Standards for objective tests.

(Submitted to AERA Meeting 2007, Accepted for the session: "New Developments in Measurement Thinking", SIG-Rasch Measurement)

Agustín Tristán
Instituto de Evaluación e Ingeniería Avanzada, S.C.
IEIA, Mexico

Rafael Vidal
Centro Nacional de Evaluación para la Educación Superior, A.C.
CENEVAL, Mexico.

A new book of standards for quality of tests has been published in Spanish, filling a gap on this field. The book includes standards, comments, a questionnaire for self-evaluation and a planning schedule; with those tools a non-expert may understand the standards, and easily follow some procedures to design or to improve a test.

Key words: Quality, standards, test.

Test designers need to accomplish some requirements concerning validity, objectivity and reliability for the items and for the test itself; they also have to follow some logistic and safety procedures. It is a common practice that the test's requirements specified by the designer do not follow a prescribed set of quality standards.

Standards for quality of tests are available in English (AERA-APA-NCME), and a Spanish version was published by the Ceneval (Centro Nacional de Evaluación para la Educación Superior, 2000), from Mexico. As it could be seen in several educational environments, the standards are useful recommendations for the specialist, but they are very complicated for other people, such as teachers or institutional practitioners.

Designers need comprehensible standards, with explanations, examples and a guide to follow the specifications' set to design or analyze a test. Our experience was that the published materials did not provide simple procedures and they could not be followed by the test designer, in fact the interpretation of the standards implies a certain amount of background and experience; but even under these circumstances, it produces subjective interpretations.

In a particular experience in 2001-2002, eight tests were evaluated in the Ceneval following the AERA-APA-NCME and the Mexican standards, but due to the subjective approach, many good characteristics of the tests were not shown and some defects of the test remained hidden.

Correspondence may be addressed to: ici_kalt@yahoo.com

According to the previous experience, a new Spanish version of the "Quality Standards for Objective Tests" has been recently published (Tristan & Vidal, 2006). The book is a new version of the former edition of standards published by the Ceneval, but it has a different approach to introduce and organize the standards, it includes several improvements in comparison with the previous version, and also compared to other books of standards from USA and Canada. In particular, the book follows a new concept in its presentation specially addressed to non-evaluators.

The book is based on references from USA, Canada, Mexico and a couple of Latin-American countries, including quality, codes of fair testing, principles of good practices and codes of ethics from several organizations. The standards in the book are not limited to educational or psychological tests, they are also useful for other fields: biostatistics, nursing and medical areas, as well. In addition, the set of standards is not linked to a specific calibration or scoring model, so the responsible of the project may easily begin with classical test theory and evolve to logistic models at its own pace.

The book of standards includes a questionnaire that has been used in Colombia, El Salvador and Mexico for assessment and planning of tests; it was used also to evaluate the evolution of a test and to prescribe the activities to do in the next future.

We have followed the proposed methodology of the book to assess the quality of tests, because it provides an objective way to improve the development of the measurement instruments.

Table of contents of the book

The book proposes a material that can be easily followed by non expert readers, in four parts:

a) The core of standards and comments, with examples and explanations that simplify the ideas contained in the standards
b) A self-evaluation questionnaire, helping the responsible of a project to check what is accomplished or what has to be done.
c) A self-scoring procedure.
d) A planning or schedule to accomplish the set of standards in no more than 18 months.

Organization of the core of the standards

The core of standards includes eleven sections, corresponding to the elements that have to be satisfied by the person in charge of the development of the test:

0. Person in charge or coordinator of the project
1. Organization of the project
    - 1.1 Organizational chart of the institution
    - 1.2 General council or committee
    - 1.3 Specific committees
2. Technical manual and planning of the test
    - 2.1 Technical manual
    - 2.2 Profile of the applicant
    - 2.3 Instruments per profile
    - 2.4 Tables of specifications (test blueprint)
    - 2.5 Planning of the test
    - 2.6 Accommodations
    - 2.7 Avoiding bias
    - 2.8 Verification process
3. Validity associated with the test
    - 3.1 Evidences of validity
    - 3.2 Content validity
    - 3.3 Criterion validity (methodology, specifications)
    - 3.4 Criterion validity (values and evidences)
    - 3.5 Construct validity (methodology, specifications)
    - 3.6 Construct validity (values and evidences)
4. Items and objectivity
    - 4.1 Manual for items design
    - 4.2 Types of items
    - 4.3 Item validation
    - 4.4 Pilot validation of items and tests
    - 4.5 Item calibration
    - 4.6 Description of the item bank
    - 4.7 Considerations of the items' use.
    - 4.8 Diffusion and items open to the public
5. Reliability related to the test
    - 5.1 Reliability and measurement error
    - 5.2 Error vs number of items per variable
    - 5.3 Item verification regarding reliability
    - 5.4 Sampling for reliability analysis
    - 5.5 Reliability of the criterion
    - 5.6 Reliability values
    - 5.7 Bias analysis (methodology, specifications)
    - 5.8 Bias analysis (values and evidences)
6. Test construction
    - 6.1 Test generation
    - 6.2 Tests versions of forms
    - 6.3 Use of versions
    - 6.4 Versions equating
    - 6.5 Design values of the tests
7. Test scoring and interpretation of the results
    - 7.1 Test Scoring
    - 7.2 Scale of the test
    - 7.3 Cut off points
    - 7.4 Scoring for norm-referenced tests
    - 7.5 Scoring for criterion-referenced tests
    - 7.6 Combination of instruments
8. Materials for the test
    - 8.1 Support materials
    - 8.2 Security and confidentiality
9. Application and logistics
    - 9.1 Registration and application
    - 9.2 Facilities
    - 9.3 Protocol for application
    - 9.4 Codification and response reading
    - 9.5 Exceptions and frauds
    - 9.6 Review of the applicant's responses
10. Presentation of the results
    - 10.1 Reports and use
    - 10.2 Analysis of the results
    - 10.3 Delivery time for analysis
    - 10.4 Delivery time for the public
    - 10.5 Users of the test
    - 10.6 Publication of the results
    - 10.7 Use of the results
    - 10.8 Research with the results
    - 10.9 Statistical results
    - 10.10 Training for users
11. Promotion and contracting
    - 11.1 Promotion of the test
    - 11.2 Training on the use of the test
    - 11.3 Contracting

Description of the standards

An example of the standards is the following:

7.4 Scoring for norm-referenced tests	Priority: Medium	Comments
1. The scoring procedure for norm-referenced tests must be described, including the descriptive parameters of the population or sample. 2. Specify the date and time validity of the results, specially when updates have to be developed in prescribed periods or when some changes on the project specifications occur.	a) On norm-referenced tests, associated to the population or group of applicants, the test scoring may be performed following the classical test theory or logistic models. b) This standard allows normalized scores, following a scale (see standard [7.2]) or in raw score. The procedure to calculate the results must be included.	

It can be shown that every standard includes:

a) The standard itself (specifications)
b) Priority (High – to be accomplished in no more than 6 months, Medium – to be fulfilled in no more than 12 months, Low – to be accomplished in no more than 18 months)
c) Comments (explanations, examples, references to other standards)

Following the standards' set, a self-evaluation questionnaire allows the responsible person of the test to identify what has been done, pending activities and show the elements to be included in the technical manual of the test.

The questionnaire has a guide that specifies the standards to be fulfilled by a specific test, according to its purpose: from a teacher's made test up to an evaluation agency.

An example of the questionnaire is shown in Figure 1.

2. Technical manual and planning of the test	
The test must have a Technical Manual, with specifications, design fundamentals and purposes of the test and its planning.	

	Topic	Develop here your answer or indicate the document code in annex	Content of the annex	Priority	Reserved for supervisor
2.1	¿Does the test have a technical manual?		Technical manual	1	
2.2	¿What is the profile or characteristics of the person (student, patient) that will be assessed with the test?		Reference profile of the applicant	1	
2.3	¿What are the sub-tests included in the full assessment set? Describe the structure of each sub-test.		General description of the sub-tests	1	
2.4	Show the tables of specifications (test blueprint)		Test blueprint tables	1	
2.5	¿Does the test have a planning?		Planning schedule	2	
2.6	¿Are there some accommodations to help people with disabilities?		Description	3	
2.7	¿How do you demonstrate the appropriateness of the test regarding gender, ethnicity, socio-economic background, and so forth?		Description	3	
2.8	¿How do you update, modify, improve the test?		Procedure	3	

Signatures of the persons in charge of the test	Signatures of the supervisors

Figure 1

The standards are not linked to a specific psychometric, edumetric or biostatistical model. For instance, it is valid to present one or various of the reliability parameters, according to the experience of the responsible of the test or his own needs, from test-retest, KR20, Cronbach's alpha, Livingston

criterion reliability, generalizability models or Separation following the Rasch model.

Concerning item calibration, the responsible may use classical difficulty-discrimination models or difficulty-fit from the Rasch model.

After the questionnaire has been solved, the responsible of the test may look at the planning including at the end of the book. Once selected the standards to work on, and eliminating the rows corresponding to the standards not needed for the test, the remaining rows show the planning suggested for 6, 12 and 18 months. This suggested planning may be modified by the responsible of the test, but it provides a good foundation to start with the real planning.

A partial sample of the planning of a test is shown in Figure 2.

Name of the test	

			Semestre 1						Semestre 2						Semestre 3					
	Activities	Standard	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	Test organization	1.1																		
2	General Committee	1.2																		
3	Committees by specialty	1.3																		
4	Technical manual	2.1																		
5	Description of the applicant	2.2																		
6	Define sub-tests	2.3																		
7	Test blueprint	2.4																		
8	Test planning	2.5																		
9	Accomodations	2.6																		
10	Avoid bias by gender, etc	2.7																		
11	Process for revision	2.8																		
…	…	…																		

Figure 2

In order to follow this planning, the responsible may develop some activities going from classical test theory to logistic models. Even that the best practice may suggest the use of the Rasch model, other descriptive models from IRT may be employed in specific environments, and the responsible must explain and justify the reason of his choice.

Conclusions

The new book of Standards for objective tests does not provide a set of prescriptions, but also includes some tools for test evaluation, control and planning. The new organization of the standards and the questionnaire have been used in some Latin American countries, with good results, as it improves the understanding of the standards for quality of a test producing a sufficient ground for non expert persons.

References

AERA-APA-NCME (1999) Standards for educational and psychological testing. AERA.

Martínez R.F. & al. (2000) Estándares de calidad para instrumentos de evaluación educativa.Ceneval. México

Tristan L.A. & Vidal U.R. (2006) Estándares de calidad para pruebas objetivas. Editorial Magisterio. Bogotá. 158 pp.

Note: The book includes more than fifty references.