**Please cite the source as:**

Contreras, L. Á. (2002). Contemporary models, procedures and practices in learning assessment: Interview with Anthony J. Nitko. *Revista Electrónica de Investigación Educativa, 4* (1). Retrieved day month, year from: http://redie.ens.uabc.mx/vol4no1/contents-nino.html

ISSN: 1607-4041

**Revista Electrónica de Investigación Educativa**

Vol. 4, No. 1, 2002

# Modelos, procedimientos y prácticas contemporáneos en la evaluación del aprendizaje: Entrevista con Anthony J. Nitko

# Contemporary Models, Procedures and Practices in Learning Assessment: Interview with Anthony J. Nitko

Luis Ángel Contreras Niño
angel@uabc.mx
Instituto de Investigación y Desarrollo Educativo
Universidad Autónoma de Baja California

A. P. 453
C. P. 22800
Ensenada, Baja California, México

## Resumen

En la entrevista, el Dr. Anthony J. Nitko nos aporta su visión sobre los principales modelos psicométricos que sustentan las prácticas contemporáneas de evaluación del aprendizaje, particularmente el papel que desempeñan, en dicho proceso, la teoría clásica de la medida y la teoría de la respuesta al ítem. Además, comenta la noción de validez para interpretar los puntajes de una prueba, así como la evaluación del aprendizaje referida a

un criterio y la denominada evaluación auténtica, tanto en el nivel de gran escala, como la que se realiza en el salón para integrar la evaluación con la instrucción.

*Palabras clave:* Evaluación del aprendizaje, pruebas referidas a un criterio, integración de la evaluación con la instrucción, pruebas estandarizadas.

## Abstract:

In this interview, Dr. Anthony J. Nitko shares with us his vision of the main psychometric models guiding contemporary practice of learning assessment. He particularly deals with the role played by Classical Measurement Theory and Item Response Theory. He also comments about the evidence supporting valid interpretations of tests scores, as well as about criterion referenced learning assessment and the so-called authentic assessment, at large scale and the classroom level where assessment and instruction are integrated.

*Key words:* Learning assessment, criterion referenced assessment, integration of assessment and instruction, standardized tests.

Dr. Anthony J. Nitko is an adjunct professor, Department of Educational Psychology, University of Arizona, and professor emeritus and former chairman of the Department of Psychology in Education at the University of Pittsburgh. He studied his Ph.D. in Educational Psychology, Measurement, and Statistics in the University of Iowa. His research interests include curriculum-based criterion-referenced testing, integration of testing and instruction, classroom assessment, and the assessment of knowledge and higher-order thinking skills. Some of the journals in which his research has appeared include American Educational Research Journal, Applied Measurement in Education, Educational Evaluation and Policy Analysis, Educational Measurement, and Research in Development Disabilities. He has published more than six books; and has recently published the third edition of *Educational Assessment of Students*.

Dr. Nitko has been the editor of the journal Educational Measurement: Issues and Practice. He has been a member of several committees of the American Educational Research Association, was elected secretary of AERA Division D, served on committees of the National Council on Measurement in Education, and was elected to the board of directors and as president of the latter. He has served as a consultant to various government and private agencies in the United States, Bangladesh, Barbados, Botswana, Indonesia, Jamaica, Malawi, Namibia, Oman, and Singapore.

Dr. Nitko has been married to Veronica Vail for 38 years. They have four children and seven grandchildren. All 16 family members live within one kilometre of each other in Tucson, Arizona.

**Luis Ángel Contreras:** In recent years we have witnessed the emergence of a wide spectrum of concepts, models, mathematical procedures and other issues in relation with psychometrics, particularly an impressive increase in the number of specialised references about Item Response Theory and its applications.

What is your perspective about educational and psychological measurement? Will it develop toward the Item Response Theory or will it continue on the line of Classical Measurement Theory?

**Anthony J. Nitko:** Item Response Theory is an outgrowth of Classical Measurement Theory —it is a strong true score theory, meaning it uses strong mathematical assumptions to derive its results. Classical Measurement Theory can solve a great many assessment problems and can explain many measurement phenomena very well. If the strong assumptions of an Item Response Theory model can be met in practice, then Item Response Theory can be used to solve many measurement problems Classical Test Theory cannot.

Although many responsible large-scale achievement assessment developers are using Item Response Theory, they do so with appropriate caution and with research programs that check on whether their assessment results meet the requisite assumptions, and if they do not, what the likely consequences are. Many developers use a combination of Classical and Item Response Theories to craft tests.

Smaller-scale achievement assessment programs will likely continue to rely mainly on classical measurement theory for a good while longer. Those programs have fewer financial resources. They will not be in a position to provide scientifically sound research programs that provide convincing support of the validity of using Item Response Theory.

**L.A.C.:** In your opinion, what is the future of the Classical Test Theory? How do you foresee the development of Criterion-Referenced Assessment? Will it develop along with Item Response Theory?

**A.J.N.:** An achievement test that seeks to optimize criterion-referenced information will need to follow procedures that preserve the ability of the scores to be validly interpreted as descriptors of the degree to which students have achieved the learning targets of the curriculum. In some cases, Item Response Theory may be able to accomplish this validity goal. In other cases, it may not. The debate should focus on valid score interpretations, rather than on what psychometric model is uniformly best.

There is a danger that uncritically adopting an Item Response Theory model will result in eliminating some important types of items or some important types of learning targets from possible inclusion in the test. This elimination distorts the ability of the test scores to be validly referenced to the appropriate achievement domain. Other distortions might occur. For example, in some Item Response Theory models, students answering the same number of items correctly will receive quite different Item Response Theory “ability scores”. While this is legitimate under the model, it makes it difficult to link the ability scores back to a well-defined domain of learning targets.

There are comparable problems when the Classical Test Theory model is used to develop an achievement test from which criterion-referenced information is needed. For example, two students may have the same number-right score, yet each may have answered quite different items —items that represent quite different learning targets.

These problems highlight the fact that no achievement test development procedure is perfectly valid.

**L.A.C.:** What is your point of view about the notion of Construct Validity? Is it essential in Criterion-Referenced Assessment?

**A.J.N.:** From my perspective, there is only *validity*, not different types of validity. Both the test user and the test developer must provide sound coherent arguments that the test scores can be validly interpreted and used in the way that the developer and user want to use them.

The validity argument must be based on evidence from different sources (see below) and must combine evidence from these sources into a coherent and convincing argument that supports a particular intended interpretation or a particular intended use of the test scores. In this sense, a validity argument is a local one, meaning that an achievement test cannot be uniformly valid for all persons to interpret and use in any way they want to do so. Each interpretation and use must be supported through a distinct validity argument.

There are basically eight sources of evidence: (1) content representativeness and relevance; (2) thinking skills and mental processes a student is required to use to respond correctly to each item; (3) relationships among the scores on the items or parts of the test; (4) relationships among the test scores and the scores on test assessing similar and different abilities; (5) reliability of test scores over different content samples from the same test specifications, over different occasions of testing the same students, and over different assessors marking the students' responses; (6) generalisation of the test interpretations and uses when the test is used with different genders and categories of people, different age levels, etc.; (7) the value of using the test in relation to the intended and unintended consequences of the test results; and (8) the cost, practicality, and instructional features of using the test. These sources of evidence are explained in detail in the recent edition of my textbook.

Although evidence from all eight sources is needed for every argument, evidence from some sources will be more important for arguing for the validity of a particular test interpretation or a particular test use. For example, if an achievement test is to be used to provide criterion-referenced information about a student's reading comprehension, then evidence from sources (1), (2), (6), and (7) should strongly support interpreting the test as an assessment of reading comprehension. Sources (3), (4), (5), and (8) will also be needed to support the test interpretation

and use, but if evidence from sources (1), (2), (6), and (7) is not strong, it will be difficult to claim the test can be interpreted and used for obtaining criterion-referenced information about reading comprehension.

**L.A.C.:** What is your opinion about the perception that we have a psychometrics of the XXI century and a Psychology of XIX century?

**A.J.N.:** This is an interesting question, but I doubt that the perception is correct, especially in the field of cognitive achievement testing. Cognitive achievement is about learning to think and to apply one's thinking to understand the real world. Our modern psychology is helping us to understand how persons think about different aspects of the real world. As our understanding of how persons think and apply their knowledge to the real world grows, then our assessments and psychometrics will follow.

It may be that the perception described in your question is based on some persons who view the elaborate equations in the Item Response Theory models as 21st century phenomena. The basic equations are only slightly more elaborate than the normal curve model that was invented by Abraham deMoivre in 1733 —of course, modern computers are needed to estimate the parameters of those basic equations. The "ability scales" of the Item Response Theory models are often interpreted substantively as "traits" or "abilities", in ways that are similar to the psychological views of traits and abilities of the 18th (and earlier) centuries. It is difficult to argue, therefore, that these psychometric equations are 21st century phenomena.

**L.A.C.:** During the past decade we have also witnessed an accelerated growth of the authentic assessment movement, both in the classroom and in a large-scale standardized context. In many technical documents you have been identified as a theorist leading this movement.

Is this perception correct?

**A.J.N.:** Throughout my professional career, I have stressed that educational assessments should contain items and tasks that come as close as possible to the ultimate achievement we would like for students when they leave formal schooling and live their lives in the real world. That is, students should read real texts, apply mathematics to real-life situations, write using a process that real writers use, etc. This is really a continuation of a long tradition of the philosophy of educational assessment that was begun by E.F. Lindquist. Lindquist formalized his position in the first edition of Educational Measurement (1951) that he edited. Thus, I have not thought of myself as a theorist leading this movement, but only as a professional attempting to implement this educational assessment philosophy.

Some educators in the 1990s were not well schooled in the traditions of modern achievement assessment. They saw several abuses of educational achievement tests and (rightly) reacted negatively to these abuses. Because they were not

knowledgeable of the long tradition of the educational measurement philosophy developed by Lindquist, his students, and his followers, they had no educational achievement-testing framework that they could use as a platform to express their concerns. As a result, they invented their own terminology and framework, calling it authentic assessment.

**L.A.C.:** In your opinion, what is the actual status of standardized authentic assessment compared to the psychometric development reached by traditional item selected response tests?

**A.J.N.:** If you mean, what procedures are used to develop authentic assessments found in the market place, then I can respond to that. When a commercial publisher prepares an instrument for wide use in the schools, the publisher must yield to the wishes of the school authorities that are buying the instrument. A reputable publisher will also follow professional guidelines for validating the assessments' interpretations and recommended uses. The reputable publisher will also provide information on the reliability of the scores from the assessments. Further, if school authorities also want information about how the achievement of students in their schools compare with the achievement of similar students in the state or nation, reputable publishers will provide that information also. This is standard industry practice.

This means that an authentic assessment from a reputable publisher will have a scientific basis for developing and improving the instrument. By necessity, this means that assessment tasks must be tried out ("trialed") with students, revised if flaws are found, and selected to be the best and most efficient tasks for assessing students. It also means that the publisher will provide scientific evidence that the assessment results can in fact be interpreted and used as "authentic" and that the evaluations of students from these assessments are reliable.

There are many non-reputable publishers who sell assessment tasks that have no scientific basis to support their claims for being valid or reliable. They would like us to believe that their assessments are authentic simply because they say they are. I am always sceptical of such unsupported claims, and so should other persons.

**L.A.C.:** Has standardized authentic assessment met the most fundamental criteria of measurement?

**A.J.N.:** To the extent that (1) a particular assessment instrument is a representative sample of tasks from the relevant domain of authentic tasks, (2) preserves the authenticity throughout the assessment and marking process, and (3) reliably distinguishes the degree to which each person who is assessed has achieve the ability to perform these tasks, then the answer is yes. Not every published assessment instrument does these three things, of course, so each assessment instrument will need to be evaluated individually.

**L.A.C.:** A great deal of your work as a theorist has been oriented to help teachers to develop plans that integrate teaching and assessment, as a means to improve instruction through better assessment of students, to interpret properly local or state-mandated tests, and in other relevant practical issues teachers face.

Is that emphasis primarily related to the three-legged stool: standards, assessments, and consequences of the national reform of education movement in the United States, and with the need to align curriculum, instruction, and assessments with standards?

**A.J.N.:** In order to evaluate how well students have achieved, a teacher needs to be clear about what the students are expected to achieve. Standards, or learning targets as I often call them, are just one of the tools to help teachers understand what their students should achieve. I have seen curriculum guides from many countries and from many schools. A large number of these describe only the content topics that teachers are expected to "cover". Often, they set no standards or learning targets with respect to this content. In several countries, now, there are movements underway to clarify what is expected of students by developing performance standards.

Deciding the specific learning targets teachers expect students to achieve is one important step in the teaching process. Instruction may be thought of as involving three fundamental but interrelated activities:

1. Deciding what the student is expected to learn.
2. Carrying out the actual instruction.
3. Evaluating the learning.

Activity 1 requires teachers to articulate the learning targets in some way, usually by specifying learning objectives or by providing several concrete examples of the tasks students should be able to do to demonstrate that the learning targets have been reached. This may require teachers to translate the given "standards" into specific learning targets. Activity 1 informs teachers and the students about what is expected as a result of teaching and studying. A teacher's understanding of the learning targets guides the teaching plans and provides a criterion for deciding whether students have attained the desired change. The more clearly a teacher specifies the learning targets, the more the teaching efforts and students' learning efforts can be directed. Activity 2 is the heart of the teaching process itself. Here a teacher provides the conditions and activities for students to learn. Activity 3, evaluating whether learning has occurred, is essential for good teaching. Through this activity, a teacher and the students come to know whether they achieved the learning targets.

At the classroom level, if the teacher aligns instruction both with assessment and with standards, then there is an opportunity to give appropriate feedback to students concerning what they need to learn to meet the standards. The three fundamental activities are interactive rather than a straight one-two-three process.

Setting clear learning targets helps a teacher to plan teaching efficiently, conduct instruction effectively, and assess student outcomes validly. Assessing and evaluating students using clearly specified learning targets provide a teacher with information about how effective the instruction has been. This information, in turn, may be used to plan the next instructional activities or to better specify the instructional targets themselves.

**L.A.C.:** With very few exceptions, in Mexico we do not have national or statewide standardised testing. As a matter of fact, our country has been practically apart from the vigorous psychometric development around the world. In my opinion, that is a very strange phenomenon because we have always had a national curriculum in our elementary and secondary education and recently the operation of those curricula has been decentralized to the states. So, based in your experience as a consultant to various government and private agencies in the United States and other countries, and having the educational authorities in mind:

What are the main political and educational contributions of practicing national or state-wide assessments?

**A.J.N.:** Central national or country-wide examinations are set, in my experience, when one or both of two situations exist: (1) education authorities or the public are concerned that achievement standards are not being upheld at the local level or (2) there is a high-stakes decision to be made from the test results (e.g., admission to the next level of education) and there is a concern that local educational authorities (including teachers and principals) cannot be trusted to make this selection objectively or honestly. When the first situation exists, a nationally set test (or at least a standardised test that has national norms) is seen as a way to compare all locally educated students against a common achievement meter-stick. When the second situation occurs, it is felt that the nationally set test provides a fairer and more objective way to identify those students who have best learned the curriculum of syllabus, than the local education authority can provide.

**L.A.C.:** Considering that in Mexico there is a national curriculum in elementary and secondary education, do you think that the model for developing curriculum-driven criterion-referenced and norm-referenced national examinations you proposed in 1994 would be applicable?

**A.J.N.:** The model I proposed is a model of the process that could be used to assure that the examinations or tests are aligned with the official curriculum. It came out of my experience when working with some countries where the examination developers were not working closely with the curriculum developers. Thus, a country might have a new or revised curriculum, but the examination developers were crafting examinations that did not match these new curricular developments. As a result, teachers were ignoring the new curriculum in an effort to prepare their students for the national examination. This hampered curriculum reform.

The model proposes a process whereby teaching the new curriculum will be the same as preparing students for the examination. The alignment of curriculum and assessment would allow the official curriculum to determine what the examination would assess, rather than having the examination define the functional curriculum.

I am not very familiar with the educational situation in Mexico, but if there were a concern about aligning a national or state examination with the official curriculum, then the model would be applicable, I believe. Of course, local or state education authorities could develop local or state examinations assessing the national curriculum by using this model.

**L.A.C.:** When a prominent member of the international psychometric community came to Mexico in 1998, I had the opportunity to ask him his opinion about your proposal to obtain in a single test both criterial and norm referencing, if special procedures are followed to obtain representative samples from the population of students and from the learning targets specified in the curriculum. He made a warm recognition about your person and your work, but disagreed on the point. His position was that both kinds of referencing were incompatible because each of them served to different evaluative purposes. He also commented that it would be as the one who serves two masters.

What do you think about this remark?

**A.J.N.:** The dual criterion-referenced and norm-referenced information capacity of assessment results is one of the most misunderstood aspects of psychometric applications in achievement assessment. The score (or result) from an assessment requires referencing in order to be interpreted (that is, in order to be given meaning). One may reference the score to the domain or population of tasks from which the sample items on the assessment was drawn. For example, if we have a test of solving simultaneous linear equations, then a student's score could be referenced to the domain of all possible simultaneous linear equations —provided that our test was a representative sample from this domain (Mathematics educators could help us define this domain so that we would be sure to include on our test examples of each of the various types of possible problems). The quality of our information from this test about how well a student has mastered simultaneous equations depends on how well we have sampled the defined domain of problems (that is, how representative our test is of the domain) and how many items we have on the test (so our inference to the domain can be reliable). This type of information about the meaning of the assessment scores is called *criterion-referenced information*.

If we gave this *same* test to a representative national sample of high school students, we could compare one student's score from this same test to the scores of other students who took the test. This will tell us how well the student's ability to solve simultaneous linear equations compares to the performance of other high school students in this domain. The quality of information about the comparison of one student to other students in a population of similar students will depend on

how well we have sampled the domain or population of high school students (that is, how representative our sample of students is of the population of students) and how many students were included in the comparison sample (so our inference to this population can be reliable). This type of information about the meaning of the assessment scores is called *norm-referenced information*.

This demonstration I presented shows that both types of information —criterion-referenced and norm-referenced— can be obtained from the same test. Any good achievement test must be representative of the domain of learning it is supposed to assess, otherwise we cannot validly interpret the scores as achievement assessment. Thus, any achievement test that is representative of a learning domain can provide criterion-referenced information. Any achievement test that is administered to a representative national sample of students can provide norm-referenced information.

Now a word of caution. Oftentimes, the quality of both criterion-referenced and norm-referenced information for a particular test is not identical. We may have better quality criterion-referenced or better quality norm-referenced information. Thus, in designing a test, we have to make a conscious decision about which type of information is more valuable to us. After we decide this, we can apply psychometrics to optimise the quality of one or the other type of information. If we decide, that norm-referenced information is more valuable to us, then we would select test items so that each one contributes to ranking or ordering examinees. In this case, the item selection process would eliminate items that nearly everyone can answer correctly or nearly everyone would answer incorrectly, because these items do not discriminate among examinees. Non-discriminating items do not contribute to optimising the norm-referenced information from the test scores. Those items remaining after this elimination will be the items that are better for ranking individuals in the norm group.

If very easy and very difficult items are eliminated from possible selection for the test, then the sample of items that does appear on the test will not be representative of the entire learning domain. As a consequence, we will loose some quality in our criterion-referenced information (similar consequences result when we use other item selection strategies).

I believe that this issue of the quality of information is what some testing specialists have focused on when they call one test a criterion-referenced test and another a norm referenced test. They call a test that optimises criterion-referenced information a criterion-referenced test. They call a test that optimises norm-referenced information a norm-referenced test.

Calling one test a criterion-referenced test and another a norm-referenced test sets up an artificial dichotomy that leads to confusion. The confusion becomes apparent when test users want *both* criterion-referenced and norm-referenced information. Some test specialists tell test users that one test cannot provide both types of information, or that the two types of information are contradictory.

However, my previous example shows that this is not true. The two types of information are different, but *both types are usually needed* before students' scores can be validly interpreted.

There is another problem when testing specialists claim that the two types of information cannot come from the same test. Test users may come to believe they need more than one test. This is likely not to be true. Building two tests inappropriately is likely to waste resources.

Sometimes test users are led to believe, to take another example, that a standardised achievement test cannot tell them anything about a student's mastery of a subject. This is also not true.

I see it, then, as a question of priority: Which is more important to a particular test user, criterion-referenced information or norm-referenced information? Once that is decided, we can develop the test, using sound test development principles, that both optimises the type of information that is a priority and keeps some useable information from the other type.