

## Abstract

This poster presents our experience of including systematically the empirical study of face validity in the development of different tests. The procedure followed is exposed as well as the main results obtained and its usefulness.

There are several aspects usually considered in the development of a test. Nevertheless, it is infrequent to take into account the point of view of examiners and test-takers systematically and objectively (their opinions about administration time, difficulty and attractiveness of the stimulus, face validity of the test...). Information that could be drawn from this procedure could be useful to improve test quality (in tryout studies) as well as to establish its properties. Several examples of actual cases are exposed.

## Overview

Several empirical analysis are carried out in a test development process to measure its quality (item analysis, reliability, convergent validity, concurrent validity, factor analysis...). All of them are considered to improve the test and to establish its psychometric properties. However, what it is unusual to consider is examinees' and examiners' opinions about the test. Although test experts agree that examiners' and examinees' feelings and attitudes towards a test should be considered by test constructors and test users, limited work has been done to develop feedback systems designed to study examinees' reactions and perception and credibility attribute to the test (v.g. Nevo & Sfez, 1985; Li, 1995).

The information that could be extracted from these opinions could be very useful to help to improve the test (tryout studies) as well as to establish its properties (Chan, Schmitt, DeShon, Clause & Delbridge, 1997). Moreover, it is important to study the credibility ascribed to the test by the examinee and the examiner because it will affect its use in any assessment process (Baird, 1987; Stricker, Wilder & Bridgeman, 2006). This information is not usually collected in a systematic and objective way rather in a subjective and partial one.

To fill this gap, we have added the study of the examiners' and examinees' opinion in the recent development or adaptation of several test. Immediately after taking a test, examinees are requested to rate their perceived test characteristics. They have to answer a questionnaire that asks them about different aspects of the test (item attractiveness, item difficulty, item understanding, item answer format, understanding of the directions, credibility ascribed, face validity...). Also, examiners have to fill another questionnaire with similar questions about the test although with some question specific to them (v.g., simplicity and clarity in administration/scoring, usefulness perceived, face validity).

Some test development processes, where this procedure have been carried out, are exposed next as well as some examples of its use.

## Method

**Study 1: Tryout of a Vocational Interests Inventory** (219 professional activities; examinee has to express his/her interest for each one in a 4-point Likert-type scale). Sample: 446 participants (64% female; mean age 32.3, SD 11.2, range 15-59 years). After taking the inventory, examinees were asked to fill out a 10-item questionnaire. The questionnaire asked about various aspects of the test (item answer format, redundancy, understanding of the directions...). Examinees were presented with statements about the test and asked to indicate the degree to which they agreed with the statement using a five point scale: 1 = Strongly Agree, 2 = Agree, 3 = neutral, 4 = Disagree, and 5 = Strongly Disagree.

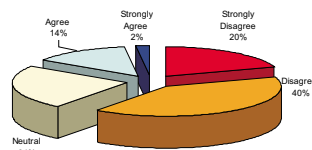
**Study 2: Tryout of the Spanish adaptation of RIAS** (Reynolds Intellectual Assessment Scales, Reynolds & Kamphaus, 2003; Spanish adaptation: Santamaría & Fernández, 2008, individual administration). Examinees' sample: 280 participants (47% female; mean age 22.6, SD 20.7, range 3-90 years). Examiners' sample: 14 examiners (each one administered the Scales to 20 participants). After administering the scales to the participants, the examiners were asked to fill out a 62-item questionnaire. The questionnaire asked about various aspects of each scale and the RIAS as a whole (item familiarity, item attractiveness, face validity, understanding of the directions...). Examinees were presented with statements about each scale and were asked to indicate the degree to which they agreed using a four point scale: 1 = Strongly Agree, 2 = Agree, 3 = Disagree, and 4 = Strongly Disagree.

**Study 3: Standardization of an abilities test: EFAl** (Santamaría, Arribas, Pereña y Seisdedos, 2005; group administration) Examinees' sample: 5865 participants (50% female; mean age 14.4, SD 1.9, range 10-49 years). Examiners' sample: 66 examiners. After taking the test, examinees were asked to fill out a 19-item questionnaire about EFAl. The questionnaire asked about various aspects of the test (item familiarity, item difficulty, item attractiveness, face validity, understanding of the directions...). Examinees were presented with statements about the test and asked to indicate the degree to which they agreed with the statement using a four point scale: 1 = Strongly Agree, 2 = Agree, 3 = Disagree, and 4 = Strongly Disagree. The same procedure was followed to the examiners.

## Results

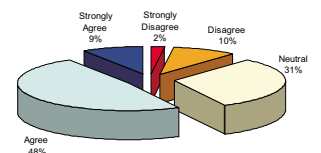
### Study 1. Tryout of a vocational Interest Inventory. Examinee's opinion: Some examples

In general, items (professional activities) are independent enough among themselves, without repeating the content, without being redundant.



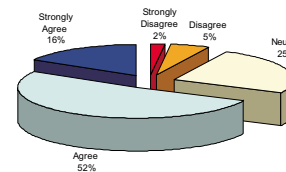
The test activities were perceived as redundant and repetitive. This was taken into account to improve the test.

The item answer format has allowed me to express my interest in each activity in an appropriate way.



The item answer format was considered acceptable, although 12% disagree. It was taken into account and improved.

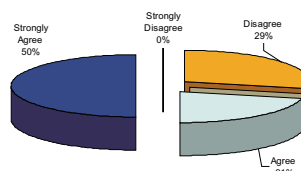
I consider that results of this test can offer interesting information to people that need guidance about what occupation to develop in the future.



The examinees considered the test useful to offer guidance to a constructor: Good face validity.

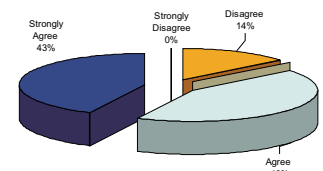
### Study 2. Tryout of the Spanish adaptation of RIAS. Examiner's opinion: Some examples

Verbal reasoning test: The examinee have a clear understanding of the task to be done.



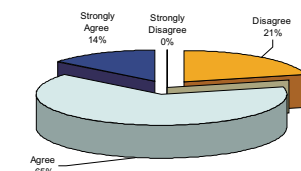
29% examiners considered that there were some problems with the understanding of the task in the verbal reasoning test.

Verbal reasoning test: The test directions are clear for the examinee and for the examiners.



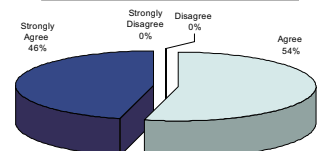
Some examiners (14%) considered the test directions were not clear enough. Together with the previously mentioned (understanding of the task), it was considered the need to improve the directions of the verbal reasoning test. More examples and directions were included and tested in the standardization version of the test.

Verbal reasoning test: The language used is familiar and accessible for the person assessed



The language used in the verbal reasoning test was not considered as familiar and accessible by 21% of the examiners. It was reviewed and improved.

I consider that the results of RIAS, as a whole, will be good indicators of the intelligence and the memory of the person assessed.

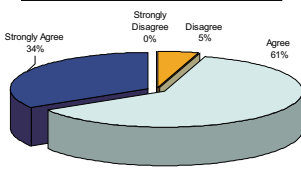


The RIAS, as a whole, showed a good face validity.

### Study 3. EFAl Standardization

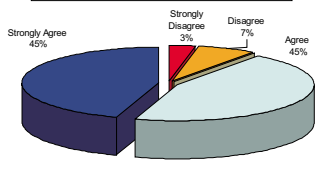
#### Examiner's opinion: Some examples

I consider that the results of this test can offer interesting information about the ability of a person to solve problems and his/her mental agility.



EFAl showed a good face validity from the examinees' and the examiners' points of view.

I consider that the results of this test can offer interesting information about the ability of a person to solve problems and his/her mental agility.



## Concluding remarks

It is infrequent to take into account the point of view of the examiners and the test-takers systematically and objectively in the development or adaptation of a test. However, the information that could be drawn from this procedure could be useful to improve the test quality (in tryout studies) as well as to establish its properties. As it has been presented in this poster, the inclusion of this complementary information has allowed us to detect specific aspects not identified by another source of information usually considered in the development of a test and improve empirically the quality of the test developed.

Readers are urged to view this method as a suggestive, rather than an established device, and they are encouraged to construct their own system to study the examiners and examinees perceptions and reactions.

REFERENCES:  
Baird, L. L. (1987). Do students think admissions tests are fair? Do tests affect their decisions?. Research in Higher Education 26 (4), pp. 373-388  
Chan D, Schmitt H, DeShon RP, Clause CS, Delbridge K. (1997). Reactors to cognitive ability tests: the relationships between race, test performance, face validity perceptions, and test taking motivation. Journal of Applied Psychology 82(2), pp. 300-310.  
Li, E. S. L. (1995). Defining face validity. In K.P.Y. Wong & C. F. Green (Ed.). Thinking language : issues in the study of language and language curriculum renewal. Hong Kong: Language Centre, Hong Kong University of Science and Technology.  
Nevo, B. & Sfez, J. (1985). Examinees' Feedback Questionnaires, Assessment & Evaluation in Higher Education, 10 (1), 236-248.  
Reynolds, C. R., & Kamphaus, R. W. (2003). RIAS: Reynolds Intellectual Assessment Scales. Lutz, FL: Psychological Assessment Resources, Inc. (Spanish adaptation: Santamaría, P. & Fernández, I., 2008, Madrid: TEA Ediciones).  
Santamaría, P., Arribas, D., Pereña, J. & Seisdedos, N. (2005). EFAl. Evaluación factorial de las aptitudes intelectuales. Madrid: TEA Ediciones.  
Stricker, L.J., Wilder, G.Z. & Bridgeman, B. (2006). Test Takers' Attitudes and Beliefs about the Graduate Management Admission Test. International Journal of Testing, 6, 255-268.