A Close Look at English Teacher Employment Examinations (ETEEs): How do raters assess?

Tomoyasu Akiyama

Department of English language & Literature, Bunkyo University
akitomo@koshigaya.bunkyo.ac.jp

Abstract

This paper focuses on an investigation of teacher employment examinations, focusing on how raters assess prospective English teachers. In order to become an English teacher at public junior and senior high schools, candidates are required to pass the first test (knowledge-based tests) and the second test (performance-based tests). This paper focuses on the second test, investigating how raters rate candidates’ demonstration of teaching skills via microteaching in TEEs.

The main original purpose of microteaching was to assist teachers in identifying their weaknesses and in developing their teaching (e.g., Monk, 2008). However, serious problems arise in cases where microteaching is used to assess candidates for English teaching positions in the high stakes context.

21 candidates undertook microteaching performance tests and four raters rated all candidates according to criteria used in the actual testing context. Data were analyzed using Facets Rasch software (Linacre, 2006), which mainly administers bias analyses based on interactions between raters and candidates. Results showed that approximately 20% of the interactions between the raters and candidates were identified as bias. Further analyses were conducted, investigating how raters assessed candidates and what might have caused biased results using think-aloud protocol.

Finally, this paper discusses issues with the use of microteaching in the high stakes context from the point of view of validity and reliability.

Keywords
microteaching, validity, reliability, teacher employment examinations,

1 Introduction

1.1 The purpose of this study

In order to become teachers at public schools in Japan, all candidates have to take teacher employment examinations (hereafter, TEEs).

With a candidate’s employment decision made on the total scores of TEEs, TEEs require candidates to undertake two types of tests, the first test (knowledge-based tests) and the second test (performance-based tests). This paper exclusively focuses on the second test, assessment of candidates’ teaching performance via microteaching and investigates how raters rate candidates’ demonstration of teaching skills.

1.2 Procedures of microteaching at the second test

There are 3 steps in microteaching testing context: 1) preparation, 2) demonstration of teaching and 3) assessment. At the preparation stage (1st step), a candidate or a cohort of candidates makes a teaching plan (e.g., an introduction of a grammatical target). Then, the candidates are required to demonstrate their teaching skills based on the teaching plan in front of assessors or the group of candidates (2nd step). After finishing his/her demonstration, raters (e.g., educational board members, school principals and English teachers) assess their performances based on assessment criteria (3rd step).

2 Research questions and data collection

2.1 Research questions

Two research questions are established in this study in order to investigate rating patterns, focusing on how raters rate candidates’ demonstration of teaching skills.

RQ 1: To what extent do raters assess candidates’ performance reliably?
RQ 2: What percentages are there in any biased interactions between raters and candidates?

2.2 Data collection

2.2.1 Participants and tasks used in this study

Twenty-one university students participated in this study. All participants were required to introduce one of the two target grammar points: 1)
past regular verb (-ed, -d) and 2) there are -. These were chosen because all are in the 1st and 2nd year English textbooks and so were actually carried out in TEEs.

2.2.2 Raters and assessment criteria

Four raters were invited to participate in this study. Two raters were junior high school teachers, who had more than 25 years of teaching experience and the others were officers at educational boards, who had assessed candidates several times in TEEs.

Six assessment criteria were chosen: 1) lesson flow, 2) instruction ability, 3) delivery, 4) personality 5) expertise and 6) overall judgment. The range of all scales was 1 to 6 points and the minimum score is 6 and the maximum score is 36 points.

3 Results and discussion

Table 1 below shows rater severity, model standard error, and infit mean square based on Rasch analysis, indicating that Rater A, who is the positive highest value among four raters, is identified as the harshest rater while Rater B with the highest negative value is the most lenient rater.

<table>
<thead>
<tr>
<th>Rater ID</th>
<th>Severity</th>
<th>Model S.E.</th>
<th>Infit Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.97</td>
<td>0.15</td>
<td>1.12</td>
</tr>
<tr>
<td>B</td>
<td>-0.46</td>
<td>0.15</td>
<td>0.67</td>
</tr>
<tr>
<td>C</td>
<td>-0.08</td>
<td>0.15</td>
<td>1.43</td>
</tr>
<tr>
<td>D</td>
<td>-0.44</td>
<td>0.15</td>
<td>0.71</td>
</tr>
<tr>
<td>Mean</td>
<td>0.00</td>
<td>0.15</td>
<td>0.98</td>
</tr>
</tbody>
</table>

(Reliability 0.95)

In terms of rater consistency, infit mean squares in the fourth are not within the acceptable range (0.7 to 1.3, McNamara, 1996) except Raters A and D. This means that Raters B and C rated candidates inconsistently. In particular, rating patterns produced by Rater C are more varied than those predicted by Rasch model (RQ1).

The next investigation (RQ2) is to identify the extent to which raters show extremely harsh or lenient but systematic rating patterns (so-called bias analysis). Multi-faceted Rasch model (Linacre, 2006) identifies such patterns and gives us a signal as Z-scores, either more than + 2 or less than - 2. The former means that a rater assesses candidate more leniently than Rasch model expected and the latter indicates the opposite. Of 84 interactions between 21 candidates ×4 raters, 16 interactions (approximately 20 %) were identified as ‘biased’ (Figure 1).

Further analyses were conducted to investigate how raters assessed candidates and what might have caused biased results using think-aloud protocol, focusing on the 16 interactions stated above. Those results will be revealed at the conference.

4. Conclusion and implications for further study

The purpose of this study was to investigate rater behaviors, focusing on the extent to which there are biased interactions between raters and candidates and the extent to which biased interactions would have on candidates. It was found that raters differed in terms of severity, but that they were relatively consistent in their overall rating patterns except those of Raters B and C. Two questions still remain:

1) What factors account for the differences in terms of rater severity or leniency when they assess test-takers’ performance in ETEEs?

2) To what extent would these rater differences have an impact on final employment decisions of English teachers?

Also there are a number of implications for further study:

- What aspects of teaching skills (i.e. task) should be assessed?
- What assessment criteria should be employed?

References


This research was supported by the Grant-in-aid for Scientific Research (C) (22520580).