Establishing a Common Score Scale for the GEPT Elementary, Intermediate, and High-Intermediate Level Listening and Reading Tests

Rachel Yi-fen Wu & Cecilia Hsiu-yu Liao
The Language Training & Testing Center
rwu@lttc.ntu.edu.tw

Abstract

In recent years, criterion-referenced assessment (CRA) has been widely adopted to meet the growing need of stakeholders for more meaningful information about what real-world activities the test takers are able to undertake. CRA attempts to "specify abilities that make up language proficiency and to define levels of proficiency" (Brindley, 1991, p. 147). Numerous proficiency scales, aiming to reflect "a hierarchical sequence of performance ranges" (Galloway, 1987, p. 27), have been developed around the world in different contexts, yet only a few have been empirically supported. The General English Proficiency Test (GEPT) is a five-level criterion-referenced EFL testing system, devised in accordance with Taiwan's education system. The GEPT level framework has enjoyed widespread acceptance during the past decade in Taiwan. The aim of the present study was to investigate the pattern of differentiation across the GEPT levels in terms of difficulty and to examine the appropriateness of the performance standards of the GEPT level framework. The paper reports procedures for vertically linking different levels of the GEPT onto a common score scale on the basis of Item Response Theory (IRT). Linking different levels of the test on the same score scale facilitates statistical depiction of the patterns of increase of test difficulty and enables empirical investigation of the relationships between the proficiency levels of the passing candidates across test levels. The common-item non-equivalent groups design was employed. Three levels of the GEPT Listening and Reading test items were grouped into two testlets, each containing shortened versions of two adjacent levels to prevent parameter estimates from being contaminated by examinees' fatigue. Concurrent and separate estimations were performed with BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003). The results showed a relatively regular pattern of increase of difficulty across the GEPT levels, and provided internal validation evidence to support the GEPT level framework. It is worth noting that vertically linking tests of different difficulty provides a weaker degree of comparability than horizontal equating relationships. Therefore, there are limitations to the inferences that vertical scaling of test levels can support. It may be appropriate to use parameter estimates for low-stake tests and less prudent to use them for higher-stake purposes.

Key Words: vertical scaling, Item Response Theory (IRT), item parameter estimates, concurrent estimation, separate estimation