**An Analysis of Variability among Raters in a High-Stakes EFL Writing Examination**

|  |  |  |
| --- | --- | --- |
| Sheila Parveen Lallmamode  Preparatory Year Programme  King Saud University | Assoc. Prof. Dr. Noorlide Abu Kassim  International Islamic University Malaysia | Ebtisam Abdul Halim  Preparatory Year Programme  King Saud University |

This study investigated the variability in the ratings of EFL essays among raters. Evidence of rater variability or rater-related effects operating in performance assessment are considered sources of construct-irrelevant variance. They may obscure the construct being measured and, thus, threaten the validity and fairness of the assessment. The study addressed the following question: what type of rater-related variability or rater effects were found among raters in a high-stakes EFL writing examination?

A small group of six trained raters, with many years of teaching experience, participated in this study which was conducted at the preparatory year programme, King Saud University. Four of the raters had English as their first language, while the two others had Arabic and Amharic (an Ethiopian language) as first languages. They were all experienced English language teachers at the Intensive English Programme and were very familiar with the rubric used to rate the writing task. The four-level (from 0 to 3) and four-part analytic rubric (which comprised of ‘task fulfillment’, ‘lexis and grammar’, ‘cohesion and coherence’, and ‘spelling and punctuation’), was developed intuitively using other existing writing rubrics. The data for analysis were ratings assigned by these raters on the final writing examination of fifty-eight preparatory year students taking the summer semester at King Saud University.

Applying the Many-facet Rasch measurement methodology, the ratings were analysed with the Rasch measurement computer programme FACETS (Linacre, 2010), where estimates of the performance of several facets of the assessment (raters, items, examinees) can be obtained. To examine a rater’s internal consistency, several indices in the FACETS analysis were looked into to determine if a rater displays halo, central, random, and extreme rating patterns.

The results indicated that among the six raters, only one showed idiosyncratic rating behavior. The other five raters were considered fit. The misfitting rater, who has English as her first language, displayed random rating patterns indicating that she used the rating scale in an unreliable manner, that is different from the other raters.

The study has implications for the utility of rater training as familiarity with a rubric and many years of teaching experience are not necessarily adequate to help raters provide reliable ratings.

[357 words]