Investigating Rating Processes in an EAP Writing Test: Insights into Scoring Validity

Jessica R W Wu and May T M Ma
The Language Training and Testing Center
jw@lttc.ntu.edu.tw; mayma@lttc.ntu.edu.tw

Written language performance is always rated against a set of evaluation criteria. However, the application of the criteria is ultimately dependent upon how raters interpret them. The purpose of this study was to investigate raters' rating processes using an analytic rating scale developed for the General English Proficiency Test (GEPT) Advanced Level Writing Test, which assesses Taiwanese learners' written performance in an EAP context.

The study adopted both qualitative and quantitative methods. Qualitative data include concurrent think-aloud protocols of four GEPT raters during the process of rating 12 essays. A coding scheme mainly based on the criteria specified in the GEPT Advanced-Writing Analytical Scale (GEPT-A-WAS), namely Relevance and Adequacy (RA), Coherence and Organization (CO), Lexical Use (LU), and Grammatical Use (GU), was developed. To obtain a better understanding of the rating behaviors, the coding scheme also included raters' verbal reports about interpretation of the criteria and the difficulties encountered when making scoring decisions. Quantitative data consisting of the analytical and global scores awarded by the same four raters were analyzed by Many-facet Rasch measurement. This allowed us to investigate whether the raters' considerations of the rating criteria were reflected in the scores they awarded.

Results show that while raters remained close to the criteria, they were also strongly influenced by their intuitive impression of each essay when they first read it. Moreover, each rater tended to focus on particular elements (e.g., content, organization) while rating an essay. Therefore, when raters produced a set of scores, they seemed to undertake a process of reconciliation of the criteria, their overall impression of the essay, and the specific features of the essay. This suggests that the rating scale does not address all the essay elements that influence raters' decisions. The raters' verbal protocols also suggest that they used certain strategies in order to decide test-takers' scores for specific criteria; for example, a rater might look for topic sentences to determine whether test-takers have a good command of organizational skills, despite the fact that the use of topic sentences is not specified in the scale. Analysis of the quantitative data indicates that despite the different paths raters might have taken to reach their scoring decisions, inter-rater reliability remains high. The quantitative data also show that while the analytical scores were highly correlated to the global scores, the CO and RA criteria tended to be scored more leniently yet received the most attention according to raters' verbal reports.

The findings are of significance in informing the ongoing improvement of the GEPT-A-WAS and rater training, especially with regard to the possible amendment of the scale wording and discussion of strategies for determining scores for each criteria during rater training. It is hoped that by adopting these measures, the variation in the rating of the GEPT Advanced Level Writing Test could be reduced further and enhance scoring validity.