Lopez, Alexis A. (2013). Alignment between standardized assessments and academic standards: The case of the Saber Mathematics Test in Colombia. *RELIEVE*, *v. 19* (2), art. 2. DOI: 10.7203/relieve.19.2.3026

e-Journal of Educational Research, Assessment and Evaluation

RELIEVE

Revista ELectrónica de Investigación y EValuación Educativa

ALIGNMENT BETWEEN STANDARDIZED ASSESSMENTS AND ACADEMIC STANDARDS: THE CASE OF THE SABER MATHEMATICS TEST IN COLOMBIA

[*Alineación entre las evaluaciones externas y los estándares académicos: El Caso de la Prueba Saber de Matemáticas en Colombia*]

by/por

Article record

About authors

HTML format

Lopez, Alexis A. (alopez@ets.org)

Ficha del artículo

Sobre los autores

Formato HTML

Abstract

In this study the concept of alignment is analyzed by presenting the results of a study that examines the degree of alignment between a standardized mathematics assessment and a set of academic standards. The alignment was examined using the Webb model and the results suggest that the degree of alignment between standardized assessments and academic content standards is not the most appropriate and therefore the results should be interpreted with great caution and stakeholders must be very careful with the type of decisions that are made based on the results of these types of assessments.

Keywords

Alignment, Standardized Tests, Academic Standards, Assessment of learning, Validity.

Resumen

En este estudio se analiza el concepto de alineación presentando los resultados de un estudio que examina el grado de alineación entre una evaluación estandarizada de matemáticas y unos estándares académicos. La alineación se examinó usando el modelo de Webb y los resultados sugieren que el grado de alineación entre las evaluaciones estandarizadas y los estándares académicos no es el más adecuado y por lo tanto los resultados se deben interpretar con mucha cautela y se debe poner mucho cuidado con el tipo de decisiones que se toman con base en los resultados.

Descriptores

Alineación, Pruebas estandarizadas, Estándares Académicos, Evaluación del aprendizaje, Validez.

Assessment is a key aspect in the teaching and learning process, as the results are used to obtain important information about what students know and can do, and to make important decisions that will affect their lives. Therefore, it is critical that these assessments are valid and appropriate for their intended purpose. For an assessment to be valid, at least, it has to be aligned with what is expected to be taught in the classroom. Alignment means that there is agreement between what is tested and what is described in the curriculum (Webb, 1997). Generally, what is taught in class is guided by what the standards say. The alignment between tests and standards helps students to learn what is stipulated in the curriculum.

The concept of alignment is intrinsically linked to the concept of validity. The validity of a test is assessed in terms of empirical evidence and theoretical rationales that support the adequacy of the interpretations and decisions made based on the results of the test (Messick, 1989). There are two fundamental aspects that can invalidate the interpretations made based on the results of a test: 1) that the content and skills tested on the exam are not well represented and 2) that the content tested includes irrelevant facets or dimensions (Messick, 1989). In the first case, the test content is not restricted or limited and includes major facets or dimensions that are described in the standards. By contrast, in the second case the content is comprehensive and assesses content or skills that are not defined in the standards.

Depending on the degree of alignment that exists between an assessment and standards, there might be four different situations. In the first situation, the entire content of the assessment is aligned with the standards, but there are many content standards not being evaluated. In the second situation, all content standards are being evaluated, but the evaluation also is measuring many aspects that are not listed in the standards. In the third situation, only part of the standards are being evaluated, omitting many, but the evaluation is measuring some aspects that are not listed in the standards. Finally, we have the ideal situation where all questions are evaluating a representative sample of all statements contained in the standards. This is difficult to reach because of many limitations: for example, the number of content standards, the number of questions in the assessment and how it is assessed (the type of questions used).

In the case of many assessments, typically the content is defined by the contents described in academic standards. The assessment results indicate how much students have learned in relation to the content of the standards. If assessments are not aligned with the content of the standards, and therefore with what teachers teach in the classroom, the results are not entirely valid to determine how much they have achieved the expectations placed for each academic grade. Therefore, it is very important to determine the degree of alignment between standardized tests and academic standards to determine how to interpret these results and to understand what kind of decisions can be based on them.

Alignment studies can also serve to identify problem areas in the assessment tools (Lopez, Webb & Stansfield, 2006; Rothman, 2004). The results allow us to identify gaps in the assessment content and help us determine whether it is sufficiently representative of the standards (Herman, 2004; Porter, 2002; Rothman, 2004; Wise, 2004).

The aim of this study is to illustrate the concept of alignment by determining the level of alignment between the Saber Mathematics Test and the Colombian Basic Skills Standards in Mathematics. To establish the degree of alignment, we examined the following research questions:

1. What is the degree of alignment between the content of the Saber Mathematics Test and content in the standards?
2. What is the degree of alignment between the cognitive demand of the questions in the Saber Mathematics Test and the cognitive demand in the standards?
3. What is the degree of alignment between the range of knowledge in the questions on the Saber Mathematics Test and the range of knowledge of the standards?
4. What is the degree of alignment between the balance of representation in the questions Saber Mathematics Test and the balance of representation of the content standards?

Before providing more information on the study, it is important to note that the term "alignment" is sometimes used in the assessment area and in the field of psychometrics to refer to several things. For example, alignment can also refer to matching scores between tests or the relationship between a test and different aspects of the educational process, including academic standards, the teaching and learning process, and classroom assessment. In this study, the concept of alignment is only used to refer to the alignment between a test and a series of academic standards and it is considered part of the validity of the test.

Similarly, it is important to clarify that the alignment presented in this study focuses solely on the alignment between tests and standards. It is also important to remember to examine the alignment between the academic standards and curriculum practices (Matore and Sireci, 2009), between the assessment and curricular practices (McGehee and Griffith, 2001), and between standardized testing and classroom assessment practices (Ruiz-Primo et al, 2012). It is also interesting to examine how standards are implemented in the classroom and how teachers use them. I hope that this study becomes a stepping stone for further studies to examine other types of alignment and the ways in which standards are implemented in class. All these types of study allow a better understanding of the alignment as the final interpretation of test scores also depend on the above aspects.

The Colombian Basic Skills Standards in Mathematics

The Colombian Basic Skills Standards in Mathematics were developed by the Ministry of Education in collaboration with the Colombian Association of Faculties of Education (Ascofade), math teachers and members of the educational community. These standards describe the competencies and skills that students should have at the end of each grade and are aimed at improving the level of education. The Basic Skills Standards in Mathematics are divided into five basic areas: 1) numerical reasoning and numerical systems, 2) spatial reasoning and geometrical systems, 3) metric reasoning and measurement systems, 4) statistical reasoning and data systems, and 5) variational reasoning and algebraic and analytic systems, and each of them has several goals (see Appendix 1).

The Saber Mathematics Test

The Saber Test is a battery of tests in the areas of Language Arts, Mathematics, Social Sciences and Natural Sciences. These are census tests students take in grades 5 and 9. According to the designers of the Saber tests, the aim of these are to "obtain, process, interpret and disseminate reliable information so as to constitute a sound basis for decision-making in the different levels of educational services" (ICFES 2008). For this study we used the 2009 test for grade 9, which consists of fifteen multiple-choice questions with four options (see more information in http://www.icfes.gov.co). It is worth mentioning that the alignment of the standards test depends on the way the test is administered and who administers it. The degree of control of the administration determines the quality of the information, so it is important that the test is consistently administered to ensure that information collected is valid.

Theoretical framework

The alignment type discussed in this study is also known as alignment or horizontal linkage between the content standards and what is being measured in an external assessment (Martineau et al., 2007). Some authors consider that the degree of alignment between assessment and academic standards is a necessary condition to validate tests that measure learning (Messick, 1989, Webb, 1997). In this study, the degree of alignment was evaluated using the model proposed by Webb (1997). Some studies have shown that the Webb methodology is most suitable to determine the degree of alignment between standardized tests and academic standards as it examines the alignment from several perspectives (Herman, Webb, and Zuniga, 2007; Martone and Griffifth, 2009). In this model, the degree of alignment is measured based on four criteria: 1) the content being assessed, 2) the cognitive demand of the questions in the assessment, 3) the number of content standards that are being assessed, and 4) the way the questions are distributed across the standards.

Alignment Based on Content

An important condition that should be considered in alignment studies is whether the content assessed on a test is related to the content described in the standards (Webb, 1997; Webb, Herman & Webb, 2007). By content we mean the issues, knowledge, skills or competencies being assessed (AERA, APA, NCME, 1999).

Alignment Based on Cognitive Demand

To assess the degree of alignment between a set of standards and a test, it is also important to review the level of cognitive demand of the content standards and the level of cognitive demand of the assessment questions (Webb, 1997, Webb et al., 2007). Below are the cognitive demand levels used by Webb (1997).

Level 1 - Recall. At level 1 the student only has to identify or remember the answer and has no need for any kind of reasoning to answer a question. Examples include : measuring an angle , finding the area of a rectangle, multiplying two numbers, converting a number in scientific notation to decimal number , identifying a diagonal in a geometric figure.

Level 2 - Skills and concepts. At level 2, the student has to do some kind of mental process that goes beyond simply remembering or reproducing a response, so this process is more complex than in Level 1. These processes require the student to make a decision about how to answer a question or solve a problem. Some examples include: extending a geometric pattern; classifying quadrilaterals; organizing a set of data and build a table; determining a strategy for estimating the number of coins in a bowl; comparing two sets of data using the mean, median and mode of each set.

Level 3 - Strategic Thinking. At level 3, the student is required to use more complex cognitive processes than in the first two levels. For example, at this level the student has to reason, plan or use evidence. At this level, the students are required to answer the questions using several steps, justify the answers, draw conclusions based on observations and explain phenomena among others. Some examples include: explaining how changes in the dimensions of a geometric figure affect its area or perimeter; giving a mathematical justification to solve a problem; interpreting information from a figure; writing a math rule for a non-routine pattern; writing and solving equations for a given problem.

Level 4 - Extended Thinking. Level 4 requires the use of complex cognitive processes. At this level, the students have to make connections between ideas and concepts, choose among alternatives to solve a problem, or implement the results of an experiment in other contexts. Many of these activities need to be done over an extended period of time, but the fact that an activity does not take enough time is not a requirement for this level. Some examples include: collecting data considering several variables and analyzing results; proposing a rule for a complex pattern and finding a phenomenon that exhibits this behavior; modeling a social phenomenon with several alternatives and chooseing a method to solve it using a mathematical model.

Alignment Based on Range of Knowledge

This criterion is used to determine if the contents included in the assessment questions cover the content in the standards (Webb, 1997). This criterion only considers the number of content standards being assessed by at least one question. This criterion does not take into account the number of times a particular content standard is being assessed.

Alignment Based on Balance of Representation

This criterion is used to indicate how much emphasis is given to a particular content standard in an assessment (Webb, 1997). Unlike the previous approach, this criterion takes into account the number of times a certain content standard is being assessed. Therefore, this criterion takes into account the difference in the proportion between each thematic content and its number of related questions.

Previous studies have indicated that there are various degrees of alignment (Bhola, Impara and Buchendahl, 2003). Herman, Webb and Zuniga (2007) found that different groups of judges can produce different results, so it is very important to properly train judges to minimize this effect. Judges should familiarize themselves with the test, the standards and the coding process. As for the test, they must know the format, item types, resources available to take the test (e.g., calculators, dictionaries, etc.), and how the test is administered (Bhola, Impara and Buchendahl, 2003). As for the standards, judges must know how they are organized, how they are used in the classroom and how to assess them (Lopez, 2009).

Similarly, the results of the study provide important information to improve the tests (Lopez, Webb and Stansfield, 2006). The alignment results provide information about the kind of inferences that can be made about the learning process, especially if you are going to make important decisions based on the results of a test (Pehuniak, 2005). On the other hand, previous studies have found that constructed-response items (e.g., open-ended questions, essay questions, etc.) tend to be aligned to more than one content standard because they generally require students to apply different types of knowledge or skills to answer them (Webb, 2002).

Method

In this section the results of a non-experimental quantitative study are reported. This information is supplemented with qualitative data on the judges' perceptions about the test items and the level of alignment between the external assessment and the standards. The data for this study were collected at a five-hour alignment session.

Participants

This study had the participation of seven judges. Table 1 shows information about each of them. Their task was to code the Mathematics Skills Standards based on the level of cognitive demand. Additionally, the judges coded all the items in the 2009 Saber Math Tests of 2009 based on the level of cognitive demand and content in the standards. Finally, each judge commented on some problematic aspects related to the test items and discussed their perceptions of the degree of alignment between the external assessment and the standards.

Table 1. Information about the judges

Judge	Educational Background	Teaching Experience
1	B.A. in Mathematics; M.A. candidate in Education	High school math teacher (2 years); higher education math teacher (3 years)
2	B.A. in Mathematics	High school math teacher (6 years); higher education math teacher (4 year)
3	B.A. in Mathematics; M.A. in Mathematics	High school math teacher (2 years); higher education math teacher (3 years)
4	B.A. in Mathematics; M.A. candidate in Mathematics	High school math teacher (3 years); higher education math teacher (3 years)
5	B.A. in Mathematics; M.A. candidate in Mathematics	High school math teacher (5 years)
6	B.A. in Mathematics	High school math teacher (2 years); Math software developer (1 year)
7	B.A. in Mathematics	High school math teacher (5 years); higher education math teacher (1 year)

Data Collection

This study was conducted in four phases to facilitate the process of assessing the degree of alignment between the Grade 9 Mathematics Skills Standards and the Saber Math Test.

Phase One. In the first phase there was a training session for judges to familiarize them with the process of alignment, with cognitive demand levels that were used to code the standards and the test, and with the Grade 9 Mathematics Skills Standards and the Saber Math Test. There were a few exercises for the judges to allow them to practice each of these aspects followed by a time for questions and clarification.

Phase two. After training, each judge coded each standard individually. Then all the judges met with the researcher to share and justify their coding. At the end of this phase, the judges reached a consensus on the level of cognitive demand of each of the content standards (see Appendix 1).

Phase Three. In the third phase of the study each judge assigned a level of cognitive demand and up to three content standards to each item on the test using a coding format (see Appendix 2). At this stage, the judges also wrote comments on each test item in the coding format.

Phase Four. In the last phase the judges met with the researcher to discuss their impressions of the degree of alignment between the Grade 9 Mathematics Skills Standards and the Saber Math Test.

Data Analysis

Quantitative data were analyzed using four different criteria to determine the degree of alignment between the Grade 9 Mathematics Skills Standards and the Saber Math Test.

Alignment based on content. According to Webb (1997), this criterion is evaluated considering the number of test items that are related to the performance indicators described in the standards. To ensure an acceptable degree of alignment at the content level, it is expected that at least six test items are assessing the contents described in each basic area. This number is taken based on an estimate of the number of items that can produce a reliable subscale to make inferences about the degree of competence or ability of the student to that content (Subkoviak, 1998). If there are six or more items, it is considered that the degree of alignment is strong. If only five items are aligned, it is considered that the degree of alignment is weak, and if there fewer than five items associated with the standards, it is considered that the degree of alignment is inadequate.

Alignment based on cognitive demand. To analyze this criterion, it is important to compare the level of cognitive demand of each item with the level of cognitive demand of each associated performance indicator in the content standards. For there to be a degree of proper alignment, 50 percent of the items have to be at least the same level of cognitive demand. If the percentage of items that is at least have the same level is between 40-50 percent, it is considered that the level alignment is weak and if the percentage of items that have at least the same level of cognitive demand is less than 40 percent, it is considered that the level of alignment between the standards and the test items is not suitable.

Alignment based on range of knowledge. To analyze how much coverage the test items have, I examined the number of performance indicators that are associated with a particular test item. If at least 50 percent of the content in a main theme has at least one item on the test, it is considered that the alignment level is adequate. If 40-50 percent of a thematic content has at least one question, it is considered that the level of alignment is weak and if the percentage is less than 40 percent, it is considered that the level of alignment between the test items and the main theme is not suitable.

Alignment based on balance of representation. To analyze how the content standards are represented in the Grade 9 Saber Math Test, I calculated an index. This index takes into account the number of performance indicators in each basic area that are related to at least one of the items. The index is calculated taking into account the difference in the proportion of content and the proportion of items related to each one of them. If the index is equal to 1 this means there is a perfect representation and assumes that the number of items related to a basic area are evenly distributed between the contents of each theme. Conversely, if the index is close to 0 this means that only a few performance indicators are being assessed by the test items. Therefore, if the ratio exceeds 0.7, it is considered that the alignment level is adequate. If the index is between 0.6 and 0.7, it is considered that the level alignment is weak and if the index is less than 0.6, it is considered that the level of alignment between the questions and the standards is not suitable.

Comments and perceptions. The comments and perceptions of the judges were qualitatively analyzed using Hatch's (2002) interpretive method. All the comments were transcribed and saved in Word documents. Then I read all the comments and perceptions to establish commonalities and differences that allowed me to make generalizations about problem areas.

To give more credibility and validity to this study, all the judges received training on the coding process. To determine the agreement between judges, I calculated Cronbach's alpha coefficient (Cronbach, 1971). In this study it was determined that the internal correlation in coding between the seven judges was .92 (see Table 2). These correlations are acceptable and suggest that there was considerable agreement among the judges to assign cognitive demand levels between items and performance indicators. The degree of relationship between pairs of judges to align performance indicators and test items was .61 and the internal correlation between the seven judges was .92. This indicates that the training the judges received was good and therefore there is high confidence in the coding they did. Webb et al. (2007) highlight the importance of proper training to judges in alignment studies.

Table 2. Interrater reliability

Grade	Among judges	Between pairs	Number of items	Number of judges
9	.92	.61	15	7

Results

Degree of Alignment at the Content Level

It was found that the Grade 9 Saber Math Test does not have a strong alignment with any of the five basic areas in the standards (see Table 3). Based on these results, I conclude that the Grade 9 Math Test is not aligned to the Colombian Basic Skills Standards in Mathematics in terms of content. This indicates that the number of test items related to each of the five basic areas is not the most appropriate.

Table 3. Alignment based on content

Basic Areas		Associated performance indicators	Alignment based on content
Name	Number of performance Indicators	Number of items	
1 – Numerical reasoning and numerical systems	4	5	Weak
2 – Spatial reasoning and geometrical systems	4	1	Inadequate
3 – Metric reasoning and measurement systems	3	1	Inadequate
4 – Statistical reasoning and data systems	9	4	Inadequate
5 – Variational reasoning and algebraic and analytic systems	9	5.	Weak

Degree of Alignment at the Cognitive Demand Level

It was found that the Grade 9 Saber Math Test has a partially adequate alignment level (see Table 4). The percentage of test items related to four of the five of the basic areas are at least at the same level of cognitive demand as the performance indicators. It should be noted that only one of these themes, "Spatial reasoning and geometric systems,' has one related test item and therefore it is difficult to make valid inferences about the degree of alignment. The only basic area that is not properly aligned in terms of cognitive demand level is 'Metric reasoning and measurement systems'. In this area only 20 percent of the test items do not have at least the same level of cognitive demand as the performance indicators.

Table 4. Alignment based on cognitive demand

Basic areas		Number of associated items	Level of the items in relation to the level of the standards			Alignment based on cognitive demand
Name	Number of performance indicators		Percentage with lower	Percentage with equal	Percentage with higher	
1 – Numerical reasoning and numerical systems	4	5	11	50	39	Adequate
2 – Spatial reasoning and geometrical systems	4	1	0	71	29	Adequate
3 – Metric reasoning and measurement systems	3	1	80	10	10	Inadequate
4 – Statistical reasoning and data systems	9	4	38	45	17	Adequate
5 – Variational reasoning and algebraic and analytic systems	9	5	20	12	68	Adequate

Degree of Alignment at the Range of Knowledge Level

In the Grade 9 Saber Math Test, only one of the basic areas, 'Numerical reasoning and numerical systems' is properly aligned at the Range of Knowledge level (see Table 5). In this area, 57 percent of the performance indicators have at least one related test item. In the other four themes we find that they do not have at least 50 percent of the performance indicators with a related test item. Consequently, we conclude that the coverage is not very good. This indicates that much of the content in the standards are not being assessed.

Table 5. Alignment based on range of knowledge

Basic areas	Number of performance indicators	Number of related items	Range of knowledge		Alignment at the range of knowledge level
			Number of related performance indicators	Percentage of total	
Name			Average	Average	
1 – Numerical reasoning and numerical systems	4	5	2.29	57	Adequate
2 – Spatial reasoning and geometrical systems	4	1	1	25	Inadequate
3 – Metric reasoning and measurement systems	3	1	0.71	24	Inadequate
4 – Statistical reasoning and data systems	9	4	2.71	30	Inadequate
5 – Variational reasoning and algebraic and analytic systems	9	5	3.57	40	Weak

Degree of Alignment at the Balance of Representation Level

Finally, it was also found that the degree of alignment of the Grade 9 Saber Math Test is not appropriate in any of the five basic areas at the balance of representation level (see Table 6). One of the basic areas, ‘Metric reasoning and measurement systems', did not meet the requirements. The index for this basic area is only .55. Another basic area, 'Numerical reasoning and numerical systems,' has an index of 66. Therefore it is concluded that the alignment level is weak. Based on this index, it follows that these two basic areas do not have sufficient representation on the test.

Tabla 6. Alignment based on balance of representation

Basic areas	Number of performance indicators	Index		Alignment based on balance of representation
		% of items related to the basic area /Total related	Index	
Name		Average	Average	
1 – Numerical reasoning and numerical systems	4	31	0.66	Weak
2 – Spatial reasoning and geometrical systems	4	6	0.86	Adequate
3 – Metric reasoning and measurement systems	3	5	0.55	Inadequate
4 – Statistical reasoning and data systems	9	26	0.85	Adequate
5 – Variational reasoning and algebraic and analytic systems	9	32	0.86	Adequate

Judges’ Comments

The judges commented on eight of the fifteen questions. In general, all comments have to do with the lack of alignment between the test items and the contents described in the standards.

Perceptions

Equally importantly, the results described in this section only indicate the perceptions of judges and do not represent a rigorous analysis of the degree of alignment between the skills described in the Grade 9 Basic Skills Standards in Mathematics and the Saber Math Test. The seven judges felt that the test assesses the most important content described in the standards, but that it does not have enough items to make important decisions about the students' skills. Similarly, the judges consider that the level of cognitive demand on the test is comparable to the cognitive level demands contained in the standards. Finally, the judges said that the Grade 9 Saber Math Test and the Basic Skills Standards in Mathematics are partially aligned, but the level of alignment between them can be improved.

Discussion

One of the most important aspects of Webb’s (1997) model is that the alignment between external standardized assessment and standards are analyzed from several perspectives. In general, it was found that the degree of alignment is not the most suitable. In terms of content, it is evident that the degree of alignment is not very strong, but this is an expected result since the number of items on the Grade 9 Saber Math Test is very low. This indicates that test users should be very careful with how they interpret the results, especially in the basic areas where there is not an adequate number of items on the test. The relationship between the content of the tests and content standards can be corrected by increasing the number of items but keep in mind that students are taking other tests on the same day. Therefore, it is important to examine how many items can be added without causing stress or fatigue in students. There is also the possibility that it is not feasible to add more items on the test. If this is the case, it would be necessary to include more complex, integrated items so that one item can be aligned to more than one basic area.

Furthermore, at the cognitive demand level, it was found that the degree of alignment is suitable. These results suggest that the level of rigor of the test is aligned to the expectations described in the performance indicators in the standards and in theory that the Grade 9 Saber Math Test is neither too difficult nor too easy. The most important thing is to have an assessment that measures different types of skills or competencies. The only concern is that the Grade 9 Saber Math Test only uses multiple-choice items. As is generally known, this type of question has many limitations, especially if the intent is to measure highly complex cognitive skills (AERA , APA, NCME , 1999) . Given this limitation, much of the content in the standards can only be assessed in the classroom. Thus, it is very important to examine how these skills or competencies are being assessed in the classroom.

It was also found that there are many contents that are not being assessed in the Grade 9 Saber Math Test, but as mentioned above, this is expected since the number of items on the test is very low and does not allow adequate range of knowledge of the contents in the standards. This indicates that the items on the Grade 9 Saber Math Test are not a representative sample of all content and skills described in the five basic areas. Thus, test users should be cautious when interpreting the results on the test. This also implies that the other contents described in the standards should be assessed in the classroom before making any decision that has to do with the teaching and learning process. Especially if there is a tendency to focus on teaching only what is measured in the assessments, running the risk that many teachers focus on some content and skills in the standards and ignore others (Herman, 2004). This consequence or impact of the tests, though not necessarily intentional, must be avoided at all costs (Herman, 2004; Lopez, 2008; Messick, 1989).

As far as the balance of representation of the content standards being assessed, it was found that the items on the Grade 9 Saber Math Test are represented proportionally. This is an important result, but with the low number of items on the test, this is relatively easy to reach. It is essential to remember that if the number of questions increases or if the coverage is greater, the distribution of assessed content must be proportional, since once assessment results become public knowledge, they may indicate or suggest what should be taught in the classroom and what students should be learning. The possibility exists that unconsciously teachers would start giving higher priority to certain content or skills than others.

Although there is no formula to indicate what the degree of ideal alignment should be between external standardized assessment and standards, it is expected to be strong enough so that test users can make informed decisions about the teaching and learning process. In no way is it expected to have a degree of perfect alignment, as this is almost impossible to achieve due to different factors. In this study, it was found that there are several of them that do not allow the degree of alignment between the Grade 9 Saber Math Test and the Basic Skills Standards in Mathematics to be greater.

The first factor, as mentioned previously, is that the number of items on the Grade 9 Saber Math Test, as in many standardized tests, is relatively very low, which does not allow assessing a representative sample of the performance indicators in each of the five basic areas. Similarly, the type of questions or item type used, multiple choice items, limits the number of performance indicators that can assessed as many of these cannot be measured in this way. For example, some performance indicators require students to produce results based on extensive work, such as projects, and these are very difficult to measure on an external standardized assessment. So some contents are designed to be assessed only within the classroom.

The level of specificity of the standards could also be affecting the connection between the content of the standards and the items on the Grade 9 Saber Math Test. Thus, the judges commented that some contents are global and others are very specific. Some studies have shown that the less specific the standards tend to be, the greater the degree of alignment (López, Webb & Stansfield, 2006). On the other hand, there is lack of clarity in some performance indicators and many of these tend to recur or are very similar. This makes it difficult to write items to assess them and therefore the degree of alignment between those items and the content standards are less obvious. Finally, it was also found that as there is no Saber Math Test at all grades, the possibility exists that an item is measuring contents that are described in lower grades and not necessarily measuring standards for the grade level being assessed.

Based on reviews from the judges, it can also be concluded that it is important to review the standards to verify that the contents are clear and measurable. The low degree of alignment cannot only indicate that there may be problems with an assessment instrument per se, but can also indicate that there might be some problems with the standards. So it is important to remember that during the standards design process standards, developers must think about how each of the performance indicators being proposed can be measured. It is equally important that the standards also have a section that describes how each of the contents described should be assessed. This process could not only facilitate the design of external standardized test systems, like the Saber Math Test, but also assessments that teachers themselves can use in the classroom.

Final comments

The results of this study have several implications. First, designers of standardized assessments may take into account the results of alignment studies to make changes in the assessment and to identify areas or contents that are not being assessed properly. It is very important that test designers always keep in mind the standards when designing test items. For example, test specifications should clearly state the connection between each item and the content described in the standards. These specifications should also provide information on the content, cognitive demand, item type, the way the item should be scored and how to interpret the results. Similarly, standardized test designers could also use the results of an alignment study as evidence of the content validity of the assessment.

Second, it is evident that there is a need to supplement the standardized assessment results with the results of the assessments that are developed in the classroom, as standardized test results do not necessarily show everything that the students have had the opportunity to learn from the curriculum that is being taught in the classroom. Therefore it is very important to continue working with teachers so they can familiarize themselves with the standards and they can use and assess them properly in the classroom. Another important aspect in the teaching and learning process is to train teachers to interpret the results of standardized tests so they can make informed decisions that can improve the quality of education.

Finally, it is critical to stress the importance of interpreting adequately the results of standardized tests. As is it customary, not all the content standards can be assessed in one assessment, so it is important to emphasize that no relevant decisions can be made to change the teaching and learning process based solely on the results of standardized tests. It is necessary to use several criteria, including assessments teachers are using in the classroom, to determine what and how much students are learning. Studies similar to the one presented here provide guidance on how to assess the alignment of a test and call attention to the need for more studies of this kind, based on expert judgment, to address basic aspects of a test’s construct and content validity.

References

AERA, APA & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.). *Educational Measurement (2nd Ed.)*. Washington, D. C.: American Council on Education.

Hatch, J. A. (2002). *Doing qualitative research in education settings*. Albany, NY: SUNNY Press.

Herman, J. (2004). The effects of testing in instruction. En S. Fuhrman and R. Elmore (Eds.), *Redesigning accountability systems for education (pp. 141-166)*. New York: Teachers College Press.

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), 101-176.

ICFES (2008). *Documentos y Marco Legal pruebas SABER*. Bajado el 7 de agosto de 2008, de http://www.icfes.gov.co

Lopez, A. A. (2009). Estudio de alineación de las pruebas Saber de Ciencias. En J. Montoya (Eda.), *Educación para el siglo XXI: Aportes del Centro de Investigación y Formación en Educación, CIFE, 2001-2008 (pp. 551-582).* Bogotá: Ediciones Uniandes.

López, A. (2008). *Potential impact of language tests: Examining the alignment between testing and instruction*. Saarbrucken, Germany: VDM Publishing.

López, A., Webb, N., y Stansfield, C. (2006). *Alignment the New Mexico Language Arts Frameworks and the Spanish Reading Standards-Based Assessments.* Report submitted to the New Mexico Public Education Department. Rockville, MD: Second Language Testing, Inc.

Martone, A. & Sireci (2009). Evaluating alignment between curriculum, assessment and instruction. *Review of Educational Research, 79*(3), 1-76.

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(2), 28-35.

McGehee, J. J., & Griffith, L. K. (2001). Large-scale assessments combined with curriculum alignment: Agents of change. *Theory Into Practice, 40*(2), 137-144.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: Macmillan.

Pehuniak, P. (2004). Educational assessment in an era of accountability. En J. E. Wall & G. R. Walz (Eds.). *Measuring up: Assessment issues for teachers, counselors, and administrators*. Greensboro, NC: CAPS Press.

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

Rothman, R. (2004). Benchmarking and alignment of state standards and assessment. En S. Fuhrman and R. Elmore (Eds.), *Redesigning accountability systems for education (pp. 96-114)*. New York: Teachers College Press.

Ruíz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M. C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching, 49*(6), 691-712.

Subkoviak, M. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47-55.

Webb, N. (2002). *An analysis of the alignment between standards and assessments for three states.* Madison, WI: Wisconsin Center for Education Research.

Webb, N. (1997). *Criteria for alignment of expectations and assessments in mathematics* *and science education. Council of Chief State School Officers and National Institute for Science Education.* Madison, Wisconsin: Wisconsin Center for Education Research, University of Wisconsin.

Webb, N., Herman, J. & Webb, N. (2007). Alignment of mathematics state-level standards and assessment: The role of reviewer agreement. *Educational Measurement: Issues and Practice*, 26(2), 17-29.

Wise, L. (2004). Vertically-articulated content standards. Bajado el 17 de julio de 2008, de http://www.nciea.org/publications/RILS_LW04.pdf.

APPENDIX

Appendix 1

Consensus about the cognitive demand of the performance indicators in the Grade 9 Basic Skills Standards

Level	Description	CDL
1	Numerical reasoning and numerical systems	2
1.1	Use real numbers in their different representations and in various contexts	1
1.2	Solve problems and simplify calculations using properties and relationships of real numbers and the relationships and transactions between them	2
1.3	Use scientific notation to represent measures of quantities of different magnitudes	1
1.4	Identify and use powers, roots and logarithms to represent mathematical and non-mathematical situations and to solve problems	2
2	Spatial reasoning and geometric systems	3
2.1	Guess and verify properties of congruencies and similarities between two-dimensional shapes and three-dimensional objects in the solution of problems	3
2.2	Acknowledge and contrast geometric properties and relationships used in basic theorem proving (Pythagoras and Tales)	2
2.3	Apply and justify congruence and similarity criteria between triangles in the resolution and formulation of problems	3
2.4	Use geometric representations to solve and formulate problems in mathematics and other disciplines	3
3	Metric reasoning and measurement systems	2
3.1	Generalize valid calculation procedures for finding the area of plane regions and the volume of solids	3
3.2	Select and use tools and techniques to measure lengths, surface areas, volumes and angles to appropriate levels of precision	2
3.3	Justify the appropriateness of using standardized measurement units in situations taken from different sciences	2
4	Statistical reasoning and data systems	2
4.1	Recognize how different ways of presenting information can cause different interpretations	2
4.2	Analytically and critically interpret statistical information from various sources (newspapers, magazines, television, experiments, queries, interviews)	3
4.3	Interpret and use concepts of mean, median and mode	3
4.4	Select and use appropriate statistical methods depending of kind of problem, information and the level of the scale at which this is represented (nominal, ordinal, interval or ratio)	2
4.5	Compare results from randomized experiments with expected results for a probabilistic mathematical model	2
4.6	Solve and formulate problems selecting relevant information in data sets from various sources (newspapers, magazines, television, experiments, queries, interviews)	3
4.7	Recognize trends in sets of related variables	1
4.8	Calculate the probability of simple events using different methods (list, tree diagrams, counting techniques)	2
4.9	Use basic concepts of probability (sample space, event, independence, etc.)	1
5	Variational reasoning and algebraic and analytic systems	2
5.1	Identify relations between the properties of graphs and the properties of algebraic equations	1
5.2	Build algebraic expressions equivalent to a given algebraic expression	2
5.3	Use inductive and algebraic language processes to formulate and test conjectures	3
5.4	Model variation situations using polynomial functions	3
5.5	Identify different methods to solve systems of linear equations	1
5.6	Analyze the processes underlying the infinite decimal notation	2
5.7	Identify and use different ways to define and measure the slope of a curve representing the variation situations on a Cartesian plane	2
5.8	Identify the relationship between changes in the parameters of the algebraic representation of a family of functions and changes in the graphs that represent them	1
5.9	Analyze in Cartesian graphical representations the change behaviors of specific functions belonging to families of polynomial, rational, exponential and logarithmic functions	2

CDL – Cognitive Demand Level

APPENDIX 2

Coding Form to Code Test Items

Judge: ____________________________ *Date:* ____________

Item Number	CDL of Item	Primary Performance Indicator	Secondary Performance Indicator	Secondary Performance Indicator	Comments
24					
26					
27					
28					
29					
30					
40					
41					
42					
43					
44					
50					

CDL – Cognitive Demand Level

ABOUT THE AUTHORS / SOBRE LOS AUTORES

Lopez, Alexis A. (alopez@ets.org). Doctor of Education, University of Illinois at Urbana-Champaign. Currently is Associate Research Scientist at Educational Testing Service – ETS, in Princeton, New Jersey. He previously worked as an associate professor of the Center for Research and Training in Education (in Spanish, *Centro de Investigación y Formación en Educación* -CIFE) at University of the Andes (Colombia), where he also directed the Center for Evaluation. His postal address is: Educational Testing Service. 600 Rosedale Road. MS-R04. Princeton, NJ 08901 (USA). *Buscar otros artículos de esta autora en Google Académico / Find other articles by this author in Scholar Google*

ARTICLE RECORD / FICHA DEL ARTÍCULO

Reference / Referencia	Lopez, Alexis A. (2013). Alignment between standardized assessments and academic standards: The case of the Saber Mathematics Test in Colombia. *RELIEVE*, *v. 19* (2), art. 3. DOI: 10.7203/relieve.19.2.3026
Titlele / Título	Alignment between standardized assessments and academic standards: The case of the Saber Mathematics Test in Colombia. [*Alineación entre las evaluaciones externas y los estándares académicos: El Caso de la Prueba Saber de Matemáticas en Colombia*].
Authors / Autores	Lopez, Alexis A.
Review / Revista	RELIEVE (Revista ELectrónica de Investigación y EValuación Educativa), *v. 19*, n. 2
ISSN	1134-4032
Publication date / Fecha de publicación	2013 (**Reception Date**: 2012 November 17 ; **Approval Date**: 2013 September 17. **Publication Date**: 2013 September 18)
Abstract / Resumen	*In this study the concept of alignment is analyzed by presenting the results of a study that examines the degree of alignment between a standardized mathematics assessment and a set of academic standards. The alignment was examined using the Webb model and the results suggest that the degree of alignment between standardized assessments and academic content standards is not the most appropriate and therefore the results should be interpreted with great caution and stakeholders must be very careful with the type of decisions that are made based on the results of these types of assessments.* En este estudio se analiza el concepto de alineación presentando los resultados de un estudio que examina el grado de alineación entre una evaluación estandarizada de matemáticas y unos estándares académicos. La alineación se examinó usando el modelo de Webb y los resultados sugieren que el grado de alineación entre las evaluaciones estandarizadas y los estándares académicos no es el más adecuado y por lo tanto los resultados se deben interpretar con mucha cautela y se debe poner mucho cuidado con el tipo de decisiones que se toman con base en los resultados.
Keywords / Descriptores	*Alignment, Standardized Tests, Academic Standards, Assessment of learning, Validity.* Alineación, Pruebas estandarizadas, Estándares Académicos, Evaluación del aprendizaje, Validez.
Institution / Institución	Educational Testing Service (USA).
Publication site / Dirección	http://www.uv.es/RELIEVE
Language / Idioma	Español & English version (Title, abstract and keywords in English & Spanish)

RELIEVE

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa

E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]