Inter-American Development Bank

Education Division (SCL/EDU)

TECHNICAL NOTE

No. IDB-TN-738

How to select an instrument for assessing student learning

Catalina Covacevich

December 2014

How to select an instrument for assessing student learning

Catalina Covacevich

Inter-American Development Bank

2014

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Covacevich, Catalina.
How to select an instrument for assessing student learning / Catalina Covacevich.
p. cm. — (IDB Technical Note ; 738)
Includes bibliographic references.
1. Education. 2. Educational tests and measurements. I. Inter-American Development Bank. Education Division. II. Title. III. Series.
IDB-TN-738

http://www.iadb.org

The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent.

The unauthorized commercial use of Bank documents is prohibited and may be punishable under the Bank's policies and/or applicable laws.

Copyright © 2014 Inter-American Development Bank. All rights reserved; may be freely reproduced for any non-commercial purpose.

How to select an instrument for assessing student learning

Catalina Covacevich[1]

Executive Summary

The implementation of educational policies and practices is closely linked to the assessment of student learning, as it enables the monitoring of progress and achievements, the improvement of teaching in the classroom, the improvement of policies and the assessment of program effectiveness, among many other objectives. To ensure that assessment achieves its purposes, it is essential to make an appropriate choice of the learning assessment instruments to be used. In this context, teachers, policy implementers, researchers and staff of the ministries of education are frequently faced with the need to select tools for learning assessment, without necessarily having an extensive knowledge of the subject. Therefore, this technical note is aimed at people working in the education sector who do not have training in the area of learning assessment, and its objective is to provide technical, practical and ethical guidance on the elements that must be taken into account when selecting or constructing a learning assessment instrument.

JEL classifications: I200, I210, I280
Keywords: Education-Assessment-Instruments-Tests

[1] The author appreciates the support of Daniela Jiménez in obtaining the bibliography, the valuable contributions of Hugo Ñopo (SCL/EDU), and the comments and guidance during the preparation of the document received from Emiliana Vegas (SCL/EDU), Head of the Education Division.

Content

Introduction

The Sector Framework Document for Education and Early Childhood Development of the Inter-American Development Bank stresses the need for quality assurance education systems to define high learning goals that guide their work. The definition of these goals is as important as their monitoring, which can be performed at different levels of the system, whether they are large scale assessments, such as subnational, national, regional or international ones; or local assessments, implemented by those responsible for the schools or by teachers for formal, diagnostic, or training purposes. All these assessments are inputs to monitor learning, inform teaching, and incorporate this information in updates of existing policies and practices and the design of new ones. Therefore, assessment is an inherent part of a continuous improvement process (Inter-American Development Bank, 2013).

In addition, whenever educational policies and programs are implemented, and especially when there are limited resources, a crucial question is whether the programs being implemented are effective. Therefore, educational assessment has become increasingly important in this context. In recent years impact assessments, which seek to measure the impact of a program by comparing equivalent groups, of which one was beneficiary of the program and the other not, have received special interest.

Student learning is also assessed for accountability, certification of competencies obtained in a given education cycle, identification of the level of knowledge of a student to assign him or her the most appropriate training, and academic selection.

The effectiveness of all these assessments depends in large part on the adecuacy and quality of the student learning assessment instruments that are used. In order to decide how to measure learning, various factors must be considered and the pros and cons of each available instrument must be weighed, to then decide which one is the most suitable for the specific situation. It is unlikely that there is an ideal instrument, which is why one has to be prepared to compromise. To make the best decision, one must fully review the greatest amount of available instruments and collect sufficient information on each instrument (Center for Assessment and Research, James Madison University, 2014).

This paper is addressed to all the teachers, politicians, designers and implementers of educational programs who are faced with the need to choose an instrument for learning assessment[2] for any of the aforementioned goals. It emphasizes large-scale standardized instruments, but most of the principles are also applicable to instruments designed and applied locally. Also, although the emphasis is on student learning assessment instruments, many of the elements discussed are applicable to other types of instruments, for example teaching assessment instruments.

Some of the questions that often arise during the process of choosing an instrument and that will be addressed in this article are: Why do we want to evaluate? Does the instrument fit my needs? What does the instrument measure? How useful is this instrument compared to others in terms of cost and saving time? Are the scores consistent? How easy is it to administrate? Does it have cultural or gender biases? (Cohen & Swerdlik, 2009; Bart, 2009). One last important question is

[2] In this article the term "learning" is used generically, covering skills, competencies, mastery of contents and achievement of objectives.

whether the necessary test is commercially available or if it will be necessary to create a new test.

The first section of the note, *Consistency between the Assessment Objectives and the Chosen Instrument*, stresses the importance of the consistency of the selected instrument with the objective of the assessment, describes the various evaluative purposes that exist, and identifies the characteristics of the instrument that must be compared with the objective in mind. Section II, *Instrument Quality*, addresses the importance of the technical quality of the instruments, describing the two main technical characteristics that need to be kept in mind: reliability and validity. Section III, *Practical Considerations*, describes important practical topics to take into account, such as the cost of the instrument, that it should be of an appropriate length and easy to administer and score. Section IV, *Types of Instruments*, presents different ways of classifying learning assessment instruments, describing the different types of instruments that exist according to each classification. Section V, *Ethical and Fairness Considerations of the Instruments*, discusses a number of items relating to ethics, bias and fairness in learning assessment instruments and how these concepts relate to validity. The last section, *Creating an Instrument or Using an Existing One*, offers a brief discussion on determining if there is an instrument that is suitable for the evaluative situation, or if it is more suitable to design a new instrument, and discusses the pros and cons of both situations. The paper closes with conclusions.

Did you know...

It is believed that tests and assessment programs emerged in China in 2200 B.C. They were used to select candidates who were applying for government jobs. They covered topics as broad as archery, music, writing, mathematics, geography, and military strategy (Cohen & Swerdlik, 2009).

I Consistency between the Assessment Objectives and the Chosen Instrument

A key consideration for choosing a learning assessment instrument is that it must be suitable for the assessment's purpose. This section begins by describing some of the possible goals of assessment, then addressing the elements that should be reviewed in an instrument to see if they are consistent with the assessment objective.

A. Why do we want to assess

Determining the purpose or goal of assessment involves asking the questions: Why do we want to measure learning? What do we want to measure? And whom do we want to evaluate? There are many possible reasons to assess learning. Below are some of them, grouped in evaluative objectives for the entire educational system, the school or the student.Figure 1: Conceptual framework

Obtaining information at the educational system level

Taking a snapshot of how an educational system is doing. In order to be able to make decisions on education policy, an educational system, either at the national level, state or municipal, may need information on how well students are fulfilling the proposed learning objectives. To do so, the best option is to design a test that assesses the national (or local) curriculum in the subjects and grades that are considered most important. In general, these tests are given at the end of the school calendar, to measure the expected accomplishments for each cycle, but an intermediate measurement can sometimes be useful to find out the degree of progress in the learning goals and to be able to intervene in time, for example in the early stages of literacy acquisition.

If the goal is only to obtain aggregate information on what is happening with the set of schools or students, a sampling assessment is sufficient, i.e., it is not necessary to administrate the test to all the schools or students, but only to a representative sample, which makes the evaluation cheaper and simplifies logistics processes. If specific information is required at the school or student level, and not only at the level of the educational system as a whole, the assessment should be census-based, i.e., applied to all the students and schools.

In general, national tests take several months to deliver results, so the information is available the next school year rather than immediately.

Comparison with other countries. Sometimes, instead of wanting to assess the performance of a country's students compared to its national curriculum, one wishes to have information on how much the country's students are learning in comparison with those of other countries. To achieve this, a country or national subsystem can participate in international studies such as PISA,[3] TIMSS,[4] PIRLS[5] and the PERCE, SERCE and TERCE[6] tests of the Latin American Laboratory of Evaluation of School Quality. All of these studies evaluate student learning at the school level in a set of countries (or national subsystems), which allows for comparison within an assessment

[3] Programme for International Student Assessment
[4] Trends in International Mathematics and Science Study
[5] Progress in International Reading Literacy Study
[6] First, Second and Third Comparative and Explanatory Regional Study

framework known and agreed upon by the participating countries. Each study measures different subjects, in different grades and with different emphases. For example, both PISA and TIMSS assess maths and science, but while TIMSS does so with 4th and 8th grade students and with a curricular focus, PISA does so with 15-year-old students with a focus on life skills. In the case of TERCE, only Latin American countries are assessed, while in the other aforementioned studies countries from different continents participate.

These studies also collect information from students and school systems that allow for the identification of the variables that most affect learning, and also serve to compare the national curriculum with the study's assessment framework of the study, which sometimes leads to adjustments in the curriculum.

These are sampling studies as data is obtained at the system level and not at the level of each school. They are managed by renowned international agencies, such as the Organization for Economic Cooperation and Development (OECD), the International Association for the Evaluation of Educational Achievement (IEA), and the United Nations Educational, Scientific and Cultural Organization (UNESCO). They are developed, administered and analyzed under strict standards of quality that ensure that the countries' results are comparable.

In general a country's participation in an international study takes at least three years. The first year, the instruments are designed and piloted, the following year they are administrated and in general it is the year after that when the databases are released and the international report is issued. Since immediate results are not obtained, these studies are useful for policy design in the medium and long term, but not in the short term.

Measuring the evolution of a system over time. Both the international and national assessments provide a snapshot of the status of learning outcomes at a certain point in time, but they may also serve to monitor learning progress over time. For this reason these tests are administered on a regular basis, generally on a yearly basis in the case of national tests, and in cycles of three or more years in the case of international tests. This enables the monitoring of progress in each country's average learning and also enables the identification of other trends, such as if gender gaps or gaps between socio-economic groups have decreased, or if there have been evident improvements in certain subject areas or in certain types of schools after certain interventions, such as changes in the curriculum or interventions in certain groups of schools. International studies can also compare the progress (or lack of progress) of a country with the progress of other countries during the same period.

Obtaining information at the educational system level

On other occasions, one wishes to obtain information on the learning outcomes that are being achieved at the individual school level. This may be for a variety of objectives, such as giving feedback to the teachers and management team to improve learning, the accountability of those who are responsible for school performance, informing the parents and community of a school's performance, or assessing the impact of specific programs in certain schools. School level evaluations are often conducted by the national or local government, but they can also be executed by other agencies, such as universities contracted to assess the effectiveness of a

program, or even a school or group of schools who wish to assess themselves in order to monitor their performance and improve their practices. Depending on the objective, it may or may not be important for the evaluation to enable comparisons among the assessed schools.

The content of assessments at the school level varies according to the assessment objective: in the majority of cases, the goal will be to assess the national curriculum, but in others it may be pertinent to focus in more detail on a specific element of the curriculum known to have weaknesses, or in the case of a program evaluation, it may be pertinent to focus on specific elements that this program sought to promote.

Sometimes national tests are designed to provide school-specific information; in other cases it may be necessary to design special tests, either because this information is not available or because the focus of the national assessment is not adequate for the assessment purpose.

Obtain information at the school level

In some cases student-level information is desired. This may be necessary as part of a continual system improvement process, to identify the weaknesses of a particular student so that his/her teachers and parents can support him or her. In other contexts, tests are administered to validate learning, for example end of cycle tests, such as a secondary school diploma or a professional examination, or a test certifying a level of proficiency in a foreign language. Individual learning instruments can also be used for student selection, for example, university admission tests. On other occasions, students are evaluated at the beginning of a school year or educational program in order to have a diagnosis of their strengths and weaknesses and reinforce the necessary elements or offer leveling courses.

B. Aligning the objectives of the assessment and the instrument

A key element when choosing a learning assessment instrument is its suitability for the objectives of the assessment. This alignment is crucial because it affects the usefulness of the information to be obtained. If the alignment is low, the assessment results will yield little or limited information (Center for Assessment and Research, James Madison University, 2014). Therefore the instrument's purpose, content and target population must be reviewed to ensure that they are aligned to the assessment's purpose.

Purpose alignment

A first question to ask is what the purpose of the test is and if this fits with the purpose of assessment (New Zealand Ministry of Education, 2014). An instrument may have several purposes, such as to perform a diagnosis, measure achievement, measure potential or aptitude, or identify preparation for a certain program or school stage (placement testing), which can also be used to place a student in a certain program or learning track (Cohen, Manion & Morrison, 2000).

Some instruments are designed to perform diagnostic, formative or summative assessments. A diagnostic assessment is an in-depth assessment of a student's strengths and weaknesses. In general it includes many items that delve deeply into a single topic to accurately identify the learning difficulties, and is criterion-referenced.[7] The formative assessment, on the other hand, takes place during a school year or program and is designed to monitor the student's progress during this period, to measure achievements of specific sections of the curriculum, and to diagnose strengths and weaknesses. It is generally criterion-referenced. The summative evaluation is administered at the end of the course or program, and is designed to measure achievements or outcomes. It can be norm-referenced or criterion-referenced, depending to some extent on how the assessment will be used (e.g., to award certificates or degrees) (Cohen, Manion & Morrison, 2000). Therefore, if the goal of assessment is to make a diagnosis, probably an instrument designed as a formative assessment is not the most appropriate.

Content Alignment

A second question is what the instrument measures and if this fits with what is going to be assessed. This analysis is not something general, like, for example, " maths skills in elementary school." One must examine in detail the content of the test, the levels of learning covered, and the age groups that it is designed for, and compare it carefully with the assessment objectives. For example, the purpose of an assessment might be to assess the implementation of the national maths curriculum, which may or may not have elements in common with a maths test designed for universal use, since the curriculum may measure geometry learning, while a test may measure learning in arithmetic. It is very important to analyze the content of the test in detail and see if it properly covers the contents that are going to be assessed, and also to ensure that it does not assess items that are not part of the assessment target. This analysis is addressed again in the section on arguments concerning content-based evidence in the section on instrument validity.

It is also necessary to examine the instrument content in detail to ensure that it is appropriate for the cognitive level, reading level, and other skills of those being assessed. For example, tests that require an advanced reading level cannot be administered to users with little reading comprehension (Timmons et al, 2005).

Alignment with the target population

A third element to consider is the coherence between the population for which the instrument was designed and the population to be assessed. In the case of norm-referenced tests, the population for which these norms were designed should also be reviewed (New Zealand Ministry of Education, 2014).

When assessment was beginning, although many of the published instruments were designed expressly for a specific population, these were administered - improperly - to people from

[7] In the "Types of Instruments" section the difference between criterion-referenced and norm-referenced assessment is explained.

different cultures. And, not surprisingly, those assessed belonging to minority groups tended to obtain scores lower than those for whom the instrument was designed. For example, a question on the *Wechsler Intelligence Scale for Children* (WISC) test from 1949 asked: "If your mother sends you to the store for a loaf of bread, and there is none, what do you do?" Many Latino children went often to buy tortillas, but not bread, and therefore did not understand the question or know what to answer (Cohen & Swerdlik, 2009).

Today, the developers of instruments take many precautions so that they are appropriate for the population for which they were designed. For example, a test designed for national use must be, in effect, appropriate for the national population. These precautions may include running a pilot version with students with different characteristics, asking the test administrators for their impressions (for example, their subjective opinions on the quality of the instructions), analyzing the items to see if they have a racial, cultural, or gender bias, or asking a panel of experts to review the items to look for possible bias (Cohen & Swerdlik, 2009). Students with special educational needs may also have difficulties. This issue is addressed in the section on ethics.

Sometimes there are adapted versions available of instruments that were originally designed for another population. Often, in Latin American countries, adapted versions of instruments designed in the United States are used. These adaptations include the construction of norms for the population of the specific country. These may have been constructed by the test designers themselves, or on other occasions, by local universities or other research institutions. Sometimes, it is possible to rely on the version adapted for a neighboring country, but the items should be thoroughly reviewed and a pilot test should be implemented to see if the adaptations are appropriate for the specific population requiring assessment.

II. Instrument Quality

Another important aspect that arises when analyzing an instrument is the technical quality. The better the quality of an instrument, the more useful it will be, the greater the confidence in the scores and the greater the confidence in making decisions based on these results. Therefore, it is imperative to use high-quality instruments to perform assessments (Center for Assessment and Research, James Madison University, 2014). The two main elements that account for the quality of an instrument are its validity and reliability.

A Validity

Evolution of the validity concept

The concept of validity has been undergoing transformations over time. Traditionally, the validity of an instrument has been understood as the extent to which the instrument does in fact measure what its authors claim that it measures (Cohen, Manion & Morrison 2000; Darr, 2005). However, more recently assessment specialists have asserted that validity is not a fixed and inherent property of an instrument, but rather an evidence-based trial about how appropriate the inferences or actions implemented from the scores of a test are in a given context (Salvia & Ysseldyke, 2004; Cohen & Swerdlik, 2009). Validation can be seen as the development of a strong validity argument for the proposed uses of the scores of an instrument and its relevance to the proposed use (Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association and National Council on Measurement in Education, 1999).[8] Defined in this way, validity is not an inherent property of the instrument; instead, it is related to the assessment objective.

Since what is judged is not in reality the validity of the instrument and its scores, but the interpretation of the results of the test for certain purposes, when a test is intended to be used in a number of ways, the validity of each use must be analyzed separately (Joint Committee on Standards for Educational and Psychological Testing, 1999).

As a result, if a test is valid for a particular group or population, it will not necessarily be so for others. For example, if a spatial reasoning test designed for students with an 8th grade reading level is administered to a student with 4th grade reading ability, its results may reflect both his or her reading level and spatial reasoning ability.

This new way of understanding validity is closely related to program assessment, in that it is necessary to specify the program to be assessed, and the contexts in which it will be implemented. One must rule out external variables that may affect the results, and programs are

[8] The explicit statement of inferences and assumptions based on applications and proposed interpretations has been called interpretative argument. And the assessment of the consistency of the interpretative argument and the plausibility of its inferences and uses has been called validity argument. Therefore, it is necessary to specify the interpretation and uses (Brennan, 2006).

often assessed more with a set of assessments than with an isolated study. In the same way, validity can be understood as a comprehensive assessment of the assessment's proposed uses, generating a coherent analysis of all the evidence in favor of or against a particular use and if possible, of alternative explanations (Cohen & Wollak, 2006).

Some questions concerning a test's validity inquire about the quality of the questions: Are the items an appropriate sample of the construct to be assessed? There are also questions concerning the interpretation of the results: What do these scores tell us? What is the relationship between low or high scores and the behavior of those assessed? How do these scores compare to other instruments that claim to measure the same thing? (Cohen & Swerdlik, 2009).

Validity should be a priority when designing or selecting instruments for program assessment. A critical element of the assessment is that is must allow its users to make robust, useful judgments regarding student progress that can have a positive impact. Being aware of validity and how it can be threatened can help in decision making about what assessments are worth administering and what uses can be given to these assessments (Darr, 2005).

Who is responsible for validity?

It is the responsibility of the test's developer to provide evidence on the validity of the instrument, specifying the population in which it was validated. But it is the user's responsibility to assess whether the instrument is appropriate for the particular context in which it will be applied. Occasionally, it may be appropriate for the user to conduct extra local validation studies. This local validation becomes essential when planning to make changes to the instructions or language of the instrument, if the intention is to administer the instrument to a population that is significantly different from that in which the test was standardized, or if one wishes to give it a different use from that for which it was designed (Joint Committee on Standards for Educational and Psychological Testing, 1999; Cohen & Swerdlik, 2009).

How is validity measured?

Judging whether an instrument is valid is not something that can be measured on an absolute scale. Validity is often categorized as weak versus acceptable, which reflects a judgment about how properly the test measures what it is supposed to measure (Cohen & Swerdlik, 2009). Other authors, like Darr (2005), suggest categorizing validity as weak, moderate, or strong.

Since validity is related to inferences and decisions made for a specific group within a specific context, judging the validity of an instrument requires collecting a large amount of data (Darr, 2005). There are different approaches designed to test different types of validity. These approaches are not mutually exclusive, and all contribute to the overall validity, although based on the use the test will be given, they may have different levels of importance (Cohen &

Swerdlik, 2009). Classically, one speaks about construct, content, and criterion validity. Different authors have different classifications of types of validity that should be considered, for example validity of construct, content, item, predictive, "face," criterion-referenced, concurrent, etc., making clear that it is not necessary to always use all the forms of validity (Wilson, 2005).

This document will use the approach proposed in the Standards for Educational and Psychological Testing (Joint Committee on Standards for Educational and Psychological Testing, 1999). Rather than talk about types of validity, they propose referring to types of evidence on validity, or lines of evidence, based on the content of the test, the response processes, internal structure, relationships with other variables, and the consequences (Joint Committee on Standards for Educational and Psychological Testing, 1999; Wilson, 2005). These types of evidence are described in the following table.

Table 1. Types of evidence on the validity

Evidence based on the test content	The tests cannot evaluate all of the students' knowledge, but instead only a sample. Therefore it is very important for the sample to be an appropriate sample of the area of learning that is of interest to assess. If this is achieved, it increases the possibility of making valid inferences on the learning achievements in a certain domain (Darr, 2005). This type of evidence requires looking at the content of the instrument to analyze the relation to the construct to be measured (Joint Committee on Standards for Educational and Psychological Testing, 1999; Cohen & Swerdlik, 2009). To know what a test measures, it is not enough to be guided by the name; it is essential look at the items that constitute it (Center for Assessment and Research, James Madison University, 2014). One can analyze each item in relation to the domain, or ask the opinion of experts on how well they address the domain (Joint Committee on Standards for Educational and Psychological Testing, 1999). If an instrument is good, it will have items that assess different aspects of the topic, and experts in the field, who are not familiar in advance with the items, will agree on what each item assesses. There are two risks that should be avoided. One is the under-representation of the construct, i.e., important elements of the construct that one wishes to assess are not being assessed. The other is the variance related to constructs that are irrelevant to what is being measured, for example on a reading test, prior knowledge of the subject or the emotional response to the text, or on a maths test, reading speed or vocabulary (Joint Committee on Standards for Educational and Psychological Testing, 1999).
Evidence based on the response processes	The theoretical analysis and empirical evidence on the response processes of the examinees can provide information about the relationship between these processes and the constructs to be assessed. For example, if a test seeks to assess mathematical reasoning, it is important for the test to actually assess that and not simply the application of algorithms. Observing response strategies or interviewing the examinees about the processes can provide this information (Joint Committee on Standards for Educational and Psychological Testing, 1999).
Evidence based on the internal structure	This analysis seeks to collect evidence on the degree in which relations between the items of a test and its components are suited to the construct that it allegedly seek to evaluate, which can involve a single dimension, or several. To see this, one should check if the items do in fact comply with the content map. If the construct has just one dimension, this can also be tested through the item analysis (for example, whether students with a good performance on the total test have a good performance in the item). Another method is to check that the items work differently in different groups, according to what the theory predicts (Joint Committee on Standards for Educational and Psychological Testing, 1999).

Table 1. Types of evidence on the validity (continued)

Evidence based on the relations to other variables[9]	This type of evidence is broken down into convergent[10] and discriminatory validity. The evidence relating to convergent validity involves comparing the results obtained from a given test with those obtained by the students themselves on tests that measure the same construct or similar constructs. It is expected that the scores of a certain instrument would correlate with others that claim to measure equal or similar constructs (Wilson, 2005; Joint Committee on Standards for Educational and Psychological Testing, 1999): if two assessments that supposedly measure the same construct are delivering very different results, this is cause for concern (Darr, 2005). A potential difficulty is that many times there are no other similar instruments (Wilson; Joint Committee on Standards for Educational and Psychological Testing, 1999). The evidence on the discriminatory validity is obtained comparing the results obtained on the test with other assessments that measure opposite or different constructs. In this case, it is expected that the scores would have little correlation with the results of tests that claim to measure different constructs (Wilson, 2005; Joint Committee on Standards for Educational and Psychological Testing, 1999).
Evidence based on the consequences of testing	Beyond all the technical information collected, if the use of a particular assessment has or can have negative consequences, or the consequences of using its results can go against the final educational objective, this is a consideration that must be taken into account in order to question the validity of an instrument and decide whether to use it or not (Darr, 2005; Wilson 2005; Joint Committee on Standards for Educational and Psychological Testing, 1999). This is validity from the point of view of the consequences of using the test results. For example, the weight given to the results can have an impact on ways to teach and learn. Negative consequences may include curriculum narrowing, "teaching to the test," or reduction in student motivation (Darr, 2005). To analyze this type of evidence it is necessary to consider both the intentional and the unintentional effects of tests (Wilson 2005; Joint Committee on Standards for Educational and Psychological Testing, 1999). It is also necessary to examine whether the undesirable consequences are due to the construct being measured, or the specific instrument being used to do so. To elucidate this, one should see if another instrument that measures the same construct has the same undesirable consequences. If this is the case, the problem is more likely due to the construct than the instrument (Wilson, 2005). Finally, one must distinguish the consequences that have to do with education policy decisions but not necessarily with validity. In general, consequence-related evidence is directly related to validity if it has to do with the under-representation of a construct or the construct irrelevance described above (Wilson, 2005; Joint Committee on Standards for Educational and Psychological Testing, 1999).

B Reliability

What is reliability?

Reliability is the consistency with which the instrument measures, or in other words, the degree of error in measurement (Cohen & Swerdlik, 2009). A reliable test delivers consistent results over time. For example, students with the same level of reading comprehension who take a

9 Some authors identify this type of validity as external.

10 In the traditional approach, one speaks of concurrent validity, that is, the degree to which a test score is related to another measure obtained at the same time, i.e. how much the scores of a test that measures 'x' relate to other tests that measure the same thing. Predictive validity refers to the degree to which the score on a test predicts a criterion-referenced behavior measured by another test in the future (Cohen & Swerdlik, 2009; Darr, 2005). For example, at the level of secondary or tertiary studies, a predictive assessment might make sense to see how well school-wide assessments predict future academic or work performance (Darr, 2005).

reading comprehension test have scores that are similar or identical, regardless of when they take it, assuming that their reading comprehension level has not changed (Timmons, Podmostko, Bremer, Lavin & Wills, 2005).

In theory, a perfectly reliable instrument (for example, a thermometer) always measures in the same way (Cohen & Swerdlik, 2009). However, in reality educational assessment is never free of some degree of error, since the same individual does not always perform the same way and external conditions can also lead to error (Joint Committee on Standards for Educational and Psychological Testing, 1999).

To look at reliability, as well as validity, one must understand them in specific evaluative purposes and contexts. However, since reliability refers to how much variation is expected from one measurement to another, it is understood in a more strictly statistical way than validity, which deals which the nature of the attributes being measured[11] (Haertel, 2006).

Who is responsible for reliability?

The web sites or manuals of instruments must specify their reliability. If they do not, their results should be taken with great care and not used to make high impact decisions (Timmons et al, 2005).

The information that must be present is the identification of the main sources of error, statistical summaries that quantify the size of these errors, and the degree of generalizability of scores between different forms, scorers, administrators, and other relevant dimensions. There should also be a description of the population with which these estimates were made. There must be enough detail to judge if the reliability is appropriate, because there is no single index applicable to any situation (Joint Committee on Standards for Educational and Psychological Testing, 1999)[12].

How is reliability measured?

There are a number of major statistical theoretical frameworks that have been developed to analyze reliability. The main ones are classical measurement theory, generalizability theory, and item response theory (IRT) (Haertel, 2006).

---

11 Since this document is aimed at people who do not necessarily have statistics and psychometrics knowledge, it only describes methods to estimate reliability in a very general way. To delve further into the subject, see "Reliability" by Haertel, E. and "Item Response Theory" by Yen, W. & Fitzpatrick, A., both in the book Educational Measurement, (4th edition). 2006. Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on Education. Praeger Publishers, Westport.

12 It is important to see if the analyses were carried out using raw scores or not.

Depending on which of these approaches is used, reliability is calculated in different ways and the information can also be reported in many different ways: as variance or standard deviations of errors of measurement, as one or more coefficients, or as functions of IRT (Joint Committee on Standards for Educational and Psychological Testing, 1999; Cohen & Swerdlik, 2009). These different approaches are described in the following table.

Table 2. Theoretical frameworks to analyze reliability

Classical Test Theory	In classical theory, the typical approaches to analyzing reliability are: coefficients derived from the management of parallel forms in independent sessions, coefficients obtained by administering the same instrument on separate occasions (also known as " test re-test," or "stability coefficient") and coefficients based on the relationship between scores derived from individual items or subtests within a test, information that is obtained from the same administration (also known as "internal", or "inter item" coefficient) (Joint Committee on Standards for Educational and Psychological Testing, 1999; Cohen & Swerdlik, 2009). The reliability coefficient most used in classical theory is Cronbach's Alpha, which belongs to this last category. Alpha was developed in 1951 to provide a measure of the internal consistency of a test or a scale, i.e., to identify how much the items measure the same concept. Therefore if a test has several scales it may be more appropriate to use alpha separately for each scale. If the items are correlated among themselves, the alpha value increases. But this value may also increase due to the amount of items (Webb, Shavelson & and Haertel, 2006). Its possible values range between 0 and 1. In general, an alpha of .7 or more is considered acceptable (Institute for Digital Research and Education, UCLA, n.d.), for example for program assessment (Center for Assessment and Research, James Madison University, 2014). But if the results will have individual consequences it is best to obtain values greater than .8 (Webb, Shavelson & Haertel, 2006).
Generalizability Theory	Classical theory assumes that the observed score is the sum of the true score and some specific residual error of that score. In contrast, the theory of generalizability, instead of using the real score, assumes a universe of generalization consisting of all possible observations considered equivalent (Brenan, 2006, Haertel, 2006). The coefficients used by the theory of generalizability allow the user to specify and estimate the different components of the true score variance, error variance, and variance of the observed score (Joint Committee on Standards for Educational and Psychological Testing, 1999). Two types of studies, generalizability (G-Study) and decision (D-Study) can be performed. A commonly used analysis tool is ANOVA, as well as the computer program GENOVA.
Item Response Theory (IRT)	IRT is a family of statistical models used to analyze test item data, providing a unified statistical process to estimate characteristics of items and individuals examined, and define how these characteristics interact in the performance on the items and the test. IRT has many possible uses in assessment, including item construction, scaling, equating, standard setting, and scoring. Since the 1990s it has been used in the majority of large scale student assessments. There are different IRT models but their common essence is a statistical description of the probability that an examinee with certain characteristics will have a determined response to an individual item, which, in turn, has special characteristics. The ways of calculating reliability under IRT take into account the characteristics of the individual and of the items (Yen & Fitzpatrick, 2006). When using IRT the function of test information is often used as a reliability measure. This summarizes how well the test discriminates between individuals of different levels in the feature being evaluated (Joint Committee on Standards for Educational and Psychological Testing, 1999). IRT

These three approaches are concerned with the instrument's reliability, but the sources of variance in the measurement can also be present in the scoring and interpretation of the instruments. For example, when the scoring process requires a lot of participation from the scorers (which is what happens with open ended items), consistency scores among judges are generally calculated, which is another way of analyzing reliability (Joint Committee on Standards for Educational and Psychological Testing, 1999; Cohen & Swerdlik, 2009).

C Standardization and its importance for validity and reliability

One must keep in mind that how a test is administered and analyzed can also affect its validity and reliability (Joint Committee on Standards for Educational and Psychological Testing, 1999; Cohen & Swerdlik, 2009). For a test to be valid, it is not enough for the technical features of the instrument to be valid. All the instruments must have been administered under the same standardized conditions of application. This means that the instructions, the context of application and scoring procedures have been exactly the same for all examinees. That ensures that data can be properly interpreted, compared, and used according to the principle that each user was treated fairly. Therefore, any alteration in the standardization of the application affects the comparability and the validity of the test (McCallin, 2006).

This brings us to the concept of standardization. What makes a test standardized is not the use of standardized scores, or that it is a multiple response test, but rather that the conditions have been standardized, i.e., instructions, administration conditions, and scoring are clearly defined and are the same for all examinees (New Zealand Ministry of Education, 2014; Cohen & Wollak, 2006). Standardization is important for all kinds of instruments, regardless of whether they are norm-referenced or criterion-referenced, the format they have and whether or not they have different forms. Standardized instructions ensure that all examinees have the same understanding of what is expected of them (Cohen & Wollak, 2006). Examples of alteration of the administration situation are: test administrators giving more or less time to answer the instruments; test administrators not reading the instructions (leading to confusion about how to respond or the scoring conditions, for example, if the incorrect answers are discounted from the correct answers); any failure to follow the application protocol, for example a test administrator who reads the reading comprehension questions aloud; altering the instructions (times, instructions, response format, cheating); variability in the application centers (posters with information relating to test content on the wall; interruptions during sessions; differences in the time of application, which leads to children with different level of tiredness and hunger); different conditions in temperature and lighting; and technical issues relevant to the administration (Cohen & Wollak, 2006; McCallin, 2006). It is also considered a lack of standardization if students are pressured to perform well, if they have prior exposure to items, or if the test administrator gives them some of the answers (Cohen & Wollak, 2006; McCallin, 2006).

The ethics section describes the ways in which ethical breaches on the part of the examinees or the test administrators can affect the validity of an instrument, addressing the issues of intention of deception, fraud, and cheating. It also goes deeper into the discussion on the relationship between the fairness of an instrument and validity.

III. Practical Considerations

In addition to the technical considerations already described, there are a number of practical considerations that are also important to take into account when deciding on an instrument. This section describes elements such as cost, times of application, or training required for test administrators, which may make it unviable to use a certain instrument in certain context, even if its content is very appropriate and its technical quality is excellent. Sometimes, these elements may not be a limitation, but they are factors that should be taken into account in the administration planning and budget development. For example, the time and resources required for the recruitment, selection and training of test administrators, scoring of open ended questions, or the time and number of people required for the data entry or scanning of data are often underestimated.

Costs

A central element in this analysis is the issue of cost, which can be a decisive factor when choosing an instrument. For each potential instrument one must know how many resources are needed to implement the assessment in its entirety (Center for Assessment and Research, James Madison University, 2014; New Zealand Ministry of Education, 2014; Timmons et al, 2005). There are several types of costs associated with instruments: buying the rights of the tests themselves, the answer sheets, and the processing, scoring and data analysis, by the owner of the test or independent provider. In addition, there are costs associated with the payment of salaries of those doing the administration and scoring, legal or licensing fees for these hirings, and renting a place for the assessment, storage of test material or scoring of open ended questions, if needed. (Cohen & Swerdlik, 2009).

With regard to the instrument rights, some are available free of charge, but others must be purchased from their authors or publishers (Center for Assessment and Research, James Madison University, 2014). While there is a wide range of prices for paid instruments, it is important to be aware that the best instrument is not necessarily the most expensive. The majority of the publishers of pencil and paper tests charge for manuals and other administration materials, in addition to each individual test, answer sheets, and scoring services. If a cost-effectiveness analysis is performed, there are a number of factors that must be considered. For example, some cheap tests can be useful, or very expensive instruments may have a very limited utility to a given population. It is also important to consider how many times the instrument will be used and if it is possible to partner with another institution to share costs (Timmons et al, 2005).

Application times

The instruments manuals always specify time (or time ranges) for their application. This can determine the suitability of the instrument for a certain use (New Zealand Ministry of Education, 2014). For example, the program may not have sufficient time and money for a long administration. In very young children, long tests can make them anxious and make their scores less valid than the scores on shorter tests. In other cases, fatigue can be a factor that influences results. It is important to choose instruments whose validity is not affected by this type of variable (Timmons et al, 2005).

Another practical consideration is that sometimes the assessments must conform to the duration of the school teaching schedule. For example, if there are periods of 90 minutes between recesses, perhaps it is logistically complicated to administer an instrument that lasts 120 minutes (Bart, 2009).

Therefore, it is necessary to assess the application times based on the time available and the type of student to whom the test will be administered, to assess whether or not it is appropriate for the particular context.

Required training for test administrators

The administrators of the instruments play an essential role. The tests vary in the level of expertise and training required by administrators or test scorers (New Zealand Ministry of Education, 2014; Timmons et al, 2005). When specific training or experience is required, this is specified in the manuals or websites of the instruments. Sometimes it is even necessary to submit documentation showing the test administrators' training before being able to access the tests. The administration or scoring of tests by personnel without the necessary qualifications is a serious ethics violation and may also affect the validity of the results (Timmons et al, 2005).

Some instruments require specific training for their administration (beyond the administrators' professional background). It is essential to have this information in advance, and if training is required, to know how much it costs and how long it takes (Ontario Ministry of Training, Colleges and Universities, 2011). It is also important to know if one has appropriate trainers to hold the trainings.

Easiness/difficulty in scoring and analysis

Some instruments require time and/or special training to be scored or to analyze the data, which can involve recruiting and training people. This is especially common in tests that assess writing. It is necessary to consider the time required and the costs associated with the selection and training of scorers, which includes designing and implementing a system to ensure reliability

among scorers, in addition to the time that is devoted to scoring (Cohen & Wollak, 2006). It is also necessary to take into account the physical space required for scoring.

How the data or scores will be recorded in a database must be taken into consideration (New Zealand Ministry of Education, 2014; Center for Assessment and Research, James Madison University, 2014). Test scoring is more efficient than before, and often with access to computerized scores or scores by Internet. In some situations scoring can take place immediately (Timmons et al, 2005), but in other cases the information must be uploaded. Most of the tests today are read with an optical reader, which is much faster than data entry. For this, it is necessary to have the appropriate answer sheets and the appropriate optical reader (Cohen & Wollak, 2006).

Group or individual administration

Some instruments require one-on-one administration (one test administrator per examinee) while others are designed for group application. In practical terms, the easiest option is to use group application instruments, i.e., having only one or two test administrators per room.

However, many instruments that require observation on the part of the test administrator or that evaluate very young children are individual administration instruments, i.e. one to one between the test administrator and the examinee. This may be impractical from the practical point of view due to cost and time constraints. On other occasions, an individual administration instrument may be the only way to assess what one wishes to measure. In this case, it is necessary to have sufficient resources. Most instruments that have been designed for individual administration cannot be used easily with a group (New Zealand Ministry of Education, 2014).

Easy/difficult to use

Tests should be as easy to use as possible, since students' performances may be affected if they do not understand the instructions. Young children, above all, can spend valuable time trying to understand the process and not the content. For example, complex answer sheets can confuse the user and a student might realize in the middle of the test that he or she has been answering in the wrong section (Timmons et al, 2005). For first and second graders experience recommends not using answer sheets, because the instructions may confuse them. It is better for them to respond directly on the test.

IV. Types of Instruments

There are various ways of classifying learning assessment instruments. For example, Cueto (2007) proposes that one way to classify assessment systems is according to the implications of their results. There are low implication systems, which generate information for training purposes, without direct consequences for those involved, and high implication systems, which use test results for various purposes that do have consequences for those involved, such as those that define teacher incentives, define educational advancement of students, or inform the public about student performance in schools. These often use census administration and are linked to accountability in education.[13]

Another way of classifying instruments is according to their technical characteristics, which is the approach that will be used in this section. One way to classify using this approach is according to whether they are direct or indirect measures of learning. Direct measures are those in which a product of the student's work, such as research projects or tests of knowledge, is observed. Indirect measures, on the other hand, are those that are not based directly on the student's work but on the perceptions of students, teachers, and other agents. This document deals mainly with learning tests, which are a direct measure of learning.

Below are three widely used classifications: parametric and non-parametric instruments, norm-referenced and criterion-referenced instruments, and open-ended and closed-ended response instruments. Other types of instruments, specifically adaptive tests and value added tests, are also briefly discussed. It is important to know all of these categories, which regularly appear in the specialized literature, and to be familiar with the characteristics of each type of instrument in order to understand their differences and be clear about what kind of instrument would be the most appropriate for the assessment purpose in mind.

A Parametric and non-Parametric Instruments

Parametric tests are designed to represent the general population, for example, of a given country or of a certain age group within a country. Normally parametric tests are commercially available, have been piloted and standardized in the entire population and offer many details about their sampling, validity and reliability. Non-parametric tests, on the other hand, have been designed for a specific group and do not consider the general population[14] (Cohen, Manion & Morrison 2000).

---

13 Cueto also argues that there are major criticisms of high implication system due to issues with the validity of the inferences that can be made. The objections revolve around issues such as up to what point are the comparisons between educational centers that cater to different populations, in different contexts, with different human resources and materials, fair.

14 Parametric tests make two big assumptions about the general population and its characteristics: i) that there is a normal curve of distribution in the population (which is noted for example, on standardized IQ scores or achievements in reading of the total population), and ii) that there is a real zero and continuous and equivalent

The advantages of the parametric tests are that they enable the comparison of sub-groups with the general population (for example, the results of a specific school against the national average score), sophisticated data analysis, and inferences. Non-parametric instruments, on the other hand, are particularly useful for small and specific samples, for example a certain class in a certain school, and can be designed specifically for what is being measured in the situation. In addition, the analyses that can be done are necessarily less sophisticated, but they are easier for any user to perform (Cohen, Manion & Morrison 2000).

An example of a parametric test would be a national census test, in which the results obtained by each school can be compared with the average results obtained by the country's schools. A non-parametric example would be a test designed by a school to diagnose their students' maths knowledge at the start of the school year.

B Norm-referenced vs. criterion-referenced instruments

Norm-referenced assessment

Norm-referenced assessments are those in which the score obtained by an individual is compared with the score obtained by a group, allowing the user to see how the performance of a student or group of students compares to others (of the same age, level of studies or another feature in common) (Cueto, 2007; Cohen & Swerdlik, 2009). The score is thus understood in a relative manner, in comparison to the scores obtained by others.[15] An important objective of this type of evaluation is to do rankings (Cohen & Swerdlik, 2009). A classic example is a college admissions test, like the Scholastic Assessment Test (SAT) test.

In this context, "norm" is understood as the usual, customary, expected, or typical behavior. In the context of psychometrics, norms are the data on the performance of a specific group that are used as a reference to evaluate or interpret individual scores. The members of this sampling with which the score is calculated are considered typical of the group, and the distribution of scores in this group is used as the test norms against which the individual scores are compared (they can be gross[16] or standardized scores). The specific group (the "others," with which the individual

---

intervals between scores (unlike typical non ordinal scores on questionnaires where one cannot assume equal distances between intervals). The non-parametric tests, on the other hand, makes few or no assumptions on the distribution of the population or their characteristics, and do not assume a normal distribution, and therefore cannot be compared with the general population. Therefore, in this case non-parametric statistics should be used (Cohen, Manion & Morrison 2000).

15 This model comes from psychology, where the variables of interest, such as intelligence, are often measured in levels of intensity and not by an absolute lack or absolute mastery. Applied to education, this model led to the development of tests that sought to achieve a normal distribution in performance, and presented data by analyzing the relative position of a group of students against another. Thus, results were presented indicating in which percentile the performance of a student or group of students fell, which enabled the user to know what percentage was above and below the described group (Cueto, 2007).

16 The raw score is the first quantitative result obtained by correcting a test. It usually corresponds to the number of correct answers. Raw scores, in themselves, lack significance and must be transformed into derived scores to

scores will be compared) can be as broad as the population of a country, or as specific as the female patients of a certain hospital (Cohen & Swerdlik, 2009).

There are many types of norms: by age, by grade, national, local, subgroup, by a reference group, and percentile (raw data from standardized sample converted to percentiles (Cohen & Swerdlik, 2009).

Standardizing a test at the national level is very expensive, so some tests do not use a representative sample at the national level to determine their norms, but rather only use the statistical descriptive of a group of individuals in a specific context. These are called user or program norms (Cohen & Swerdlik, 2009).

Criterion-referenced assessment

As has just been described in the norm-referenced assessment, a way to give meaning to a score is to compare it to a group score. The other way to give it meaning is to compare the score with respect to a criterion or to a predetermined learning standard. For example, in educational assessment, tests are often used to determine if an individual has achieved a certain level of competence in a field of knowledge or skill set. Criterion-referenced assessment gives meaning to the result of a single score by comparing it to a standard, which can be referenced to knowledge, skills, or abilities. Here the focus is on the performance of the person being assessed: what matters is what he or she can or cannot do, what he or she has learned, and if he or she meets or does not meet specific criteria expected for their group. No matter how he or she has fared in reference to the rest of the group, an individual score does not affect the performance of the others (Cohen & Swerdlik, 2009; Urbina, 2004).

Criterion-referenced assessment is also known as assessment referenced to content, domains, goals or competencies (Urbina, 2004; Virginia Tech, School of Education, 2014; Cohen & Swerdlik, 2009). These different names are due to the fact that this type of assessment can address a set of knowledge in a certain domain, proven on a standardized test, or the demonstration of a level of competence, or compliance with certain learning objectives, and may even refer to the relationship between certain scores on a test and performance levels expected in a certain criterion (Urbina, 2004).

This type of test can provide quantitative data, such as percentages of correct responses. There may also be qualitative categorizations, whether all or nothing with regard to a certain level of achievement (for example, a driving test is either pass or fail), or offering several levels of possible intermediate performance (Urbina, 2004), as in the international PISA test.

---

be interpreted. A raw score can only be interpreted by being contrasted with one or several normed groups, which will enable its transformation to Norms in Standard Scores, Percentile or T-score.

This type of assessment requires a thorough description of what the expected standard is and where to put the cut scores (Cohen & Swerdlik, 2009). It is important to clearly define the content of what is being assessed and to ensure that the test effectively assesses all elements defined as important. To do this, you can make tables of specifications detailing the quantity of items for content or aim of learning (Urbina, 2004). These elements are very important to consider if choosing a criterion-referenced test: the test criteria and cut score must be related to the assessment objectives and content.

Pros and cons of norm-referenced and criterion-referenced assessments

The advantages or disadvantages of a criterion-referenced or norm-referenced assessment have to do with the assessment objective.

In general assessment focused on learning is criterion-referenced assessment, since it is directly linked to learning objectives and a series of performance objectives. In other words, it assesses how much students have achieved the proposed objectives, and it measures competencies with respect to an instructional objective. Additionally, the information obtained can also serve the student him or herself or teachers to improve on those issues identified as weak (Virginia Tech, School of Education, 2014). Norm-referenced assessment, however, is particularly useful if the goal is to order students, schools or countries, either simply to rank or to select the best. However, it is difficult to establish how much students know in relation to a particular standard or minimum acceptable level (Cueto, 2007).

A disadvantage of the criterion-referenced assessment is that it is gives no information on the relative performance of the student compared to the rest, and if the test is not designed to collect this information, information about students who are at the extreme ends, either more advanced or less advanced, can be lost. For example, a test could be focused on measuring basic reading skills, but would not identify students who have very advanced reading skills. Instead, a norm-referenced test could identify a student who stands out among the rest (Virginia Tech, School of Education, 2014). Or conversely, a test designed to describe advanced mathematics learning provides very little information on what students who fall under the most basic achievement level know how to do. According to the assessment objective, this may or may not be a problem.

It is important to emphasize that, although many instruments belong to one category or another, the two approaches are not necessarily mutually exclusive (Cohen & Swerdlik, 2009). An example of this is the PISA test, which enables the user to rank the scores, and also describes the percentage of students located at each performance level.

C Open-ended or closed-ended response instruments

The instruments that are used to measure learning can be separated into those that employ open-ended responses and those with closed-ended responses. The latter involve choosing alternatives

within a series of provided options. The most well-known example is standardized tests with multiple-choice questions. Open-ended response instruments, on the other hand, are assessments in which the response must be created, such as essays, reports, or oral examinations (Center for Assessment and Research, James Madison University, 2014). Open-ended tests also fall in this category.

The advantages of multiple-selection instruments are that they can be administered quickly and easily to large numbers of students and address many topics. However, they offer less richness, less depth and tend to focus on the ability to remember and other low cognitive skills. The advantages of the constructed response are that they provide a deeper information about what students know and can do, but they are expensive because they must be scored by personnel trained especially for this purpose, which takes quite some time and generally requires longer application times (Center for Assessment and Research, James Madison University, 2014). In addition, if the scoring is not done correctly and in the same way by all the test administrators, the score comparison is affected, which does not occur in the case of closed-ended tests.

There are some tests that combine both types of items. For example, a test may be mainly formed by multiple choice items, but then complements the information collected with one or two open-ended questions, requiring the student to write a text and create arguments, thus assessing skills that cannot be collected in the same manner with only multiple-choice items.

D Other types of instruments

Adaptive tests

Normally tests are composed of a set of items that must be answered by all students. However, in some contexts it may be useful to use adaptive tests, which have the feature of being able to choose the degree of difficulty of the items that correspond to each student from the answers that have already been given. This enables much greater precision in the measurement and also solves the problem of regular tests, where if the student knows very little, the majority of the items is very difficult for him or her and will require some guessing, and if he or she knows a lot, the majority of the questions are too easy. Other advantages are that adaptive tests reduce the amount of items that each student responds to, they allow each examinee to progress at their own pace, and students feel that the task is challenging but achievable.

Adaptive tests are necessarily administered on a computer, and not with paper and pencil. Responding in the computer negates the need for answer sheets and therefore data analysis is easier (there is no need for data entry) and the security of the information is greater. They are particularly useful in large-scale assessments in which students are expected to have a wide range of skills.

Some disadvantages are that the skills required to operate the computer may be different from those assessed by the test, computer use can increase or decrease the level of motivation of

examinees, and there are more external elements that can influence the results, such as the slowness of the computer, its brightness, etc. There may also be technical factors in the data analysis, related to the calculation of the standard error.[17] Other technical difficulties are that the test requires a very large pool of items of varying degrees of difficulty, which may affect its quality, and that it is necessary to very thoroughly calculate the rules of assigning items based on rigorous pilot testing that provides information on the difficulty of the items (Cohen, Manion & Morrison, 2000).

Pre-tests, post-tests and added value

An important question in assessment is not only if students can demonstrate learning at the end of a course, program or school year, but how much of that learning was obtained during the course, school year or program in question.

This suggests the need to measure students' skills when they enter and leave the program, something that can be achieved by applying a pre-test and a post-test. Different types of instruments can fulfill this purpose; portfolios are particularly suitable. On some occasions, standardized tests developed for a specific group can also serve this purpose. However, if exactly the same test is applied, the improvement may be due simply to the fact that the students were already acquainted with the test. On the other hand, if different tests are administered, it is difficult to ensure their comparability (Indiana University Southeast, 2006).

Sometimes, only post-tests, i.e. tests that are applied once the process is completed to identify whether the students achieved the proposed learning objectives, are used, without applying a pre-test (Virginia Tech, School of Education, 2014). For example, if the objective is to certify the learning obtained, a post-test is enough, since it is not part of the objective assessment to see how much learning is obtained due to the course. In other cases this question is in fact relevant, and to see how much of the learning was due to the specific program or course, it would also be necessary to apply the post-test to a control group. Otherwise it will not be known whether the results were due to the program or course or to other variables, such as the passage of time. Another alternative is highly specialized tests, called "value added," that some suppliers provide. These are administered at the beginning and end of the school year and enable identification of how much of the learning has been achieved due to the effect of the school or teacher.[18]

---

17 For more information on how to calculate this error, see Thissen, D. (1990) Reliability and measurement precision. In H. Wainer (ed.) *Computer Adaptive Testing: A Primer*. Hillsdale, NJ: Erlbaum, 161–86.

18 To calculate the aggregate value, one must have aggregated data of equivalent students, against which to compare the results, to be able to estimate what the performance would have been if the intervention had not been made.

V. Ethical and Fairness Considerations of the Instruments

This section discusses issues related to ethics and fairness in the selection and administration of learning assessment instruments. These are fundamental issues at the value level, in order to ensure equity in assessments, and are at the same time inexorably linked to technical aspects such as validity. The section first addresses the ethical practices of those who design and administer instruments and of those who are assessed. It then discusses aspects relating to the fairness and bias of the assessments and describes the changes in the assessment situation that are sometimes implemented to accommodate students with special educational needs, with the intention of carrying out fairer assessments.

A. Ethical Considerations

On the one hand, it is the assessment developer's responsibility to design the test with strict ethical standards in all its stages: in the design and selection of the test, the revision of norms and standards to ensure suitability for the group to which it will be administered, ensuring that the modifications for students with special educational needs have been done correctly, etc. Since many of these elements require technical expertise, many ethical breaches can be involuntary and due to a lack of technical knowledge. On the other hand, it is also the responsibility of those who are being evaluated to ethically participate in the assessment, although there is ample evidence of test takers attempting to falsify results (Cohen, Manion & Morrison, 2000). Below the breaches — voluntary and involuntary — that may be committed by the different individuals involved are described, along with some guides for conducting assessments that meet ethical standards.

Ethical practices on the part of those being assessed

The greater the consequences of an assessment, the more unethical conducts appear on the part of those who are being evaluated or those that will be accountable for the results (Cohen, Manion & Morrison, 2000). One can identify three types of fraud: a) the examinee cheating while responding, b) theft of items, and c) a third party changing the scores, either by impersonating another in taking the test, or changing the answers given by the test taker (Impara & Foster, 2006).

a. Cheating can manifest in different ways: interacting with others during the test via cell phones or other means, using summaries or prohibited material, using more time than permitted, using calculators if they are banned, etc. (Impara & Foster, 2006).

b. Item theft consists of reproducing the questions to then lend them or give them away (this may be done by the examinees or anyone else who has had access to the material). In some contexts, there are pirates who take a test only to memorize the items and sell them: if they work in a group, they manage to obtain many items with only one application (Cohen & Wollak, 2006; Impara & Foster, 2006). There are also cases of teachers who train their students to memorize certain items and thus reconstruct a test.

c. With respect to changing answers, there are many cases of teachers and even principals who, due to pressure to have good results, change the answers given by students on

answer sheets (Cohen & Wollak, 2006; Phillips & Camara, 2006). Another fairly common school phenomenon is to ask some underachieving students not to attend the day the test is going to be administered, or directly provide them the correct answers during the evaluation (Cohen & Wollak, 2006). And although it is less common, there are also cases in which students from more advanced courses take a test instead of the course that should actually take it, i.e. group cases of impersonation.

Another important source of data distortion related to the teacher and principal role is the preparation that students have received for the test, also known as teaching to the test (Cohen & Wollak, 2006). This can manifest in many ways, including: ensuring coverage of the program content and objectives that will be included in the assessment; restricting coverage only to those elements that will be assessed; preparing students to take this type of test; practicing with similar tests; telling the students in advance what will be on the test, and practicing with the test itself without teacher support or reviewing the test with the teacher (Cohen, Manion & Morrison, 2000).

There has been much debate about how appropriate teaching to the test is. The barrier between what is legitimate preparation for a test and what is not is diffuse. For example, in the case of tests with consequences for students, they are left at a disadvantage if they are not prepared (Cohen, Manion & Morrison, 2000). In fact, for university admissions tests, part of the rules of the game is that students are prepared. However, preparation is not something desirable in program assessment. A definition of unethical practices (Gipps, 1994, in Cohen, Manion & Morrison, 2000) is when scores increase but the reliable inferences about performance do not, and when different groups of students are prepared differently for the test, giving some unfair advantages over others. Gipps suggests that it is right for teachers to teach content beyond what is included in the test, and not to prepare for the test: ideally, only the best instruction is acceptable. Other authors propose that it is inappropriate if there is a lot of preparation or practice, and that it is appropriate if the contents are taught (Phillips & Camara, 2006).

From a technical point of view, when there is fraud it affects the validity of scores, since the scores do not properly reflect the skills measured by the test. In other words, they are a source of irrelevant variance for the construct being measured. In addition, fraud has problems associated with the assessment's reputation and with fairness for those who did not cheat. In the event that the test must be replicated, it can be harmful or stressful for the examinees to be re-assessed (Impara & Foster, 2006).

Ethical practices on the part of those responsible for the evaluation

The elements that have been described in this article as technical issues to take into account, along with other issues related to quality that have not been included here, can also be considered ethical duties (Cronbach 1970; Hanna 1993; Cunningham 1998, all in Cohen, Manion & Morrison, 2000). For example, for an evaluation to be ethical, the instruments must be valid and

reliable; their administration, scoring and analysis must only be done by people who are prepared and do not have vested interests; access to the material must be controlled; and the scoring processes must be transparent.

When choosing instruments, it is the responsibility of the person making the selection to ensure that the contents are aligned with the assessment objectives and that the format is relevant. In the case of norm-referenced tests, the population used for the sample from which the norm was constructed must be representative of the population being generalized. In the case of the assessment of young children, it must be ensured that the test is appropriate for their age (MCME Code, in Phillips & Camara, 2006).

Many times preventing fraud is also in the hands of the person administering the assessment, by setting strict quality controls in the handling of material before, during and after application, defining that those who have access to the material are not affected by the results and are not responsible for the results that are being measured. For this, it is essential to select appropriate test administrators who do not have conflicts of interest. For example, a common mistake is to use the teacher of the class that is being evaluated as a test administrator, because it is a cheaper option. It must also be ensured that test administrators rigorously follow the instructions and have the required capabilities; for example, they must be able to clearly read the instructions aloud (McCallin, 2006). Preventing distribution of test items prior to application and discouraging unethical practices in preparation for the test are the responsibility of the government in cases in which the assessment is administered by the government (Phillips & Camara, 2006).

Another relevant ethical issue is that students being assessed have rights. Some are based on commonly accepted national legislation or school rules, others on common sense, commonly accepted ethical standards, and respect. In some instances, permission given by the parents or guardians is required. Therefore it is important to learn about the legal context of the place where the assessment will be administered as well as the students' cultural context when selecting the assessment method, and to provide the necessary adaptations for students with special educational needs (The Joint Committee on Standards for Educational Evaluation, 2003). It is also important to respect their privacy and dignity, ensuring that the assessment does not harm them, and their informed consent should be requested before their participation (Cohen, Manion & Morrison, 2000).

There are codes of ethics applicable to researchers who do educational assessment, such as those of the American Educational Research Association, which include elements such as giving information to the research participants (Phillips & Camara, 2006). In the United States, there is legislation in this regard that can be used as a guide for applications in other places. For example, the National Research Act of 1974 requires that the study has to be approved by an external organization prior to its implementation, do no harm, include informed consent and parent or guardian permission must be obtained. The Family Educational Rights and Privacy Act mandates privacy in student registration, which can be achieved by assigning them a code not linked to

their real identity. The National Council on Measurement in Education (NCME) has established a code to guide its members involved in educational assessment and to offer a guide for those who are not members (http://ncme.org). The section addressed to those who select instruments suggests to:

- Conduct a thorough review of the types of assessments and instruments that could be used for the proposed use.
- Select tests based on the public technical documentation on its quality.
- Make clear any relationship they may have with the test owners.
- Inform decision makers and potential users about the suitability of the instrument for its intended purpose, the likely consequences of its use, and the rights of examinees, costs, assessment requirements, and known limitations.
- Veto instruments that may produce invalid results for specific groups on grounds of race, gender, socioeconomic status, etc.
- Comply with all safety precautions in the handling of the material.
- Report any attempts to pressure or manipulate the selection of a certain instrument.
- Avoid using test preparation material that may affect the results obtained by the students.

The American Educational Research Association along with the American Psychological Association and the National Council on Measurement in Education have created a set of standards for educational and psychological assessment, whose 1999 revised version is widely recognized, and addresses issues related to the construction of instruments, assessment fairness, and administration in different contexts (Joint Committee on Standards for Educational and Psychological Testing, 1999).

Did you know...

Since the first formal exams in China 2000 years ago, there have been cases of cheating on exams, using the same techniques that are seen today, such as cheat sheets in the folds of clothing. The punishment could even be beheading.

B. Fairness

Another issue closely related to ethics in assessment is fairness. How does one ensure that a test is fair? How does fairness relate to bias and validity? How can assessments be made fair for students with special educational needs?

Fairness, bias and validity

Since the 1960s fairness, understood as items devoid of bias, has been a major topic in psychometrics (Zieky, 2006). Understood statistically, bias is a systematic error between two

measurements when these should be equal (Camilli, 2006). Or, put another way, it is when irrelevant or arbitrary factors systematically affect the inferences and judgments made based on an assessment in a way that differently affects a student or group of students. These factors may be related to culture, language, developmental or physical difficulties, socioeconomic status, gender or race (The Joint Committee on Standards for Educational Evaluation, 2003). So, a fair test seeks to identify differences, if there are any; and if there are differences, these are due to differences in skill, not to differences in the way of measuring. In other words, if both groups do in fact show differences in their skill level, and the test detects it, this is not an unfair test[19] (Camilli, 2006). Bias can be calculated mathematically, using a metric measurement of the performance of different groups. For example the Differential Item Functioning (DIF) can compare performance on an item by two groups who have the same skill level[20] (Camilli, 2006; Zieky, 2006).

Even today, some authors use the term fair as a synonym for unbiased. Understood as such, a test is unfair when two groups of individuals obtain different measurements on an instrument, while having the same ability in the domain being measured (Camilli, 2006).

For other authors, the terms validity, fairness, and bias of tests are different concepts. While psychometrically bias is a property of a test that systematically prevents accurate and impartial assessments, and is calculated mathematically, fairness refers to the extent to which the text is used in an impartial, fair and equitable manner (Cohen & Swerdlik, 2009). For others, "fair" tests are those that are both unbiased and meet recognized ethical and administration standards. For example, standardized tests should be administered exactly according to their application instructions and everyone should receive the same instructions (Zieky, 2006; Timmons et al, 2005). And, since the quality of the scores obtained will depend on the quality of the norms, it is important to learn everything possible about groups with whom or for whom the norms were calculated. Do the norms represent one's student sample? (Bart, 2009).

Zieky (2006) argues that the most useful technical definition is that which links fairness with validity. Any element that is not relevant to the construct is invalid. Hence, fairness requires that features irrelevant to the construct of examinees do not affect the results (this also includes emotional responses). In other words, anything that reduces the validity of an item reduces its fairness. Linking fairness with validity also implies taking into account the purpose of the test and the use that is being given. A fair test can be used unfairly. Under this perspective, the

---

[19] It is popularly understood that a test is a fair one when it yields the same scores for different groups, for example men and women. However, this definition is wrong. As has just been explained, for psychometrists the difference between group scores says nothing about how fair the test is: for example, a group of men can be taller than a group of women, but this does not reflect unfairness in the instrument that was used to measure height, instead it reflects the effective differences between the groups. Judging the fairness of an item to see if it seems to favor some groups also is wrong, since it is a subjective perception and not necessarily correlated with the actual performance of the people (Zieky, 2006).

[20] For this analysis it is essential to have identified two groups that effectively have the same skill level, otherwise an item can display DIF because the match of skill between the two groups was not done well, and not because there really is bias.

fairness reviews must focus on the validity of the items, not on whether they are politically correct. To be able to determine what is irrelevant to the construct, it is necessary to have a very good definition of the construct to be measured (Zieky, 2006).

Fairness can also be affected if the user has little familiarity with the assessment process, for example, if a person has never used an answer sheet and does not understand how it works, his or her score may be affected. Other factors could be familiarity with the test language, its administration format (pencil and paper or computer), and previous experiences or family experience with the assessment. Practice sessions can be useful in these cases. Test administrators should also be aware if any child, especially the youngest, is having difficulties with answer sheets (Timmons et al, 2005).

For a test to be used in a fair manner, considering the cultural and linguistic differences of the test takers is an essential element. In the United States, there has been an emphasis on ensuring justice for different groups: races, minorities, gender, disability, religion, socioeconomic status and age. It cannot be taken as a given that members of different communities will find certain elements or items appropriate, nor can it be assumed that the same test can be used for everyone when assessing individuals of different cultural and linguistic backgrounds. It can also not be assumed that because a test has been translated into another language, it is exactly the same as the original in all other respects, or that the assumptions in the foundation of the test will affect groups from different cultures in the same way (Cohen & Swerdlik, 2009). The Fairness Review Guidelines (www.ets.org) propose, as guidelines, treating people with respect, minimizing the effect of irrelevant skills to the construct, avoiding offensive or unnecessarily controversial material, using appropriate terminology for people, avoiding stereotypes, and representing a diversity of people in the examples. These guidelines serve for any culture, but must be interpreted according to each culture. For example, something may be inappropriate in Saudi Arabia and acceptable in Sweden (Zieky, 2006). It is always a good idea to analyze the items available to see they could contain material offensive to specific students or words that may have more than one meaning (Bart, 2009).

Did you know that…

Latin American and Caribbean countries use different food measurements, for example, buying apples per kilo or by units, and rice by box or per kilo, which can affect the understanding of maths tests based on daily life examples.

Accommodations

An element closely related to assessment fairness is the changes to the application conditions to allow special educational needs students (SEN) to participate in the same assessments as their peers and be assessed in a fair manner.

Whether these changes are accommodations or modifications, they have been controversial because it is difficult to do this without interfering with the construct being assessed (Zieky, 2006; Thurlow, Thompson, & Lazarius, 2006; Phillips & Camara, 2006).

Both accommodations and modifications involve changes to the standardized administration situation. Accommodations do not interfere with the construct being assessed, and may be associated with: presentation (for example, Braille instead of written words); response times for students with difficulties with language; help marking the answer sheet in the case of children who are unable to mark them by themselves; or changes to the location (holding the test on the first floor so that a student with reduced mobility can have access). In order for these accommodations to be applied correctly, they must necessary be for a SEN student to take a test, without affecting the validity and comparability of their scores. These accommodations should not be associated with the skills related to the construct being measured. For example, helping a paraplegic mark the answers on a maths test is only an accommodation (Phillips & Camara, 2006). However, there is little research evidence on the effects of accommodations. It is difficult to know whether what is being eliminated is irrelevant to the construct, because it is not known if it affects the construct or not. For example, all tests have a reading comprehension component (Thurlow, Thompson, & Lazarius, 2006). Modifications, on the other hand, do change the construct being assessed and therefore standardization and comparability are lost.

This is an issue that has political and legal aspects in addition to the technical aspects. Often there is pressure by interest groups to ensure assessments come with appropriate accommodations so as not to exclude or be detrimental to SEN students.

Moreover, in the United States there is much legislation on this subject, which is necessary to take into account if assessments are being given in this country. If the assessment is given in another country, it is recommendable to check if there is national legislation in this regard (Thurlow, Thompson & Lazarius, 2006).

VI Creating an instrument or using an existing one

There are many instruments available for commercial use that can be used for evaluative purposes. From the practical point of view, using an existing test saves a lot of time and resources (Center for Assessment and Research, James Madison University, 2014; Cohen, Manion & Morrison, 2000). Other advantages are that they are generally technically solid instruments, i.e. have been piloted and standardized in a population described in detail, they state their validity and reliability, they cover a wide range of content, they tend to be parametric tests and therefore sophisticated analyses can be done, they include detailed instructions for their administration, they are generally easy to administer and score, and in general they include guidelines for the interpretation of the results (Cohen, Manion & Morrison, 2000).

The possible disadvantages include: they are expensive; they are often are aimed at a very specific population and may not conform to the required assessment purpose; some have restricted availability so it may be necessary to become a member of a certain institution to use them, which may require compliance with certain requirements; and by definition the available tests are intended for a general population and are not tailored to local needs.

To use one of these tests, it is important to be sure that the goals, purposes and content of the test are aligned with the assessment objectives. The Standards for Educational and Psychological Testing state that for a researcher to decide to use an existing instrument, the golden rule is that he or she must be able to demonstrate its suitability for the purpose (Cohen, Manion & Morrison, 2000). However, it is difficult to find an instrument that conforms exactly to the specific objectives of a program (Center for Assessment and Research, James Madison University, 2014).

Some common errors that are committed when selecting instruments that affect their validity and therefore should be avoided are: using a certain instrument because it has a good reputation or has been used before; using information because it is available; using unfamiliar methods without proper training; not providing accommodations for SEN students or students who do not speak the language (Joint Committee on Standards for Educational and Psychological Testing, 1999).

If the decision is made to design an instrument, the main advantage is that it is created to be perfectly aligned with the program objectives (Center for Assessment and Research, James Madison University, 2014) and it will be adapted with precision to the local and institutional context (Cohen, Manion & Morrison, 2000). Another factor to consider is that while designing an instrument involves huge investment of resources, the instrument belongs to the one who constructs it, so it can be a good investment if it will be applied to many students or for a long period of time, whilst if an existing instrument is used, many times it is necessary to pay for each administration (Center for Assessment and Research, James Madison University, 2014).

However, creating instruments is expensive, slow, and because it will probably be non-parametric, the range of possible analyses will be more limited than in the case of a parametric test (Cohen, Manion & Morrison, 2000). In addition, often one does not have access to

specialized staff that can design a technically rigorous instrument (Center for Assessment and Research, James Madison University, 2014).

In the United States there are a number of organizations that have developed standards or recommendations with practices for test design, interpretation and use. Perhaps the most famous are the standards for educational and psychological tests, a joint effort of the American Psychological Association and the National Council on Measurement in Education, the last version being the one from 1999 (Buckendahl & Plake, 2006). A chapter of the Standards for Educational and Psychological Testing (Joint Committee on Standards for Educational and Psychological Testing, 1999) provides the information that test designers should give, which in general is taken as a guide for commercially available tests. The purpose of these recommendations is to provide users with relevant information (Buckendal & Plake, 2006). It highlights the need to state:

a. The purpose of the test, the suggested uses of its results, including the age group and the qualifications of those interpreting the data. This purpose should be compared with the assessment purpose (Buckendal & Plake, 2006).
b. Information about how was the test constructed.
c. Technical information on standards, scaling, and detailed information about the sample that was used to construct the norm (it must be compared with the group). Evidence about the generalizability of the scores and validity (Buckendal & Plake, 2006).

VII Conclusions

As foreseen in the introduction, choosing the assessment instrument most suitable for the evaluative purpose in mind is not an easy task. A number of technical, ethical, practical, and sometimes political considerations come together, and these may also point in different directions. Compromises will most likely have to be made, and there will not be an ideal instrument, but it is essential to make an informed decision and clearly understand the risks involved.

The main consideration to take into account is that the instrument chosen must be consistent with the assessment objective. This may seem obvious, but in many cases people are tempted to choose an instrument for other reasons: because it is close at hand, they have used it before, they know how to use it, or a major reference group uses it. There may alos be political pressure to use certain instruments or types of instruments.

A major difficulty is that there is no single solution applicable to all situations, nor a checklist of items that can be quickly reviewed each time an instrument is reviewed. Each evaluative objective requires a detailed analysis to see if the instrument is suitable for the particular context, considering whether the instrument effectively assesses what is sought by the evaluative objective and if the population for which the instrument was designed is equivalent to the population being assessed, and taking into account the available resources. Even the quality of the instruments is context-dependent, since the latest definitions of validity do not understand it as the instrument validity of the score validity, but instead as the interpretation of the results of the test for certain uses in certain contexts, and alos because validity and reliability are at stake in the implementation of the instruments as well as in their design.

It is especially important to be able to apply ethical criteria that make the chosen test as fair as possible. It is very difficult to separate these elements from the technical elements, since many times what is correct from the point of view of validity and psychometrics coincides with what is ethical and fair. It is important to have the technical expertise to be able to make informed decisions: otherwise, one risks not only of making technically incorrect decisions, but also ethically questionable decisions, where all groups of students are not being given the same opportunities to demonstrate their knowledge, or some groups are allowed to obtain inflated scores as a result of cheating. The fairest option also depends on each particular situation.

If it is decided that the available instruments do not comply with the requirements that justify their use, it will be necessary to specially design an instrument. This can be a great opportunity to create an instrument that is perfectly suited to the needs of the evaluative objective and to apply the different recommendations outlined in this document. On the other hand, it requires significant investment and specialists who can create such a test.

Bibliography

Banco Interamericano de Desarrollo, División de Educación, 2013. *Documento de Marco Sectorial de Educación y Desarrollo Infantil Temprano.*

Bart, M. 2009. What You Need to Know When Evaluating Assessment Instruments. Available at http://www.facultyfocus.com/articles/educational-assessment/what-you-need-to-know-when-evaluating-assessment-instruments/

Brennan, R. 2006. Perspectives on the Evolution and Future of Educational Measurement. In *Educational Measurement* (4th edition). 2006. Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on education. Praeger publishers, Westport.

Buckendahl C., & Plake, B., 2006. Evaluating tests. In *Handbook of Test Development*. Downing, S., & Haladyna, T., Ed. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc., Publishers.

Camilli, G., 2006. Test Fairness. En *Educational Measurement,* (4th edition). 2006. Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on Education. Praeger Publishers, Westport.

Center for Assessment and Research, James Madison University. 2014. The Programme Assessment Support Services. Downloaded September 10th from http://www.jmu.edu/assessment/pass/assmntresources/instruments.htm#ExistingInstruments

Cueto, S. 2007. Las evaluaciones nacionales e internacionales de rendimiento escolar en el Perú: balance y perspectivas. En *Investigación, políticas y desarrollo en el Perú. Lima: GRADE*. p. 405-455. Available at http://www.grade.org.pe/download/pubs/InvPolitDesarr-10.pdf

Cohen, A. & Wollak, J. 2006. Test Administration, Security, Scoring, and Reporting. In *Test Administration, Scoring and Reporting.* Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on Education. Praeger publishers, Westport.

Cohen, L., Manion, L., & Morrison, K. 2000. *Research Methods in Education* (6th edition). London, RoutledgeFalmer.

Cohen, R. & Swerdlik, M. 2009. *Psychological Testing and Assessment: An Introduction to Tests and Measurement* (7th Edition). Boston: McGraw-Hill Higher Education

Darr, C., 2005. A Hitchhiker's Guide to Validity. Available at: http://toolselector.tki.org.nz/Assessment-fundamentals/Criteria-for-choosing-an-assessment-tool

Haertel, E. 2006. Reliability. In *Educational Measurement,* (4th edition). 2006. Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on education. Praeger publishers, Westport.

Joint Committee on Standards for Educational and Psychological Testing, 1999. *Standards for Educational and Psychological Testing.* American Educational Research Association, Washington DC.

Indiana University Southeast. 2006. The Indiana University Southeast Faculty Assessment Handbook. Available at: http://www.ius.edu/oie/program-assessment/assessment-handbook.html

Institute for Digital Research and Education, UCLA (n.d.). SPSS FAQ. What does Cronbach's alpha mean? Available at: http://www.ats.ucla.edu/stat/spss/faq/alpha.html

Impara, J. & Foster, D., 2006. Item and Test Development Strategies to Minimize Fraud. In *Handbook of Test Development*. Downing, S., & Haladyna, T., Ed. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc., Publishers.

McCallin, R., (2006). Test Administration. In *Handbook of Test Development*. Downing, S., & Haladyna, T., Ed. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc., Publishers.

Ministry of Education of New Zealand, 2014. *Criteria for choosing an assessment tool.* Downloaded July 20th, 2014 from http://toolselector.tki.org.nz/Assessment-fundamentals/Criteria-for-choosing-an-assessment-tool

National Council on Measurement in Education (NCME), 2104. Code of professional responsabilties in educational measurement. Available at: http://ncme.org/resource-center/code-of-professional-responsibilities-in-educational-measurement/

Ontario Ministry of Training, Colleges and Universities, 2011. Selected assessment Tools. Downloaded July 20th, 2014 from

http://www.tcu.gov.on.ca/eng/eopg/publications/OALCF_Selected_Assessment_Tools_Mar_11.pdf

Phillips, S., & Camara, W., 2006. Legal and Ethical Issues. In *Educational Measurement*, (4th edition). 2006. Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on education. Praeger publishers, Westport.

The Joint Committee on Standards for Educational Evaluation, 2003. *The Student Evaluation Standards.* Corwin Press Inc. Thousand Oaks, California

Timmons, J., Podmostko, M., Bremer, C., Lavin, D., & Wills, J. (2005). *Career planning begins with assessment: A guide for professionals serving youth with educational & career development challenges (Rev. Ed.).* Washington, D.C. Downloaded from http://www.ncwd-youth.info/career-planning-begins-with-assessment

Thurlow, M., Thompson, S., & Lazarius, S., 2006. Considerations for the administration of tests to special needs students: accommodations, modifications, and more. In *Handbook of Test Development*. Downing, S., & Haladyna, T., Ed. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc., Publishers.

Urbina, S. 2004. *Essentials of Psychological Testing*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Virginia Tech, School of Education, 2014. *Introduction to Instructional Design. Lesson 7, Assessment Instruments*. Available at: http://www.itma.vt.edu/modules/spring03/instrdes/lesson7.htm

Webb, N., Shavelson R., & and Haertel, E., 2006. Reliability Coefficients and Generalizability Theory. In *Handbook of Statistics, Volume 26: Psychometrics*. Rao, C & Sinhara, R. eds. Available at: http://web.stanford.edu/dept/SUSE/SEAL/Reports_Papers/methods_papers/G%20Theory%20Hdbk%20of%20Statistics.pdf

Wilson, M. 2005. *Constructing measures, an item response modeling approach*. Lawrence Erlbaum Associates Inc., Publishers. Mahwah, New Jersey.

Yen, W. & Fitzpatrick, 2006. Item Response Theory. In *Educational Measurement*,(4th edition). 2006. Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on Education. Praeger Publishers, Westport.

Zieky, M., 2006. Fairness reviews in assessment. In *Handbook of Test Development*. Downing, S., & Haladyna, T., Ed. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc., Publishers.**Annex**