, taxonomía revisada de Bloom, taxonomía de Porter.*

Recibido: 20/09/07 *Aceptación inicial:* 18/10/07 *Aceptación final:* 07/04/08

Abstract

Introduction. In a standards-based school-system alignment of policy documents with standards and assessment is important. To be able to evaluate whether schools and students have reached the standards, the assessment should focus on the standards. Different models and methods can be used for measuring alignment, i.e. the correspondence between standards and assessment. Based on the assumption that a model must be able to include content and cognitive complexity, nine different models are identified and these models are then scrutinized with reference to defined theoretical criteria. The conclusion is that Bloom's revised taxonomy and Porter's taxonomy are the most appropriate models.

Method. Bloom's revised taxonomy and Porter's taxonomy are compared based on empirical data from standards and assessment in a chemistry course in upper secondary schools in Sweden. The comparison is based on five rules and of inter-rater reliability.

Results. Bloom's revised taxonomy was more inclusive and exclusive than Porter's taxonomy. The inter-rater reliability for classification of standards was significantly better for Bloom's revised taxonomy than for Porter's taxonomy.

Conclusion. Based on the five rules, the conclusion is that Bloom's revised taxonomy is the best model.

Keywords: *Alignment, Standards, Assessment, Bloom's revised taxonomy, Porter's taxonomy*

Received: 09/20/07 *Initial Acceptance:* 10/18/07 *Definitive Acceptance:* 04/07/08

Introduction

The importance of alignment

The concept of alignment involves a description of the relationship between three components in an educational system: standards defined in policy documents, teaching, and assessment. In this kind of educational system, a standards-based school-system, the students are supposed to reach the standards. To be able to evaluate whether the students have reached those standards, assessments should measure the standards, i.e. the assessment should be aligned with the standards. Alignment between standards and assessment is important for the effectiveness of an educational system (Webb, 1997), students' learning (Anderson, 2002; Biggs, 2003; Farenga, Joyce & Ness, 2002; La Marca, Redfield, Winter, Bailey & Hansche, 2000), accountability decisions (Koretz & Hamilton, 2006; La Marca, 2001), evaluation of educational reforms (Herman, Webb & Zuniga, 2007), validation of interpretation of assessment scores (La Marca, 2001; Rothman, 2003), information to students, parents, the public and politicians (Herman, Webb & Zuniga, 2007). Thus, alignment is a fundament in standards-based education (Fuhrman, 2001) and the question of interest in this context is how to study alignment. In this article, a theoretical and an empirical investigation of possible tools for alignment studies are presented.

Alignment in a standards-based school-system

Standards are descriptions, in policy documents, of what or how well a student should be able to master a certain knowledge and ability. Standards are commonly divided into two categories: content standards and performance standards (Hambleton, 2001). Content standards refer to what the students are expected to know or be able to do. Performance standards describe *how well* the students are expected to know or be able to do in relation to the content standards. The educational process aims to make it possible for the students to reach the standards and the process of assessment aims to measure the standards that, in turn, are related to the curriculum.

Expressed in general terms, alignment can be described as a situation where things are brought into a straight line (Baker, 2004). For an educational system, this means that the components in the system (standards, education and assessment) are arranged in a line, with the standards in the first position. One possible way of obtaining alignment in this general

meaning is to start with the curriculum, then define the standards and use the standards as a basis for defining blueprints. These blueprints, according to Baker, can then be used as a point of departure for teaching and assessment. Very often this procedure will result in very specific and detailed descriptions and, since both teaching and assessment can be regarded as dynamic and cyclic processes, the conclusion is that there is a need for more general methods as a basis for determining the degree of alignment (La Marca, Redfield, Winter, Bailey and Hansche, 2000).

Different terms are also used to describe the concept of alignment. Alignment appear when two or all three components in a certain education system are consistent (e.g. Biggs, 1999; Blank, Porter & Smithson, 2001), in agreement (e.g. Bhola, Impara & Buckendahl, 2003; Webb, 1997), matched (e.g. La Marca, 2001; Olson, 2003) or work together (e.g. Ananda, 2003; Roach, Elliot & Webb, 2005). Most commonly, alignment between standards and assessment has been analyzed (e.g. Bhola, Impara & Buckendahl, 2003; Herman, Webb & Zuniga, 2007), but alignment between standards and instruction as well as between instruction and assessment has also been studied (e.g. Porter, 2002). Several methods see alignment between standards and assessments as a means of increasing student learning, for example Webb (1997), Hansche (1998) and Roach, Elliot and Webb (2005). One way to deal with these differences in terminology is to focus on measurement and design, i.e., how alignment is measured and the design for alignment studies. This will be the focus in this article.

Methods for studying alignment

Methods that are used in alignment studies have varied and can, with reference to measurement, be classified according to complexity (Bhola, Impara & Buckendahl, 2003). Regardless of the methods of measurement, the design includes in most cases a classification of assessment compared with standards. In the least complex methods the model for measurement implies that content standards and assessment are matched. In methods with moderate complexity, standards and assessment are simultaneously classified and matched regarding content and cognitive complexity. Methods with high complexity add more complex criteria than just matching for evaluating the degree of alignment.

In all alignment studies, certain criteria are defined as a basis for classification (Bhola, Impara & Buckendahl, 2003). The criteria that are defined are quite similar, even if the terms

used differ to some extent. This conclusion is based on an analysis with design and model for measurement as a point of departure. Four categories of commonly used criteria are identified and these criteria are presented and defined in Table 1. The two most frequently occurring categories are content and cognitive complexity and the other categories, range and balance, are less common.

Content is defined, by most authors, by referring to topics and subtopics in a subject, e.g. NCTM (1989), Porter & Smithson (2001a), Mullis et al (2001) and Porter (2002). The number of topics and subtopics can be considerably large. Porter & Smithson (2001a), for example, defined 177 topics in science. The strategy of using topics and subtopics also implies that most sets of alignment criteria are bound to a single or a few, narrow subjects, i.e. these methods for alignment studies cannot be used as a general strategy for all subjects. Bloom's revised taxonomy (Anderson & Krathwohl, 2001) is probably, so far, the only model that defines content in general terms, i.e. in terms of different forms of knowledge.

The term cognitive complexity is found in all the studied sets of criteria, but complexity is given different names (see Table 1). Also the number of levels of cognitive complexity also differs. For example, Webb (1999) defines four levels, Porter (2002) five levels, Bloom's revised taxonomy six levels, and Porter & Smithson (2001a) nine levels. Cognitive complexity is assumed to be related to a continuous scale ranging from a low to a high degree of complexity (Anderson & Krathwohl, 2001; Webb, 1999).

It can also be added that some of the studies, referred to in Table 1, have proposed other criteria than the four mentioned above. The criteria focused on are, for example, equity, fairness and pedagogical implications (Webb, 1997), sources of challenge (Webb, 2007), and accessibility (La Marca, Redfield, Winter & Bailey, 2000).

The four alignment criteria from the review referred to above are either quantitative or qualitative. Range and balance are quantitative variables that compare the assessment as a whole with the standards as a whole. Content and cognitive complexity are based on a qualitative classification of individual standards and assessment questions. Some kind of framework or taxonomy is needed for this latter classification.

Table 1. The categories of the most commonly used criteria in alignment studies, including definitions and references to studies.

Category of criterion	Definitions	Studies
Content	Often defined by topics and subtopics. Content can also be defined as kinds of knowledge.	NCTM; 1989; Webb, 1997; La Marca, Redfield, Winter & Bailey, 2000; Anderson & Krathwohl, 2001; Mullis et al, 2001; Porter & Smithson, 2001a; Porter, 2002; Rothman, Slattery, Vranek & Resnick, 2002; Herman, Webb & Zuniga, 2005
Cognitive complexity	The level of complexity of what the students are supposed to do with information, but also the cognitive complexity of the information.	NCTM; 1989; Webb, 1997; La Marca, Redfield, Winter & Bailey, 2000; Anderson & Krathwohl, 2001; Mullis et al, 2001; Porter & Smithson, 2001a; Porter, 2002; Rothman, Slattery, Vranek & Resnick, 2002; Herman, Webb & Zuniga, 2005
Range	How many of the standards an assessment covers, in at least one question	Webb, 1997; La Marca, Redfield, Winter & Bailey, 2000; Porter, 2002; Rothman, Slattery, Vranek & Resnick, 2002
Balance	How well an assessment reflects the emphasis the standards give to a particular content	Webb, 1997; La Marca, Redfield, Winter & Bailey, 2000; Rothman, Slattery, Vranek & Resnick, 2002

Based on the finding that most standards are of two kinds, content standards and performance standards (Hambleton, 2001) the conclusion is also that performance standards include a cognitive dimension. Another finding is that most standards and assessments in schools today include content specifications as well as descriptions of cognitive levels. A closer look at the Swedish system also confirms this conclusion. This is another reason for focusing content as well as cognitive complexity in this article.

A theoretical investigation, that includes all frameworks and taxonomies that can categorize both content and cognitive complexity, is described in the next section.

Theoretical investigation

Nine taxonomies and frameworks are found to be useful for categorizing both content and cognitive complexity. These nine frameworks and taxonomies are summarized in table 2. These frameworks and taxonomies are: Bloom's revised taxonomy (Anderson & Krathwohl, 2001), DeBlock (de Landsheere, 1990), De Corte (de Landsheere, 1990), Guilford (1967), Marzano (2001), Merrill's performance-content matrix (1994); PISA (OECD, 1999), Porter (Porter & Smithson, 2001a, 2001b) and TIMSS (Robitaille et al., 1993).

It is true that content criteria are included in Webb's framework or taxonomy (Webb, 1997) but he does not offer any tool for categorizing content. Therefore, Webb's taxonomy is excluded in this study. The remaining frameworks and taxonomies will be scrutinized in the following section.

Marzano (2001) has developed a taxonomy that is focused on the development of thinking. He defines three systems of thinking that are hierarchically ordered and placed on a scale that varies from low to high level of consciousness. This means that the basis of his taxonomy is consciousness of how to process thinking, not cognitive complexity.

In TIMSS (Robitaille et al., 1993) the aspect of performance expectations can be regarded as comparable with cognitive complexity, but the authors claim that there is no relation between the categories in this aspect. Cognitive complexity, by the definition in the alignment review, is a gradient and there has to be some kind of relationship between the categories of this dimension. For PISA (OECD, 1999), the authors also claim that the scien-

tific processes, which are comparable with cognitive complexity, are not hierarchical. Merrill (1997) did not state that the performance categories have any kind of relationship between the categories or are placed on a scale of cognitive complexity.

Table 2. Frameworks and taxonomies with at least two dimensions for cognitive domain, useful in chemistry, and names of their dimensions that can be connected to the criteria of content and of cognitive complexity. Any third or fourth dimension is added.

Framework/taxonomy	Content criteria	Cognitive complexity	More dimensions
Bloom's revised taxonomy	Knowledge dimension	Cognitive process dimension	-
De Block	Content	Method	Transfer
De Corte	Subject matter	Operation	The domain The product
Guilford	Product	Operational categories	Content categories
Marzano	Domains of knowledge	Level of thinking system	-
Merrill	Content categories	Performance categories	
PISA	Scientific concepts	Scientific process	Scientific situations
Porter	Topics	Expectations of students' performance (cognitive demands)	Mode of presentation (is not used by Porter)
TIMSS	Content	Performance expectations	Perspectives

The taxonomies of De Block (de Landsheere, 1990), De Corte (de Landsheere, 1990) and Guilford (1967) have at least three dimensions in their frames of reference or taxonomies. Therefore, to be able to use these taxonomies according to the alignment criteria, one or two dimensions have to be excluded. There is a risk involved in excluding one or more dimensions since the excluded dimension could be a part of the other dimensions and therefore the taxonomy will not be inclusive enough. According to Hauenstein (1998) it is important that a taxonomy includes all possible categories. Thus, the conclusion is that elimination of dimensions includes a risk.

Porter has three dimensions in his taxonomy. The third dimension, mode of presentation, has been found to be of little use in alignment analysis (Porter & Smithson, 2001a) and is therefore excluded from the taxonomy. Thus, the Porter taxonomy is two-dimensional including content and cognitive complexity, which are ordered on a scale.

Based on the examination of the nine frameworks or taxonomies in Table 2, the summarised conclusion is that there are two remaining models: Bloom's revised taxonomy and Porter's taxonomy. These two taxonomies are able to categorize both content and cognitive complexity, the latter also liying on a continuous scale. The next question to ask is: which model should we use? When answering this question it is relevant to refer to the concept of usefulness, i.e. usefulness as a tool for description of alignment. The basis for answering this question is empirical in this article, i.e. these two taxonomies are empirically tested for their usefulness in alignment analysis. The taxonomies have been applied to one syllabus and one assessment in chemistry in upper secondary schools in Sweden.

Objective

The main purpose of the empirical investigation is to compare Bloom´s and Porter's models in order to identify the most useful model for studying alignment. Five rules are defined for this comparison (Hauerstein, 1998).

Method

The theoretical investigation resulted in two taxonomies. The usefulness of these taxonomies for categorizing both standards and assessment questions was then empirically tested through individual categorizations made by two judges of one set of standards and of the items in one assessment in chemistry with both taxonomies. The usefulness of the taxonomy was investigated based on criteria presented below and on the level of inter-judge consistency.

Criteria for usefulness

Hauenstein's (1998) five rules for taxonomies are used for the empirical comparison. These five rules are: 1) applicable 2) totally inclusive; 3) mutually exclusive; 4) following a

consistent principle of order; 5) the terms used in categories and subcategories are representative of those used in the field. When translating these rules into the conditions for the empirical comparison in this study, five questions can be asked: 1) Is the taxonomy applicable to alignment analysis? 2) Can all the standards and assessment questions be included in the taxonomy? 3) Is there any overlap between the categories, or subcategories, in the taxonomy, which will lead to categorizations of a single standard and/or assessment question in more than one category? 4) Are categories arranged in a consistent principle of order? 5) Are the terms in categories and subcategories representative of those used in the field?

Empirical data in this study can be used to answer the question of applicability (1), totally inclusive (2) and mutually exclusive (3). Hauenstein's remaining two rules (4, 5) are discussed with reference to knowledge of the practical conditions in upper secondary school in Sweden.

Material

Standards

The set of standards in this study is formed by a division of the standards that make up one syllabus in chemistry for upper secondary schools in Sweden. The syllabus contains a total of 23 original standards, of which 14 are content standards and 9 are performance standards. The division of the original standards is based on the principle in Bloom's revised taxonomy, which states that a standard should be formed by a noun and a verb (Anderson & Krathwohl, 2001). This division resulted in 102 sub-standards, which form the set of standards used in this study.

Assessment

The assessment that is used in this study is commonly used in upper secondary schools in Sweden. The purpose of this assessment is to support teachers' grading of their students and is supposed to be an interpretation of the studied syllabus. Every year, different versions of the assessment are offered to schools. The particular assessment was given in spring 2005. It consisted of 58 questions.

Instruments

Bloom's revised taxonomy

Bloom's revised taxonomy (Anderson & Krathwohl, 2001) has two dimensions, the knowledge dimension and the cognitive process dimension. In the knowledge dimension the content is defined as different kinds of knowledge. The categories of the knowledge dimensions are *factual knowledge, conceptual knowledge, procedure knowledge* and *metacognitive knowledge*. The categories in the knowledge dimension lie along a continuum, from concrete; as in *factual knowledge,* to abstract, as in *metacognitive knowledge*. There is no clear-cut border between *conceptual* and *procedural knowledge*.

The dimension of cognitive processes is focused on how the knowledge is used. The categories of the dimension of cognitive processes are *remember, understand, apply, analyse, evaluate* and *create*. The underlying continuum in the dimension of cognitive complexity is cognitive complexity, ranging from little cognitive complexity in *remember* to the most cognitive complexity in *create*. This dimension represents the alignment category *cognitive complexity.*

Bloom's revised taxonomy offers a two-dimensional taxonomy table with 4x6=24 cells. The rows in the taxonomy table represent the four main categories of the knowledge dimension and the columns the six main categories of the cognitive process dimension. Based on Bloom's revised taxonomy a standard can, for example, be categorized according to the two dimensions and placed in a certain cell in the taxonomy table.

Porter's taxonomy for alignment analysis

Porter's taxonomy was developed to offer a systematic and uniform language for describing content (Porter, 2002) and for making detailed quantitative comparisons of standards, assessment and teaching in science and mathematics (Porter & Smithson, 2002). The results of analyses are presented in content maps, which resemble topographic maps. The "height gradient" is represented by the percentage of total time for each content.

Initially the taxonomy had three dimensions: topics, expectations of student performance and modes of presentation (Porter & Smithson, 2001a). The dimension of modes of presentation had seven categories: *exposition, pictorial models, concrete models, equa-*

tions/formulas, graphical, laboratory work, and *fieldwork.* However, the authors found it difficult to integrate modes of presentation in the alignment analyses and their conclusion was to exclude this dimension, i.e. to use a two-dimensional approach.

The dimension of topics is a list of topics in either mathematics or science. There is no hierarchy in this dimension. In science, the dimension of topics contains 177 topics for upper secondary schools, divided into 25 content areas (Porter & Smithson, 2001a). The categories used in the empirical study are the content areas.

The number of categories in the second dimension, expectations of student performance, has varied between four and nine (Porter & Smithson, 2001b). In their first approach, nine categories were used, but the number of possible combinations of the two dimensions was too large to handle in educational settings. Therefore, and in order to make the classification easier, the number of categories was reduced (Porter & Smithson, 2001a). When applying Porter's model to the empirical data (Chemistry) in this study the approach with seven categories will be used, because this approach is the only one which offers definitions that can be applied to science (Porter & Smithson, 2001b). It can also be added that the categories differ somewhat between mathematics and science. For science the categories are: 1) memorize facts, definitions, formulas; 2) understand concepts; 3) perform procedures; 4) generate questions/hypotheses; 5) collect data; 6) analyze and interpret information; 7) use information to make connections. This dimension is ordered by level of cognitive demand.

The two dimensions form a two-dimensional matrix with 25x7=175 cells. The rows will represent the 25 content areas of the dimension of topics and the columns the 7 categories of expectations of student performance. In a way similar to Bloom's revised taxonomy, a standard can, for example, be categorised by the intersection of the two dimensions of Porter's taxonomy and placed in the corresponding cell.

Judges

The panel of judges consisted of two assessment experts, with relevant education and experience of teaching Chemistry. The two assessment experts have experience of both developing and constructing national assessments in upper secondary schools.

Procedure

The procedure was carried out in two stages. In the first stage, the training stage, the judges received general information about the Bloom's revised taxonomy and Porter's taxonomy and examples of standards, and authentic assessment questions, to practice categorization on. The practice categorization was made individually and followed by a consensus discussion. In the second stage, the two judges categorized the set of standards and the assessment questions individually. The obtained result is the basis for the empirical investigation.

Statistical description and analysis

Statistics for inclusiveness, mutual exclusiveness and range of each taxonomy are reported. The proportion of standards and assessment questions that can be categorized indicates the degree of total inclusiveness. The proportion of standards and assessment questions that are categorized in more than one category indicates the degree of mutual exclusiveness. The proportion of cells with categorized standards and/or assessment questions is reported and indicates to what extent each taxonomy is applicable.

The inter-rater reliability is indicated by Fleiss' kappa: K_f (Fleiss, 1971) as well as by proportions of agreement between judges. Fleiss' kappa was chosen because the data is on a nominal level (Stemler, 2004) because content categories in both taxonomies are not fully ordered along a continuum. According to Landis and Koch (1977), kappa values between 0.01 and 0.20 represent slight agreement, those between 0.21 and 0.40 fair agreement, those between 0.41 and 0.60 moderate agreement, and those greater than 0.60 substantial agreement. Differences in percentage of agreement are tested if they are significant with χ^2 at 95 % confidence level.

Results

Results for inclusiveness, mutually exclusiveness, applicability and inter-rater reliability are presented and compared for Bloom's revised taxonomy and Porter's taxonomy. Table 3 gives an overview of obtained data from the empirical comparison of Bloom's revised taxonomy and Porter's taxonomy.

Table 3. Results from the classification by two judges of standards and the assessment

		No. not classified		No. classified in ≥2 categories		Range (% of categories)		K_f	Equal categorization
	Taxonomy	Judge 1	Judge 2	Judge 1	Judge 2	Judge 1	Judge 2		
Standards	Bloom's revised	0	4	0	1	58 %	71 %	.46	53%[1]
	Porter's	11	8	54	45	35 %	32 %	.07	37%[1]
Assessment questions	Bloom's revised	0	0	0	0	29 %	21 %	.36	48%[2]
	Porter's	0	0	20	14	9 %	6 %	.30	60%[2]

1) At 95% confidence level, X^2 shows a significant difference; 2) No significant difference

The obtained data indicates that more standards remained uncategorized with Porter's taxonomy than with Bloom's revised taxonomy. About 10 percent of the standards were not classified in Porter's taxonomy, while one judge (Judge 1) classified all standards and one judge all but four standards with Bloom's revised taxonomy. The data in Table 3 also indicatea that there was no difference between the two models regarding the assessment questions, as all questions were categorized.

Regarding the proportion of standards categorized in more than one category, the data in Table 3 indicates a big difference between the two models. All standards, except one (Judge 2), were placed in only one category in Bloom's revised taxonomy. The result was quite the opposite for Porter's taxonomy - a large proportion of the standards was placed in two or more (≥2) categories. This difference between the two models was also obtained for the assessment, i.e. all questions in the assessment were placed into only one category in Bloom's revised taxonomy, but not in Porter's taxonomy.

When considering the applicability of a taxonomy, the range of categories used for classification is a good estimate. If most of the categories are used, the taxonomy will give a more diversified picture of what is classified. The proportion of categories used is larger for Bloom's revised taxonomy than for Porter's taxonomy, both for standards and assessment (see Table 3).

Fleiss' kappa indicates that the inter-rater reliability levels are much higher for standards, and marginally higher for the assessment questions, for Bloom's revised taxonomy as compared to Porter's taxonomy (see Table 3). The kappa coefficients for the standards indicate a moderate agreement for Bloom's revised taxonomy and a slight agreement for Porter's taxonomy. For the assessment questions, the kappa coefficients for both taxonomies indicate a fair agreement. For the standards, the proportion of agreement between the two judges is significantly higher for Bloom's revised taxonomy than for Porter's taxonomy. In contrary, for the assessment questions, the proportion of agreement is higher for Porter's taxonomy. However, the difference in agreement for the assessment is not significant.

To summarize, the data obtained from the empirical investigation indicated that Bloom's revised taxonomy categorized more standards and assessment questions, that only one standard was placed in more than one category, and that it had a larger proportion of used categories and higher levels of inter-rater reliability compared to Porter's taxonomy.

Conclusions

The purpose of this study was to find the most useful tool for alignment studies when focusing on the relationship between standards and assessment. The assumption was that alignment should include content and cognitive complexity, which resulted in nine possible models. However, only two of the models, Bloom's revised taxonomy and Porter's taxonomy, also supported the assumption that cognitive complexity lies on a continuous scale, i.e. categories are ordered. An empirical comparison of these two taxonomies considering the usefulness and inter-rater reliability levels indicated that Bloom's revised taxonomy was more useful than Porter's taxonomy. This conclusion was based on an evaluation of usefulness based on Hauenstein's (1998) five rules.

The first rule, i.e. whether the taxonomies are applicable, was partially answered by the theoretical selection of taxonomies. In the alignment review a taxonomy should be able to classify both content and cognitive complexity. Both Bloom's revised taxonomy and Porter's taxonomy fulfilled these criteria. Amer (2006) also found Bloom's revised taxonomy useful for alignment analyses. Another aspect of the applicability is what percentage of the categories is used when standards and assessment questions are classified. The results showed that a larger proportion of the categories in Bloom's revised taxonomy than in Porter's taxonomy was used. Therefore Bloom's revised taxonomy can be considered more applicable than Porter's taxonomy.

The range of used categories was smaller for assessment than for standards, regardless of taxonomy. Based on this fact, one conclusion is that the degree of alignment of standards and assessment may increase if the assessment is completed with new questions that are related to standards.

The second rule deals with total inclusiveness, i.e. that every standard and assessment question should be classifiable by means of the taxonomy. Both taxonomies were totally inclusive for assessment questions, but neither of the taxonomies was totally inclusive for the standards. Bloom's revised taxonomy was, however, more inclusive than Porter's taxonomy. Therefore the conclusion is that Porter's taxonomy is a less useful tool for classifying present standards and, as a consequence, less useful in alignment studies. But it must be added that Bloom's revised taxonomy is not totally perfect in this respect. One judge could not classify 4 standards with Bloom's revised taxonomy.

The third rule deals with mutual exclusiveness, i.e. standards and assessment questions should be classified in only one category. In this aspect, Bloom's revised taxonomy is better than Porter's taxonomy. For Bloom's revised taxonomy one judge classified one standard in more than one category, but all the other standards and assessment questions were classified in only one category. For Porter's taxonomy a large percentage of both standards and assessment questions were classified as belonging to more than one category. According to the third rule, Bloom's revised taxonomy is superior to Porter's taxonomy.

The fourth rule, i.e. if the categories are ordered by a consistent principle, is answered by the selection of the two taxonomies. The alignment criterion of cognitive complexity is by definition on a scale from low to high cognitive complexity and the empirically investigated taxonomies both fulfilled this definition. However, the scale steps are not the same for the two taxonomies. For the content criteria the principles of order differ more between the two taxonomies. The knowledge dimension in Bloom's revised taxonomy is, according to the authors, ordered from concrete in factual knowledge to abstract in metacognitive knowledge, with some overlaps in conceptual and procedural knowledge. In Porter's taxonomy content is categorized in topics, which are only ordered in broad categories, so called content area, without any relation among them.

The fifth question, i.e. whether the terms in the taxonomies are representative of the field is discussed by the authors of the taxonomies. It has been questioned whether the terms in Porter's taxonomy represent the terms in the present national standards in the US (Porter & Smithson, 2001b). The authors reply that a reform-neutral language is better for a taxonomy that has a chance of surviving new reforms. In Bloom's revised taxonomy, the aims were to

use a common language, to be consistent with current psychological and educational research and to use realistic examples. The group of authors of this taxonomy consisted of representatives of the fields of cognitive psychologists, curriculum theorists and instructional researchers, as well as testing and assessment specialists. In a study of learning frameworks, Mosely et al. (2005) gave credit to Bloom's revised taxonomy for the vocabulary. Both taxonomies are therefore assumed to use terms that are representative of the field.

For the standards, the level of inter-rater reliability was higher for Bloom's revised taxonomy than for Porter's taxonomy, indicating that Bloom's revised taxonomy was the more reliable taxonomy. However, for the assessment questions the levels were about the same for both taxonomies.

The conclusion from the discussion above is that Bloom's revised taxonomy is more useful than Porter's taxonomy as a classification tool in alignment studies and therefore, the most useful classification tool today.

Discussion

One interesting current issue is how content is stated in standards and what kinds of categories an alignment tool should have. The historical trend is, according to Ward, Stocker & Murray-Ward (1984), that content standards are becoming broader and less detailed. Baartman, Bastiaens, Kirschner & van der Vleuten (2006) state that content standards nowadays are more competence-based. A useful alignment tool should therefore be able to classify standards that are broad and more competence-based. A consequence is that taxonomies and frameworks that include categories that are based on topics and subtopics will be less useful. Thus, the conclusion is that to be able to classify modern kinds of standards, there is a need to have more general categories and probably also categories for meta-cognitive aspects. Bloom's revised taxonomy classifies content as forms of knowledge and therefore this model also allows for classification of more general standards.

Based on the assumptions that a model for analyzing alignment of standards and assessments should be able to categorize both content and cognitive complexity, and that the categories of cognitive complexity are ordered on a continuous scale, Bloom's revised taxon-

omy seem to be a very useful model. However, the question is whether this taxonomy can be further improved in order to obtain an even better tool for alignment analyses of standards and assessment? One possible strategy worth testing is to establish better defined categories, for example by tailoring definitions and examples to different categories or topics.

In this context it is also relevant to mention that Thurston, Grant, and Topping (2006) have also emphasized the importance of questioning in teaching science and that there is a need to develop such questions. A classification tool, like Bloom's revised taxonomy, can be of help to construct questions related to the standards in a more effective way, by analyzing the standards first and then constructing questions (Martineau, Paek, Keene & Hirsch, 2007).

It is also relevant to question whether it is an optimal strategy to add new categories to Bloom's revised taxonomy? In this context it is worth mentioning that there is a risk involved when the number of categories increases. An increase of categories in a taxonomy may lead to a decrease in inter-rater reliability, at least when the number of categories extend seven (Wolf, 1997). In this context it can also be noted that the number of categories in Porter's taxonomy is more than seven times larger than in Bloom's revised taxonomy. Therefore, the conclusion is that a strategy for refining Bloom's model by adding new categories, should be very carefully considered before being implemented.

References

Amer, A. (2006). Reflections on Bloom's revised taxonomy. *Electronic Journal of Research in Educational Psychology, 4*(1), 213-230.

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

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

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A & van der Vleuten, C. P. M. (2006). The wheel of compentency assessment: presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation, 32*(2), 153-170.

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

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

Biggs, J. (1999). *Teaching for quality learning at university.* Birmingham: Open University Press.

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

Blank, R. K., Porter, A., & Smithson, J. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics & science. Results from survey of enacted curriculum project. Final report.* Washington: Council of Chief State School Officers.

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.

de Landsheere, V. (1990). Taxonomies of educational objectives. In Herbert J. Walberg & Geneva D. Haertel (Ed.), *The international encyclopedia of educational evaluation* (pp. 179-188). Oxford: Pergamon Press.

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. Bybee (Ed.), *Learning science and the science of learning*. Arlington: NSTA press.

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.

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

Hambleton, R.K. (2001). Setting performance standards on educational assessment and criteria for evaluating the process. In G.J. Cizek (Ed.), *Setting performance standards: concepts, methods, and perspectives* (pp. 89-116). Mahwah: Lawrence Erlbaum.

Hansche, L. N. (1998). *Meeting the requirements of title I: Handbook for the development of performance standards.* Washington: U.S. Department of Education.

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. (2005). *Measurement issues in the alignment of standards and assessments: A case study* (CSE report 653). Los Angeles: National Center for Research on Evaluation, Standards, and Student testing.

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.

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.

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.

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

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

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.

Merrill, M. D. (1994). *Instructional design theory.* Englewood Cliffs: Educational Technology Publications.

Mosely, D., Baumfield, V., Higgins, S., Lin, M., Miller, D., Robson, S., Elliot, J. & Gregson, M. (2004). *Thinking skill frameworks for post-16 learners: an evaluation. A research report for the learning and skills research centre.* Trowbridge: Learning & Skills research centre.

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.

NCTM. (1989). *Curriculum and evaluation standards for school mathematics.* Reston: NCTM.

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

Olson, L. (2003). Standards and tests: Keeping them aligned. *Research Points, 1*(1), 1-4.

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. (2001a). *Defining, developing, and using curriculum indicators* (CPRE Research report series RR-048). Philadelphia: Consortium for Policy Research in Education.

Porter, A. C. & Smithson, J. L. (2001b). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S. H. Fuhrmans (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. L. (2002). *Alignment of assessments, standards and instruction: Using curriculum indicator data.* Paper presented at the Annual meeting of the American Educational Research Association, New Orleans, April 1-5 2002.

Roach, A. T., Elliot, S. N., & Webb, N. L. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin alternate assessment. *The Journal of Special education, 38*(4), 218-231.

Robitallie, D. F., Schmidt, W. H., Raizen, S., McKnight C., Britton, E. & Nicol, C. (1993). *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.R. (2002). *Benchmarking and alignment of standards and testing* (CSE Technical report 566). Los Angeles: National Center for Research on Evaluation, Standards, and Student testing.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4).

Thurston, A., Grant, G. & Topping, K. J. (2006). Constructing understanding in primary science: An exploration of process and outcomes in the topic areas of light and the earth in space. *Electronic Journal of Research in Educational Psychology, 4*(1), 1-34.

Ward, A. W., Stocker, H.W., & Murray-Ward, M. (1984). *Educational measurement. Origins, theories and explications. Volume II. Theories and applications.* Lanham: University Press of America.

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. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research monograph, No. 18). Madison: National Institute for Science Education.

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

Wolf, R. M. (1997). Rating scales. In J. P. Keeves (ed.) *Educational research, methodology, and measurement: An international handbook* (pp.958-965). Cambridge: Pergamon.