DEVELOPMENT OF TEFL TEACHERS' SKILLS IN ASSESSING STUDENT WRITING THROUGH PRE- TO IN-SERVICE

Kong, Yunjun

Doctoral School of Education, University of Szeged

Research has shown that TEFL (teaching English as a foreign language) raters differ from each other when grading student writing. For example, experienced teacher raters tend to be severer than less experienced ones, trainee self-raters assess their own texts lower than their peers do, etc. However, the changes in teachers' assessment skills from pre-service to in-service does not get enough attention in this regard. The current cross-sectional study aimed to compare teachers' and trainees' evaluation of one selected student text by an English learner. Three samples at different stages of professional development in China were involved in this study: 59 pre-practicum trainees, 31 post-practicum trainees, and 32 teachers. They were asked to rate the same student text on a five-point Likert scale from six aspects: holistic, content, structure, style, grammar, and mechanics. Many-Facet Rasch Measurement (MFRM) was used to analyze their severity toward the text as well as the assessment criteria. Results showed that most pre-practicum trainees were severe toward the text, while the minority were relatively lenient, indicating somewhat not equal severity between inter-raters (separation index=1.72, separation reliability=.75, χ^2=213.2, p<.05). Pre-practicum trainees were harshest on criteria regarding the structure and were the most lenient on holistic scoring. Similarly, post-practicum trainees rated harshly but with a wide range of measure logits from -4.02 to 3.27, illustrating significant differences between raters (separation index=2.06, separation reliability=.81, χ^2=165.3, p<.01). As for the criteria, they scored most severely on style and more leniently on content, structure, and the holistic, compared to grammar and mechanics. By contrast, most teachers were lenient toward the text with measure logits below 0, but still, significant differences were found between teacher raters (separation index=2.33, separation reliability=.84, χ^2=203.4, p<.01). Interestingly, teachers also graded most strictly on style and tended to be most lenient on content and holistic scoring. Furthermore, results also showed that some raters in each group tended to misfit or overfit the MFRM model. Namely, 12 out of 59 pre-practicum trainees, 3 out of 31 post-practicum trainees, and 5 out of 32 teachers were respectively above the misfit values (greater than 1.5), indicating an overly inconsistent rating behaviors among these raters; and 7 of the pre-practicum trainees, 4 of the post-practicum trainees, and 7 of the teachers were respectively below the overfit values (smaller than .5), indicating they were too consistent. The findings show that there is an obvious change of participants' severity from pre- to in-service when evaluating the student text. Yet it is noticeable that inconsistent rating performance was found between raters in each sample. The findings may support training in both pre- and in-service programs to target the development of assessment skills.