Creating a Common Score Scale for the GEPT to Support Interpretation of Learning Progress

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

ABSTRACT

Level-based tests have advantages over norm-referenced assessment for monitoring and tracking progress in order to support instructional planning and promote continued learning. These tests assess the degree of individual test takers' mastery in the specified domain that the test is designed to assess and compare each test taker's ability against predetermined performance standards, rather than against the ability level of other test takers as norm-referenced assessment does. The General English Proficiency Test (GEPT) is a level-based EFL testing system, developed with reference to the English curriculum in Taiwan to provide accessible attainment targets for English learners at different stages. GEPT scores have been widely used in school settings to measure learning outcome and evaluate students' progress in learning English. Students who pass one level of the GEPT are often motivated to take the test at the next highest level. When they take a higher level of the test, their scores cannot be compared straightforwardly with the scores they earned at lower levels. To facilitate comparisons of test results across levels, the scores from different levels must be converted and placed onto a common score scale.

The present study reports procedures for constructing vertical scales, also known as 'developmental score scales' (Tong & Kolen 2007:228), for four levels of the GEPT, spanning CEFR A2 to C1 levels. This study applied common-item non-equivalent groups design to link scores from different levels of the GEPT Listening and Reading tests onto two separate vertical scales based on the Rasch model estimation. A total of 1,270 students participated in this study. The results showed that the scaling of both Listening and Reading tests was effective, and that the relationships between the examinees' IRT ability estimates and their operational scores across the four GEPT levels were relatively linear. This paper documents the issues and complexities involved in creating vertical scales, suggests possible applications of the vertical scales, and discusses the limitations of the present study.

Key words: level-based tests, norm-referenced measurement, common-item non-equivalent design, vertical scales, Item Response Theory (IRT)