Constructing a common scale for a multi-level test to enhance interpretation of learning outcomes

Jessica Wu and Rachel Yi-fen Wu
The Language Training and Testing Center
jw@lttc.ntu.edu.tw; rwu@lttc.ntu.edu.tw

Abstract

A criterion-referenced test may contain multiple levels. When a learner takes a higher level of a given test, his or her progress cannot be measured by directly comparing the scores from two different levels. The General English Proficiency Test (GEPT), a five-level criterion-referenced EFL testing system, was developed with reference to the English curriculum in Taiwan. The GEPT is taken by more than 500,000 learners annually for various purposes, among which assessing learning outcomes is the most common. Recently, an increasing number of schools have begun to encourage their students with better English proficiency to take the GEPT at a higher level. Under this circumstance, scores for more than one GEPT level are used in a single school setting. In other words, these schools compare the learning outcomes of students who do not necessarily take the same level of the GEPT. To establish more constructive relationships among teaching, learning, and assessment, it is desirable to create a common scale across all levels of the GEPT by vertical scaling (Kolen & Brennan, 2004).

To this end, the present paper reports procedures for constructing a vertical scale for four GEPT levels (Elementary, Intermediate, High-Intermediate, Advanced; roughly equivalent to CEFR A2-C1, respectively) through the use of Item Response Theory. The study employed non-equivalent groups with anchor test design, and the test items were grouped into three testlets, each containing shortened versions of two adjacent levels to prevent data from being contaminated due to test-taker fatigue. The between-item multidimensionality model was used to analyze the score data from a total of 1,270 learners. According to the concurrent estimations of learners' ability (θ) on the common score scale, the corresponding$\theta$of GEPT scores across levels were estimated (in logit). Thus, scores from different levels can be compared on the vertical scale. Major findings include

1. A good fit with Rasch model was obtained.
2. An increase of difficulty across the GEPT levels provided internal validation evidence to support the hierarchical sequence.
3. A difference of 1.0 logit was found between the Elementary level and the Intermediate level, and between the Intermediate level and the High-Intermediate

level. However, a wider gap of 1.2 logit was found between the High-Intermediate level and the Advanced level, suggesting that more learning is required for students to achieve the Advanced level.

However, it is worth noting that vertically linking tests of different levels provides a weaker degree of comparability than horizontally linking or equating. Despite the constraints on the inferences that vertical linking can support, it is recommended that test-takers' logit scores on the common scale be provided as a supplement to their GEPT scores when the test results are reported. In this way, the usefulness of the GEPT can be enhanced in order to better inform teaching and learning.

Key words：between-item multidimensionality, item response theory, Rasch model, vertical scaling

Reference:

Kolen, M.J. and Brennan, R.L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices (2nd ed.)*. New York: Springer-Verlang.