A contribution to the evidence of the scale validity of the PISA test

(Technical Report IEIA-INEE)
Submitted to the International Meeting of Evaluation in Higher Education.
Veracruz, Mexico. September 19, 2008
Colegio de Bachilleres de Veracruz
Instituto de Evaluación e Ingeniería Avanzada, S.C.

Agustín Tristán-López[1]. *Director*
Liliana Mendoza-González

Instituto de Evaluación e Ingeniería Avanzada, S.C.
IEIA. México.

María Antonieta Díaz-Gutiérrez. *Director*
Gustavo Flores-Vázquez
Roberto Solís-González
Damián Canales-Sánchez
Plácido Morelos-Mora
Yesenia de la C. Hernández

Internacional Projects.
Instituto Nacional para la Evaluación de la Educación.
INEE

[1] Correspondence may be addressed to: A. Tristán-López. ici_kalt@yahoo.com

A contribution to the evidence of the scale validity of the PISA test

The international OECD PISA 2006 test focused on the performance of Sciences of 15 years old students. The unsatisfactory results from Mexico were submitted to analysis, including multilevel models, to explain the origin of their deficiencies. It was clear that a differential functioning behavior or a design skewed against a specific country would produce serious validity problems. Besides the conceptual design of the test blueprint and the intrinsic quality of the items, an evidence of the quality of the test is provided by the Test Design Line, grounded on the distribution of the item difficulties calibrated with the Rasch model. The results of this approach show that the scale validity of the test is appropriate for Mexican students, an information that has to be known by practitioners and educational authorities.

Key words: Validity, Rasch model, testing, scale.

Mexico has participated during several years in the large international assessment of academic and non-academic skills for 15-year old youth among member and non-member countries of the OECD, named Programme for International Assessment (PISA). In the 2006 version of test, PISA explored the reading, mathematics and science skills, mainly focused on this third area, while the former two were the main purpose of previous versions in 2000 and 2003.

The PISA assessments interest field is the kind of literacy skills potentially needed at the post-secondary education or the labor market, instead of the contents or abilities provided by a school curriculum.

The design of the PISA sampling considers the possibility to compare the results of the participant countries, to show some ranking, specially on science skills and to provide ideas of opportunity areas of improvement for each country. An additional sampling has been requested by some countries, (for instance Canada and Mexico), in order to compare results at the intra-country level.

The test design is under the responsibility of the International Project Managers, while a translation and customization are considered for each country. Practically no local influence is added to the test, as it must follow exactly the same characteristics defined at the international level, including the item design, the test blueprint, the item scoring and calibration.

The main results from the PISA test are produced in a report by the OECD, but only at the international level. After the presentation of this report, every country has the option to produce its local report, comparison and analysis. The international report lacks of a specific chapter concerning the technical report of the test, for instance no evidence is provided concerning the validity or reliability of the test, excepting the possibility to obtain the data base of the results and every country, if needed, may produce its own interpretation of the data and the quality or the test.

Mexican results are unsatisfactory since the first administration of PISA, locating the national mean very low compared to the mean of the OCDE or even compared to Spain and other Latin American countries. Several reasons of these low results are under analysis at this moment by a task force specially organized by the Mexican Management of PISA, headed at the National Institute for the Educational Evaluation (Instituto Nacional para la Evaluacion de la Educacion, INEE). The set of hypothesis under preparation will be proved mainly using multilevel models, considering as

explanatory variables: socio-economical, parental, school background among other variables and indices included in the PISA data base.

A comprehensive multivariable study including structural equations, factor analysis and hierarchical linear models, has been performed, to show the relationship among socio-economical variables and the results on PISA (Tristan et al, 2008).

One of the hypothesis that has been managed since the first PISA version is based on this research question: How appropriate is the design of the PISA test and its items to Mexican students? This question rather corresponds to a suspicion than to a hypothesis, due to several reasons: some of the items do not match the Mexican context (train stations, airports, communications and some home possessions are out of the reach of many of the 15-year old students and their families and schools); Mexican education and assessments traditionally do not develop some problem solving or analysis skills (for instance, to provide a response to situations where a explicit question is not involved); but some other skills developed in Mexican schools are not part of the PISA test (attitudes vis-a-vis of the family, short-cut solution of some problems and social values are not considered on the test).

In addition of the item design, there are some technical requests concerning the test calibration and its difficulty. Evidences regarding the quality of the test are needed to confirm that there is no influence of test bias, differential functioning against some countries (Mexico in this case). Once these evidences could be provided, the purpose of the study and the reliability of the procedure could be acceptable.

The scale of the PISA test

An uniform distribution of the difficulties of the items is a requirement for validity, according to Wright and Stone (1979) or Bond and Fox (2001), following previous ideas suggested by former authors such as Loevinger (1947), with some practical implications of the model provided by Messick (1998). A valid scale fulfills the following criteria:

a. The latent variable defines a unidimensional variable in a single cartesian axis (from minus to plus or from low to high)
b. Items describing the latent variable may be located on the unidimensional scale. Its position corresponds to the mean of correct answers (raw score or logits).
c. The order of the items make sense for the latent variable, a higher measure of the item means “more” of the latent trait.
d. Items must be centered around the mean of difficulty of the latent variable, corresponding to the mean ability.
e. The set of items of the test must be distributed on the whole range of measure of the latent variable.
f. The measurement error at every point of the scale must be as constant as possible. This implies that the items must be uniformly distributed in the whole measurement range.

Figure 1 shows the Wright Map of the PISA 2006 output, according to the Rasch model. This model has been chosen because it is the most comprehensive tool to obtain objective measures of the

students and the items, and also because it is the model used by the OECD in the project. It is evident the bell shaped distribution of the students' measures and a quite uniform distribution of the items across the entire range of abilities.

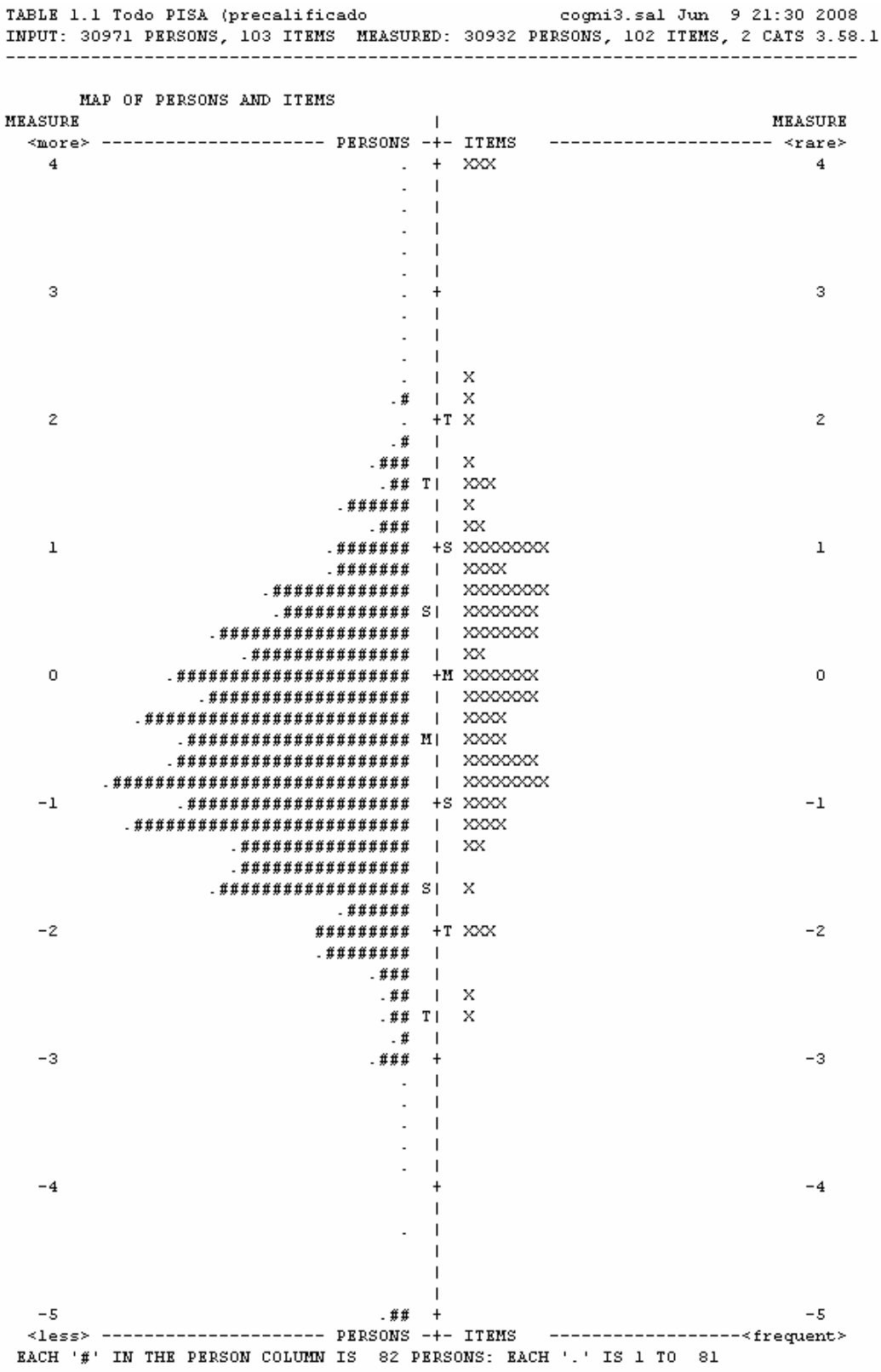

Figure 1. Wright map of the PISA test

Despite the importance of the Wright Map, it does not allow to verify the uniformity of the difficulties of the items, so another tool is needed, and the Test Design Line concept has been used.

Evidence of scale validity using Rasch calibrations

An evidence of quality is needed in addition of the many reasons that the OECD could provide, or the knowledge the countries may have about the seriousness and experience of the persons in charge of the test design. The objective evidence provided by the Test Design Line has been used here, following Wright & Stone (2004), Bond & Fox (2001) and Tristan & Vidal (2006).

The Rasch model mathematically provides a linear scale from any set of items, but in addition an evidence of the scale validity passes by the concept of scale validity, combining the attributes of invariance for a validity-centered design and the order and fit validities for the item distribution on a test. The Test Design Line is a paradigmatic model (not only a descriptive one), that gives the evidence for the following elements: (a) uniform distribution of items difficulties; (b) range of difficulties covering all the spectrum of person's abilities; (c) mean difficulty at the center of the interval; (d) no bias of the test difficulty.

To test the quality of the design, the model includes the mean absolute difference as a quantitative parameter that reflects the distribution of the items and the limits of the design of a given test.

According to Tristan & Vidal, the "Test Design Line" is defined by three values: TDL[LL,UL,MAD]. the equation of the line for N items, in the plane Difficulty versus number of items, is:

$$D=W(I-1)/(N-1)-LL \qquad [1]$$

Where:
D = Difficulty of item I (from 1 to N).
LL = Lower difficulty of the test, in logits. Suggested: LL = -1.5 logits.
W = Width of the test in logits (W = 2 x abs(LL)). Suggested: W=3.0 logits.

The discrepancies between observed and expected item difficulties can be calculated and the mean absolute difference (MAD) for N items is:

$$MAD = [\Sigma\, Abs(D_{observed} - D_{expected})]/N \qquad [2]$$

Where:
MAD = Mean absolute difference
D = Item difficulty

For real tests a MAD above 0.25 logits (called "¼ logit rule") indicates a high discrepancy among difficulties. Discrepancies may be due to: (a) the items are not uniformly distributed, (b) the mean of difficulties is far from zero or far from the persons mean, or (c) the width of the test is bigger than 3 logits. The thumb rule to identify an unacceptable test bias is more than 1 logit.

The graphical representation, and the MAD provide an evidence of the validity of the scale, as shown in the next section.

Evidence of scale validity of the PISA 2006 test

Using the data base provided by PISA, item calibration has been performed using Winsteps, with the default option centering the mean difficulty of the items in 0. Figure 2 compares the TDL[-1.5,+1.5] for PISA 2006 on the science skills. In this case MAD=0.13 indicating a very high fit to the theoretical distribution of the items according to the model.

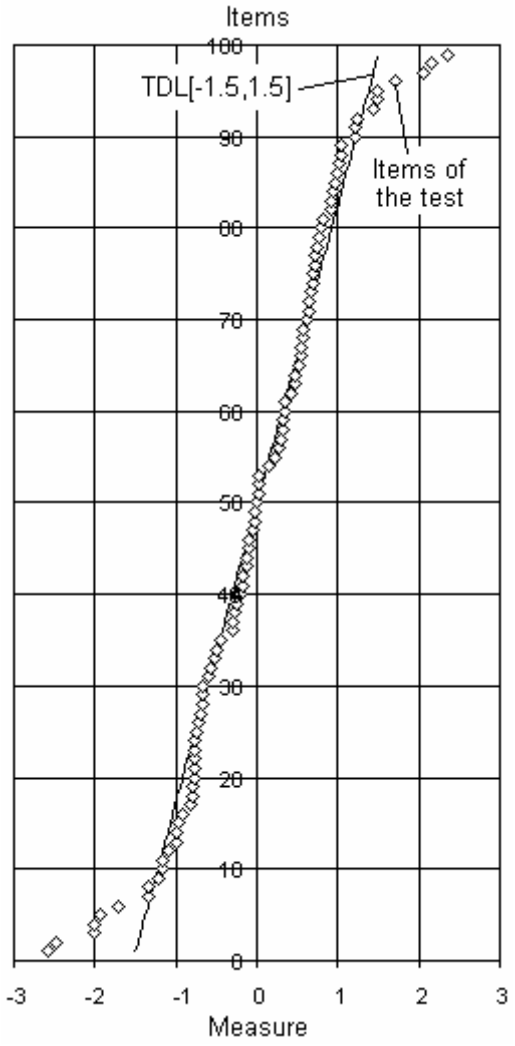

Figure 2. Theoretical and observed TDL (centered on the mean difficulty)

Winsteps allows centering the persons on their mean of ability in science, in such case the item calibrations now are correspond to the real mean difficulty of the test compared to the performance of the students.

The second snapshot of the test is shown in Figure 3, with the TDL[-1.5,+1.5], but now the test fit is not acceptable as the MAD=0.46, indicating that the test is difficult for the students tested with PISA 2006. It can be seen this fact by the shift of the observed item distribution to the right of the TDL, confirming that the test is more difficult than expected for the population, but this difficulty is less than a half logit.

The shift reported by MAD=0.46 is not a reason to invalidate the test or to consider that it is out of the reach of the population. Moreover, it is possible to suggest that the ability of the Mexican students must be increased, step by step, at least to the amount of this shift, following some pedagogical and political measures to improve the academic performance of the studens (not their ability to answer the items of the test).

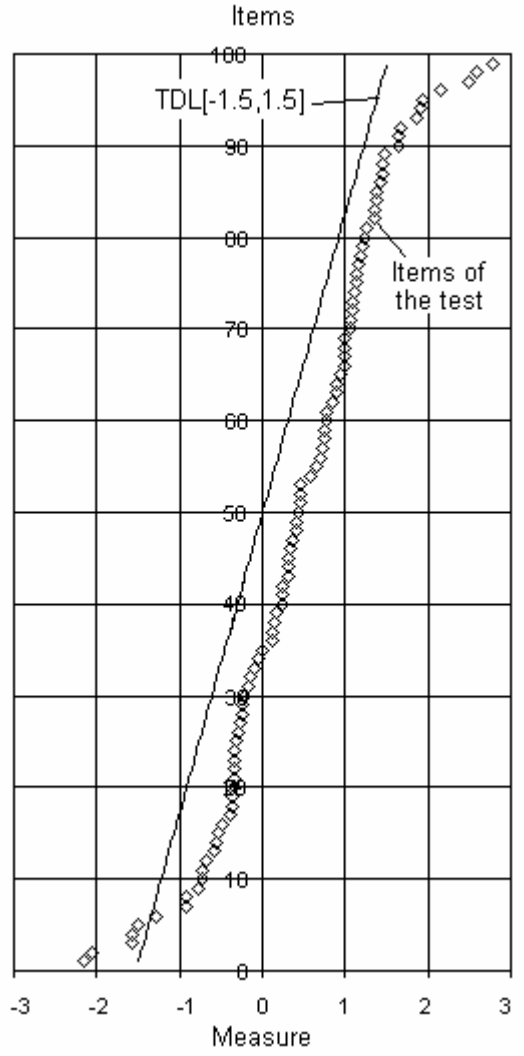

Figure 3. Theoretical and observed TDL (centered on the mean students' measure)

Conclusions

Two evidences of the validity of the PISA 2006 test have been obtained using the concept of the TDL combined with the Rasch calibration of the items. With these evidences, the multilevel studies to be performed in Mexico will confidently focus on the possible factors producing the low performance of 15-year old students, but nothing on the Mexican results can be imputed to a general defect of the test.

It is very interesting to verify that the TDL[-1.5,+1.5] provides a good theoretical model for the item distribution and the verification of the quality of a test produced independently by an international agency.

References

Bond T.G. & Fox C.M. (2001) *Applying the Rasch model.* Erlbaum, NJ Pp. 4-8.

Diaz G.M.A., Flores V.F. & Martinez R.F. (2007) *PISA 2006 en Mexico.* INEE, Mexico. 343 pp.

Linacre J.M. (2006) A user's *guide to Winsteps.* Available at Winsteps.com

Loevinger, J. (1947) *A systematic approach to the construction and evaluation of tests of ability.* Psychological Monographs. (61)4,14-32.

Messick, S. (1998) *Validity.* En R.L.Linn (ed) Educational measurement. pp 13-103. Washington DC: American Council on Education and National Council on Measurement in Education.

Tristan L.A. & Vidal U.R. (2006) *Linear Model to Assess the scale's validity of a test.* AERA Meeting, 2007. Session: New developments in Measurement Thinking. SIG-Rasch Measurement. Available as ERIC Document: ED501232.

Tristan, L.A. et al (2008) *Análisis multinivel de la calidad educativa en México ante los datos de PISA 2006,* Instituto Nacional para la Evaluación de la Educación, INEE. México. Pp. 188

Wright B.D. & Stone M.H. (1979) *Best test design.* MESA Press. Chicago.pp 133-140

Wright B.D. & Stone M.H. (1988) *Validity in Rasch measurement.* Research memorandum 54. MESA. University of Chicago. 12 pp.

Wright B.D. & Stone M.H. (2004) *Making measures*. The Phaneron Press.Chicago. USA. Pp.35-39.