

Rasch 모형을 활용한 국어교사의 쓰기 평가 특성 분석

—중학생 설명문 쓰기 평가를 중심으로—

박영민* · 최숙기**

<차 례>

- I. 서론
- II. 국어교사 요인과 쓰기 평가
- III. 연구 과정 및 연구 방법
- IV. 연구 결과 및 논의
- V. 결론

I. 서론

국어교사는 학생 글을 평가할 때 자신의 배경지식과 같은 인지 요인, 동기나 신념, 태도와 같은 정의 요인, 피로도 등과 같은 상황 요인 등으로부터 영향을 받는다. 이러한 요인이 국어교사의 쓰기 평가에 영향을 미치는 것은 매우 자연스러운 일이지만, 쓰기 평가의 신뢰도가 흔들리는 문제를 초래하기도 한다. 그래서 이러한 요인들은 엄격하게 통제되어야 할 요인으로 간주되어 왔다(Huot, 1990, 1996 ; Penny, Johnson, & Gordon, 2000 ; 박영민, 1999, 2008 ; 박영민, 2009a, 2009b ; 박영민 · 최숙기, 2010a, 2010b).

그러나 쓰기 평가에 영향을 미치는 요인을 효과적으로 통제하기 위해

* 제1저자, 교신저자, 한국교원대학교 국어교육과 교수

** 제2저자, 한국교원대학교 국어교육과 강사

서는 국어교사가 보이는 쓰기 평가의 특성이 어떠한지를 파악하는 것이 우선되어야 한다. 국어교사가 보이는 쓰기 평가 특성이 어떠한지를 파악해야 쓰기 평가 과정에서 통제되어야 할 영향 요인이 더 구체적으로 드러날 수 있기 때문이다. 이 점에서 국어교사의 쓰기 평가 특성을 분석하는 것은 중요한 의미가 있다.

쓰기 평가의 영향 요인을 통제하려는 목적이 아닐지라도, 학생 글에 대한 평가자로서 국어교사의 쓰기 평가 특성을 분석하는 것은 그 자체로서도 큰 의미가 있다. 학교에서 쓰기 평가가 지속적으로 이루어지고 있음에도 불구하고, 쓰기 능력에 대한 직접 평가의 중요성이 강조되고 쓰기 평가의 적용 범위가 점점 더 확대되고 있음에도 불구하고, 쓰기 평가에서 드러나는 국어교사의 특성에 대한 정보는 거의 확보되어 있지 못하기 때문이다. 예를 들면, 우리나라의 국어교사들은 학생 글을 얼마나 엄격하게 평가하고 있는지, 평가 기준을 얼마나 일관성 있게 적용하고 있는지, 어떤 평가 요인에서 평가 점수가 가장 높은지 또는 낮은지 등에 대한 정보가 거의 없는 실정이다.

그러므로 이 연구에서는 중학생 설명문 평가 결과를 자료로 활용하여 국어교사가 보이는 쓰기 평가의 특성을 엄격성(severity), 일관성(consistency)으로 구분하여 분석하고자 한다. 더 나아가 국어교사의 성(性)과 경력에 따라 쓰기 평가의 엄격성과 일관성이 어떠한 차이가 있는지를 분석함으로써 국어교사의 쓰기 평가 특성을 구체적으로 확인하고자 한다. 이 과정에서 쓰기 평가자로서 국어교사들이 가지고 있는 판정 오차도 분석하게 될 것이다. 이러한 국어교사의 쓰기 평가 특성 분석을 위해 이 연구에서는 국어교사의 중학생 설명문 평가 결과를 자료로 활용하고자 한다. 분석에는 5국면 Rasch 부분 점수 모형이 적용될 것이다.

이 연구는 표본으로 선정된 일부 국어교사의 쓰기 평가 결과를 활용하였다는 점에서 연구 결과를 일반화하는 데 한계가 있다. 그러나 국어교사가 보이는 쓰기 평가의 엄격성과 일관성을 분석하여 국어교사의 쓰기 평가 특성을 분석했다는 점에서, 또한 이 연구의 결과는 국어교사의 쓰기 평가 특성을 이해하고, 쓰기 평가의 전문성 신장을 위한 기본적인 정보로

활용될 수 있다는 점에서 연구의 의의를 찾을 수 있다. 이 연구의 결과는 학생 쓰기 평가에 대한 국어교사의 전문성 신장을 추구하는 데 필요한 정보를 제공하게 될 것이다.

II. 국어교사 요인과 쓰기 평가

쓰기 평가에 영향을 미치는 국어교사의 요인은 매우 다양하다. 국어교사가 가지고 있는 특성적인 요인들은 쓰기 평가에 직·간접적인 영향을 미치면서 국어교사들의 평가자 특성을 형성한다. 이러한 평가자 특성은 평가 결과가 얼마나 신뢰할 수 있는지를 판별하는 근거로 쓰이기도 한다. 이에 따라 쓰기 평가 연구에서는 이러한 주제를 지속적으로 다루어 왔다.

쓰기 평가에 영향을 미치는 국어교사 요인으로 주목을 끌었던 것은 국어교사의 지식이다. 국어교사의 쓰기 관련 지식은 좋은 글의 요건이나 쓰기에 대한 관점, 문장에 대한 관점이나 인식과 관련된 내적 준거 형성에 영향을 미침으로써 학생 글 평가에 영향을 미친다. 국어교사가 암묵적으로 가지고 있는 좋은 글, 좋은 문장 등에 관한 내적 준거는 학생 글을 평가할 때 우열을 재는 중요한 척도로 기능하기 때문이다. 그래서 국어교사의 쓰기 지식은 학생 글 평가할 때 복합적이면서 메타적으로 영향을 미치며, 따라서 쓰기 평가 전문성의 주요 요인으로 다루어야 할 필요가 있다(Runder, 1992 ; Engelhard, 1994 ; 이성영, 2005).

국어교사의 인지 양식이나 쓰기 신념 등도 쓰기 평가에 영향을 미치는 것으로 알려져 있다. 민태범(1991)에 따르면, 국어교사를 장(場) 의존적 인지 양식을 보이는 집단과 장 독립적 인지 양식을 보이는 집단으로 구분하였을 때, 후자의 국어교사들은 장 의존적인 국어교사들과 달리, 학생 글에 대해 비판적 비평을 하는 경향이 강했으며, 학생 글의 성취 수준이 낮을 때에는 심지어 불쾌감을 드러내기도 하는 것으로 나타났다. 실제로 교사 130명이 채점한 논술 답안을 분석한 결과, 장 독립적 집단은 장 의존

적 집단보다 평균 점수가 낮았다. 이러한 결과는 국어교사의 인지 양식이 쓰기 평가에 영향을 미친다는 사실을 잘 보여준다.

국어교사의 쓰기 신념은 교류 신념(transactional writing belief)과 전달 신념(transmissional writing belief)으로 분류할 수 있다(White & Brunning, 2005). 쓰기 신념은 학생 글을 대하는 국어교사의 인식과 태도를 매개 요인으로 하여 쓰기 평가에 영향을 미친다. 56명의 현직 국어교사를 대상으로 하여 쓰기 신념에 따른 쓰기 평가의 차이를 분석한 가은아(2008)에 따르면, 글의 조직 요인에서 전달 신념을 지닌 국어교사들이 학생 글을 평가할 때 더 높은 점수를 부여하는 것으로 나타났다. 이러한 결과는 쓰기 신념과 같은 국어교사의 정의 요인이 쓰기 평가에 영향을 미친다는 사실을 구체적으로 보여준다.

국어교사의 성(性)도 학생 글을 평가할 때 영향을 미친다. 이 주제를 직접적으로 다룬 Peterson & Kennedy(2006)는 남녀 교사 108명에게 학생들이 작성한 서사문과 논설문을 서술의 형식으로 평가하게 하고, 그 차이를 성별에 따라 분석하였다. 연구 결과, 교사 성별에 따라 학생 글에 대한 논평(comments), 즉 평가 내용이 서로 다른 것으로 나타났다. 여교사는 글의 규칙과 조직에 대해 중점을 두었으나, 남교사는 미적인 문체에 초점을 두었다. 이러한 성별에 따른 평가의 차이는 Barnes(1990)와 Roulis(1995)에서도 확인할 수 있다. 이 두 연구에 따르면, 평가의 양적 수준에 있어서도 여교사들은 일반적으로 남교사들보다 더 많은 논평을 남겼다. 국어교사의 성별에 따른 쓰기 평가의 차이는 수량 평가에서도 일부 확인되었다. 중학생들이 작성한 논설문을 국어교사들에게 평가하게 했을 때, 통계적으로 유의하지는 않았지만 여교사가 남교사에 비해 엄격하게 평가하는 경향이 발견되었다(박영민·최숙기, 2009).

학생 글을 평가할 때는 국어교사의 학생 성별 인식도 쓰기 평가에 영향을 끼친다. 즉, 국어교사가 글을 작성한 학생을 남학생으로 인식하는가, 아니면 여학생으로 인식하는가가 쓰기 평가에 영향을 미치는 것이다. 이러한 주제를 다룬 Peterson(1998)은 여학생이 쓴 것으로 인식한 글보다는 남학생이 쓴 것으로 인식한 글에 대해 국어교사들이 더 많이 수정 사항을

지적하고 더 많이 비평한다는 사실을 밝혀냈다. 국어교사들은 여학생들보다는 남학생들에게 쓰기 평가를 통한 지도나 안내가 더 많이 필요하다고 생각하고, 그 결과 남학생이 쓴 것으로 인식한 글에 대해 설명, 제안, 지도 내용을 더 많이 남겼던 것으로 분석되었다.

이와 함께 Peterson(1998)은 국어교사가 학생 글을 평가할 때 동성(同性) 평가 절하(same-sex depreciation)의 경향을 보인다는 점을 분석함으로써 국어교사의 성별 인식의 영향력을 실증적으로 보여주었다. 국어교사의 동성 평가 절하 현상이란, 국어교사는 자신의 성과 동일한 성을 가진 학생이 쓴 글에 대해 더 엄격하게, 더 부정적으로 평가하는 현상을 일컫는다. 즉, 남자 교사는 남학생이 쓴 것으로 인식한 글에 대해서 비판을 더 많이 하고, 여자 교사도 여학생이 쓴 것으로 인식한 글에 대해서 비판을 더 많이 하는 현상을 말한다.

이러한 동성 평가 절하의 경향은 다른 선행 연구에서도 이미 확인된 바 있다(Erbaugh, Houtler, & Prasn timer, 1988; Haswell & Haswell, 1995, 1996). 선행 연구에 따르면, 남교사들은 여학생이 쓴 글보다 남학생 글에서 더 많은 수정 의견을 제시하고 비판을 하였다. 남교사들은 남학생이 썼다고 생각하는 글보다는 여학생이 쓴 것으로 보이는 글을 더 칭찬하는 경향이 있었다. 이와 달리, 여교사들은 여학생이 쓴 것으로 인식된 글보다는 남학생이 쓴 것으로 보이는 글을 더 많이 칭찬했다.

Haswell & Haswell(1995)에 따르면, 학생 글을 수량적 방법으로 평가하는 경우에도 동성 평가 절하 현상은 동일하게 존재한다. 남자 평가자는 여학생이 썼다고 생각하는 글에 더 높은 점수를 주고, 여자 평가자는 남학생이 썼다고 생각하는 글에 더 높은 점수를 부여하였다. 이러한 연구들은 교사의 학생 성별에 대한 인식이 쓰기 평가에 영향을 미친다는 점을 잘 보여준다.

Ⅲ. 연구 과정 및 연구 방법

1. 연구 대상

국어교사의 쓰기 평가 특성을 분석하기 위하여 국어교사 68명을 표본으로 하여 수집한 평가 자료를 활용하였다. 중학생 설명문 평가에 참여한 국어교사는 전국적인 수준에서 선정하였다. 이 평가 자료는 박영민·최숙기(2010a)의 모평균 추정에서도 활용되었으나, 이 연구에서는 일부 결측이 포함된 자료 1건을 제외한 68명의 자료를 분석 대상으로 삼았다.

국어교사의 하위 분석 요인으로는 국어교사의 성(性)과 경력을 설정하였으며, 표본을 선정하여 자료를 수집할 때 이를 반영하였다. 표본으로 선정된 국어교사는 남교사 12명(17.65%), 여교사 56명(82.35%)이었다. 경력은 1~5년, 6~10년, 11~20년, 20년 이상으로 구분하였다. 이를 정리하면 <표 1>과 같다.

<표 1> 연구 대상

성별 경력	남	여	계
1~5년	3	24	27
6~10년	3	24	27
11~20년	5	8	13
20년 초과	1	0	1
계	12	56	68

2. 검사 도구

이 연구에서 평가 자료를 수집할 때 적용한 학생 글은 ‘자신이 관심 있거나 좋아하는 책, 영화, 운동(스포츠), 취미 활동 등을 학급 친구들에게

소개하고 이를 좋아하게 된 이유를 설명하라’는 과제에 따라 작성된 중학생 설명문이다. 연구 과정에서는 68명의 국어교사가 중학생 설명문 35편을 모두 평가하였지만, 분석 과정에서는 그 중에서 학생 글 6편에 대한 평가 자료만 활용되었다. 이는 무료로 공개된 FACETS를 활용하여 Rasch 모형을 적용하였기 때문이다.

학생 글은 필체나 학생의 성(性) 등 주관적 영향 요인을 통제하기 위하여 워드 프로세서로 입력하여 평가하였다. 입력 과정에서 의미의 오해를 초래할 수 있는 띄어쓰기는 부분적으로 수정하였다. 그러나 맞춤법 오류는 평가 기준에 포함되어 있는 요소이므로 수정하지 않았다. 평가자인 국어교사들에게는 쓰기 평가 기준표와 채점표를 함께 제공하였다. <표 2>의 쓰기 평가 기준표는 Spandel & Culham(1996)이 제시한 총체적 평가 기준을 중학생의 설명문 평가에 적합하도록 수정하여 적용한 것이다. 쓰기 평가의 하위 요인은 내용, 조직, 표현, 단어 선택, 형식 및 어법으로 구성되었다. 평가 기준에는 1점, 3점, 5점에 해당하는 평가 기준을 제시하였으나, 1점부터 5점 사이에서 자유롭게 점수를 부여할 수 있게 하였다.

〈표 2〉 Spandel & Culham(1996)의 평가 기준표

평가 요인	평가 기준
내용	• 5점 : 글의 중심 내용(주제)가 명료하며 독자의 주의를 끈다. 세부적인 내용들은 전체적인 중심 내용(주제)과 부합한다. • 3점 : 글의 중심 내용(주제)이 다소 명확하지 못하다. 글 전체의 중심 내용 주제와 부합하지 않은 세부 내용이 들어 있다. • 1점 : 글의 중심 내용(주제)이 잘 드러나 있지 않다. 세부 내용은 글의 전체적인 중심 내용과 잘 어울리지 않는다.
조직	• 5점 : 설명 대상의 특성을 고려하여 조직되었으며, 중심 내용이 잘 드러나도록 조직되었다. 내용의 순서나 구조가 독자가 이해하기 쉽도록 되어 있다. • 3점 : 설명 대상의 특성은 고려하였으나, 글의 구조가 명확하게 드러나지 않는다. 독자가 이해하는 데 어려움이 따른다. • 1점 : 설명 대상의 특성이 고려되지 않았으며, 글의 구조가 명확하게 드러나지 않는다. 독자가 이해하는 데 어려움이 따른다.
표현 (어조 및 태도)	• 5점 : 독창적이며 흥미롭게 표현되어 있으며, 독자가 쉽고 정확하게 이해할 수 있도록 표현되었다. 필자의 주제적인 목소리도 드러난다.

평가 요인	평가 기준
표현 (어조 및 태도)	• 3점 : 독자가 쉽게 이해할 수 있도록 표현되었으나, 독창성이나 흥미는 다소 떨어진다. 필자의 주체적인 목소리도 잘 드러나지 않는다. • 1점 : 내용만을 기계적으로 나열하여 글의 생동감이 떨어지며 흥미를 주지 못한다. 독자가 쉽게 이해할 수 있도록 표현되지 않았다.
단어 선택	• 5점 : 내용을 정확히, 흥미롭게, 자연스럽게 전달할 수 있는 단어가 선택되었다. • 3점 : 대체적으로 단어 선택이 내용 전달에 무리가 없으나 부적절한 단어들도 포함되어 있다. • 1점 : 내용을 전달하는 단어가 매우 제한적이어서 단어의 선택이 풍부하지 못하다.
형식 및 어법	• 5점 : 표준적이며 모범적인 쓰기 형식이 잘 드러나 있다. (어법, 구두점, 철자, 단락 구분 등) • 3점 : 제한된 범위에서만 글의 표준적 형식이 확인된다. • 1점 : 철자, 구두점, 문법에서 잘못된 것이 많아 내용을 파악하며 읽는 것이 어렵다.

3. 연구 절차

국어교사의 설명문 쓰기 평가 점수 자료를 수집하기 위하여, 현직 국어교사 68명에게 중학생들이 작성한 35편의 설명문과 평가 기준표, 채점표, 개인 기록표를 제공하였다. 평가자로 참여한 국어교사들은 평가자 훈련을 거치지 않고 개별적으로 평가를 수행하였다. 국어교사의 평가 자료는 2009년 1월부터 3월까지 수집하였다. 국어교사들은 평가 기준에 따라 5점 척도로 내용, 조직, 표현, 단어 선택, 형식 및 어법 점수를 각각 채점하였다.

4. 분석 방법

국어교사의 쓰기 평가 특성을 파악하기 위하여 이 연구에서는 평가자로서의 국어교사, 국어교사의 성별, 경력, 설명문 평가 요인, 평가 대상으로서의 중학생이라는 5국면을 Rasch 모형에 적용하여 분석하였다. 이를 통해, 중학생 설명문에 대한 쓰기 평가에서 성(性)이 p 이고 경력이 q 인 국

어교사(평가자) j 가 설명문 쓰기 능력이 n 인 중학생(피험자)의 설명문 쓰기 평가 요인 i 에 대해 평가 점수가 $k-1$ 이 아닌 k 를 부여할 확률과, 그 확률을 $\log$ 로 변환한 값을 얻을 수 있다(Linacre, 1989). 이러한 분석 모형을 정리하면 <표 3>과 같다. 쓰기 평가 결과를 Rasch 모형에 따라 분석할 때에는 평가 요인이나 평가 기준을 문항으로 간주하므로, 이 연구에서는 <표 2>의 평가 요인을 각각의 문항으로 처리하였다.

<표 3> 쓰기 평가 특성 분석을 위한 Rasch 분석 모형

$$\log(P_{nijpqk} / P_{nijpq(k-1)}) = B_n - D_i - C_j - U_p - T_q - F_k$$

P_{nijpqk} = 성(性)이 p 이고 경력이 q 인 국어교사(평가자) j 가 평가 요인 i 에 대해 점수 k 를 줄 확률

$P_{nijpq(k-1)}$ = 성(性)이 p 이고 경력이 q 인 국어교사(평가자) j 가 평가 요인 i 에 점수 $k-1$ 을 줄 확률

B_n = 중학생(피험자) n 의 설명문 쓰기 능력

D_i = 설명문 평가 요인 i 의 난도(難度)

C_j = 국어교사(평가자) j 의 엄격성

U_p = 성(性)이 p 인 국어교사(평가자) j 의 엄격성

T_q = 경력이 q 인 국어교사(평가자) j 의 엄격성

F_k = 평가 척도 $k-1$ 에 대한 척도 k 의 난도

<표 3>과 같이 각 국면의 분석 결과를 logit 척도로 변환하면, 중학생 설명문 쓰기 능력(점수), 평가 요인의 난도(難度), 성과 경력에 따른 국어교사의 엄격성 수준을 높은 수준으로부터 내림차순으로 정리할 수 있어 국어교사의 쓰기 평가 특성을 효과적으로 파악할 수 있다. 특히 국어교사(평가자)에 대한 모형 적합도 수치는 국어교사의 평가 점수가 중학생(피험자)의 설명문 쓰기 능력을 얼마나 정확하게 추정하는지에 대한 정보를 제공해 준다. 성별 및 경력별 적합도 통계치는 평가자인 국어교사가 피험자인 중학생들의 설명문 쓰기 능력을 일관성 있게 평가하고 있는지에 대한 정보를 제공하여 준다.

연구 과정을 통해 수집한 자료는 무료로 공개된 FACETS Version 2.0 프로그램(Linacre, 1999)을 이용하여 분석하였다. 분석은 국어교사의 쓰기 평

가 특성에 해당하는 엄격성과 일관성을 중심으로 수행되었으며, 평가 기준포의 하위 요인이 단일한 차원을 측정하는지에 대한 적합도 분석도 수행되었다. 분석 모형에 대한 자료의 적합도는 내적합(infit) 지수와 외적합(outfit) 지수를 이용하여 검증을 실시하였다(Wright & Masters, 1982). 내적합 및 외적합 지수는 기대치가 1.0인 X^2 의 분포를 이루는데, 지수가 1.0이면 자료가 모형에 적절하다는 것을 뜻한다. 적합도 지수가 1.3보다 크면 일관성이 없는 것으로, 0.7보다 작으면 평가 결과가 거의 변화가 없거나 거의 종속적으로 평가된 것으로 해석한다(Wright, & Linacre, 1990).

IV. 연구 결과 및 논의

1. 국어교사의 엄격성 및 일관성 분석

국어교사의 쓰기 평가 특성 중 우선적으로 검토해야 할 것은 엄격성 수준이다. 국어교사들이 중학생 설명문에 대해 얼마나 엄격하게 채점하는지, 또는 얼마나 관대하게 채점하는지에 따라 평가 결과가 크게 달라질 수 있기 때문이다. FACETS 프로그램 분석을 통해 얻은 logit 분포를 도식화하여 제시하면 <그림 1>과 같다. <그림 1>은 각 국면별 logit 모수 추정치를 도식화한 것이다. 첫째 열의 지수(Measr)는 문항 반응 이론에서 다루는 공통 척도(logit)를 의미하며, 국어교사(평가자)의 엄격성, 학생(피험자)의 능력, 평가 요인의 분포를 판단할 수 있는 근거로 쓰인다. 둘째 열(+Examinees)은 평가 대상이었던 학생 글 6편의 logit 분포를, 셋째 열(-rater)은 평가자인 국어교사 68명의 logit 분포를, 넷째 열(-gender)은 국어교사의 성별 logit 분포를, 다섯째 열(-career)은 국어교사의 경력별 logit 분포를, 여섯째 열(-criteria)은 평가 요인의 logit 분포를, 일곱째 열(Scale)은 평가 요인의 점수 척도별 logit 분포를 도식화한 것이다.¹⁾

〈그림 1〉 학생×국어교사×성별×경력×평가 요인의 분포도

[Measr]+[Examinees]-[rater]	[gender]-[career]	[criteria]	[Scale]
0 +			(5)
2 + 3			4
1 -			---
0 + 1		organization	3
-1 +		craft content word choice convention	---
-2 +			2
-3 +			(1)
[Measr]+[Examinees]-[rater]	[gender]-[career]	[criteria]	[Scale]

〈그림 1〉은 이 연구에서 설정한 5국면, 즉 중학생(피험자), 국어교사(평가자), 국어교사의 성별, 국어교사의 경력, 평가 기준표의 하위 요인을 동일 척도에 제시한 것이다. 따라서 각 국면의 분포 상태를 명시적으로 확인할 수 있다. 〈그림 1〉은 전체 국면의 엄격성 또는 난도에 대한 정보를 보여주는데, 특히 주목해야 할 것은 평가자, 성별, 경력별 국면이다. 평가자 국면에서는 6명의 중학생 설명문을 평가한 국어교사 68명의 엄격성 수준이 어떠한지를, 성별 국면과 경력별 국면에서는 국어교사들이 성과 경력에 따라 엄격성 수준이 어떠한지를 명시적으로 보여주기 때문이다.

국어교사의 평가 특성을 상대적인 크기를 통해 좀 더 구체적으로 파악하기 위하여 수치로 정리하면 <표 4>와 같다. 이 표는 국어교사와 평가 하위 요인이 평가자의 평가 결과 및 평가 결과의 일관성에 미치는 영

- 1) 이는 다음과 같은 의미로 정리할 수 있다. 둘째 열(+Examinees)은 학생의 쓰기 능력, 셋째 열(-rater)은 국어교사의 엄격성, 넷째 열(-gender)은 국어교사의 성별 엄격성, 다섯째 열(-career)은 국어교사의 경력별 엄격성, 여섯째 열(-criteria)은 평가 요인의 엄격성, 즉 난도(難度)를 뜻한다. 둘째 열의 'Examinees'은 다른 열과 달리 '+'로 부호화되어 있다는 점에 유의할 필요가 있다.

항을 조정한 logit 점수, 그 logit 점수의 표준오차, 내적합(infit) 지수, 외적합(outfit) 지수를 제시한 것이며, logit 값의 크기에 따라 내림치순으로 정리한 것이다.²⁾

〈표 4〉 국어교사(평가자 국면)의 평가 엄격성 수준과 적합도

평가자	엄격성 (logit)	SE	Infit MS	Outfit MS	평가자	엄격성 (logit)	SE	Infit MS	Outfit MS
r40	2.48	0.33	1.06	1.02	r21	0.07	0.25	1.91	1.90
r09	2.13	0.3	0.48	0.44	r49	0.07	0.25	1.2	1.20
r12	1.62	0.28	0.45	0.50	r61	0.05	0.27	0.89	0.88
r47	1.57	0.28	1.79	1.96	r66	0.01	0.28	0.33	0.33
r60	1.33	0.26	0.5	0.49	r20	-0.04	0.25	0.82	0.82
r43	1.12	0.27	0.8	0.79	r17	-0.13	0.25	0.74	0.74
r35	1.1	0.26	0.73	0.68	r28	-0.13	0.25	0.68	0.71
r46	0.99	0.26	0.59	0.59	r62	-0.16	0.27	0.49	0.50
r59	0.91	0.26	0.65	0.65	r54	-0.23	0.25	0.71	0.71
r57	0.89	0.26	2.29	2.25	r03	-0.37	0.25	0.39	0.38
r56	0.78	0.26	0.69	0.67	r13	-0.37	0.25	0.68	0.70
r10	0.76	0.26	0.95	0.97	r36	-0.49	0.25	1.05	1.04
r04	0.72	0.26	0.42	0.42	r34	-0.5	0.25	0.85	0.85
r31	0.7	0.26	1.18	1.15	r25	-0.61	0.25	0.65	0.66
r02	0.64	0.25	0.71	0.69	r55	-0.61	0.25	0.7	0.70
r64	0.6	0.28	3.34	3.20	r18	-0.63	0.25	0.54	0.54
r05	0.59	0.25	1.11	1.06	r26	-0.66	0.25	1.02	1.01
r45	0.59	0.25	0.7	0.72	r14	-0.68	0.25	0.49	0.50
r24	0.39	0.25	1.49	1.44	r52	-0.68	0.26	0.85	1.01
r19	0.38	0.25	0.95	0.96	r08	-0.73	0.25	0.87	0.87
r27	0.34	0.25	3.05	3.03	r39	-0.74	0.25	0.66	0.66
r23	0.33	0.25	1.01	1.09	r51	-0.74	0.25	1.35	1.36
r30	0.33	0.25	0.96	0.93	r68	-0.77	0.28	0.75	0.76

2) <표 4>에서 내적합 지수나 외적합 지수가 0.7~1.3의 범위를 벗어나면 평가 결과가 변별력이 없거나 일관성이 없는 것으로 해석한다. 즉, 지수가 0.7 이하(과적합)이면 평가 점수의 변별이 없이, 진위형 문항처럼 1점이나 5점을 부여한 것으로 해석하며, 1.3 이상(부적합)이면 일관성 없이 쓰기 능력이 우수한 학생에게 낮은 점수를 주고 쓰기 능력이 낮은 학생에게 높게 점수는 준 것으로 해석한다.

평가자	엄격성 (logit)	SE	Infit MS	Outfit MS	평가자	엄격성 (logit)	SE	Infit MS	Outfit MS
r37	0.33	0.25	0.96	0.97	r33	-0.87	0.25	0.83	0.85
r42	0.33	0.25	1.71	1.66	r38	-0.89	0.25	0.86	0.85
r50	0.32	0.25	0.51	0.53	r11	-0.92	0.25	0.82	0.82
r1	0.27	0.25	0.55	0.59	r32	-1.08	0.25	0.95	0.94
r41	0.26	0.25	0.28	0.30	r06	-1.11	0.25	0.61	0.60
r48	0.22	0.25	0.94	1.00	r65	-1.12	0.28	5.4	5.72
r67	0.22	0.28	1.09	1.08	r63	-1.3	0.28	0.41	0.41
r29	0.19	0.25	0.91	0.9	r07	-1.77	0.26	1.12	1.07
r58	0.14	0.25	1.16	1.14	r53	-1.78	0.26	0.86	0.84
r15	0.07	0.25	1.32	1.30	r44	-1.82	0.26	1.07	1.04
r16	0.07	0.25	0.52	0.53	r22	-1.96	0.26	1.06	1.29
$\bar{X}$	logit = .00		S.E. = .26		Infit MS=1.01 Outfit MS= 1.01				
s	logit = .89		S.E. = .01		Infit MS=.76 Outfit MS= .78				

<표 4>에 따르면, 국어교사가 보인 평가자로서의 엄격성은 -1.96 logit (SE=.26)부터 2.48 logit(SE=.33)까지의 범위에 분포한다. 이러한 분포는 국어교사 사이에 엄격성이 차이가 있음을 뜻한다. 이러한 차이는 $X^2=745.8$ ($p<.001$), 분리신뢰도 $R=.92$ 로서 통계적으로 유의하였다.

<표 4>를 통해 볼 때, 국어교사 22는 학생 설명문을 가장 관대하게 평가하였으며(-1.96 logit), 국어교사 40은 학생 설명문을 가장 엄격하게 평가하였다(2.48 logit). 모형 적합도 분석에 의하면 국어교사 47, 57, 64, 24, 27, 42, 15, 21, 51, 65(이상 10명, logit 내림차순)는 외적합 지수가 1.3 이상이므로 부적합(misfit)으로 판정할 수 있다. 이 국어교사들은 학생 설명문을 평가할 때 쓰기 능력이 낮은 글에 대해 예상치 않게 높은 점수를 부여하였거나, 혹은 그와 반대로 쓰기 능력이 우수한 글에 대해 낮은 점수를 부여하였다는 뜻이다.

또한, 국어교사 09, 12, 60, 35, 46, 59, 04, 02, 56, 50, 01, 41, 16, 66, 62, 03, 25, 18, 14, 39, 06, 63(이상 22명, logit 내림차순)은 외적합 지수가 0.7 보다 낮은 것으로 나타나 과적합(overfit)으로 분석되었다. 이 국어교사들은

학생 설명문을 변별력 있게 평가하지 못하고, 종속적인 양상으로 평가한 것으로 해석된다. 즉, 어느 정도 잘 썼다고 판단한 글은 모두 5점을 주고, 못 썼다고 판단한 글은 1점을 주었다는 뜻이다. 점수 척도가 1~5점으로 구성되었음에도 불구하고, 진위형 문항을 채점하듯 잘 쓴 것 같으면 5점, 못 쓴 것 같으면 1점을 부여하였다는 의미이다.

이외 국어교사들의 외적합 지수는 모두 0.7과 1.3의 범위 내에 속한다. 학생 글 평가의 엄격성이 일관성 있게 유지되었다고 할 수 있다. 그러므로 국어교사 전체 68명 중 36명(52.94%)의 평가 결과는 엄격성이 일관성 있게 적용된, 안정적인 수준을 보이고 있다고 할 수 있다.

다음 <표 5>는 이 연구의 쓰기 평가에 참여한 국어교사들의 평가 결과로부터 얻은 관찰 점수와, Rasch 모형에 의한 기대 점수 간의 차이를 제시한 것이다.

<표 5> 평가 기댓값 이외의 반응 결과

평가자	성별	경력	피험자	평가 요소	관찰 점수	기대 점수	잔차	표준화 잔차
65	여	1~5	3	단어 선택	2	4.6	-2.6	-4.4
64	여	6~10	2	내용	5	1.9	3.1	4.2
47	여	1~5	4	형식·어법	3	1.2	1.8	4.2
65	여	1~5	3	형식·어법	2	4.5	-2.5	-4.1
65	여	1~5	4	내용	5	2.3	2.7	3.6
21	여	6~10	3	조직	1	3.7	-2.7	-3.5
65	여	1~5	4	단어 선택	5	2.4	2.6	3.5
65	여	1~5	5	단어 선택	1	3.6	-2.6	-3.5
57	여	11~20	3	단어 선택	1	3.6	-2.6	-3.4
65	여	1~5	5	형식·어법	1	3.5	-2.5	-3.4
22	남	6~10	3	조직	3	4.7	-1.7	-3.3
57	여	11~20	3	형식·어법	1	3.5	-2.5	-3.3
40	여	1~5	6	형식·어법	3	1.3	1.7	3.3
42	여	1~5	1	형식·어법	5	2.7	2.3	3.2
27	남	11~20	5	조직	5	2.6	2.4	3.2
19	여	6~10	5	내용	5	2.8	2.2	3.1
44	여	1~5	2	형식·어법	1	3.2	-2.2	-3.0

<표 5>는 국어교사가 평가한 점수, 즉 관찰 점수가 Rasch 모형에 기반을 두고 예측한 기대 점수에서 얼마나 벗어나 있는지를 알려준다. 학생 글이 각 평가 요인에서 얻을 것으로 예측된 점수와, 실제로 국어교사가 평가한 관찰 점수 사이에 존재하는 차이를 분석함으로써, 각 국어교사의 평가 결과에 존재하는 편향성을 파악할 수 있다. <표 5>의 잔차는 관찰 점수와 Rasch 모형에 의한 예측 점수 간의 차이를 의미한다.

국어교사를 기준으로 할 때, 가장 큰 편향성을 보인 평가자는 국어교사 65로 나타났다. 65번 국어교사는 경력이 1~5년인 여교사로, 3번 학생 설명문에 대하여 단어 선택 요인과 형식·어법 요인에서 기대 점수보다 훨씬 낮은 점수를 부여함으로써 매우 높은 수준의 엄격성을 보였다. 그러나 4번 학생 글에 대해서는 내용 요인과 단어 선택 요인에서 기대 점수보다 훨씬 높은 점수를 부여한 것으로 나타났다. 이에 비해 5번 학생 설명문에 대해서는 단어 선택 요인과 형식·어법 요인에서 기대 점수보다 훨씬 낮은 점수를 부여하였다. 이를 종합적으로 고려해 볼 때, 65번 국어교사는 쓰기 평가에서 학생의 설명문 쓰기 능력을 측정하는 데 큰 어려움을 겪는 것으로 보인다. 이는 국어교사 65가 쓰기 평가를 통해 중학생의 설명문 쓰기 능력을 추정하는 데 큰 문제가 있음을 의미한다고 할 수 있다.

경력을 기준으로 할 때, 편향성은 경력이 1~5년인 국어교사들에게서 뚜렷하게 나타났다. 1~5년의 경력을 지닌 국어교사들의 평가 결과가 다른 경력의 국어교사들의 결과보다 관찰 점수와 예측 점수의 차이가 컸다. 성별을 기준으로 하면, 여교사가 남교사에 비해 쓰기 평가의 편향성이 더 큰 것으로 보인다. 그러나 이 연구의 쓰기 평가에 참여한 국어교사의 성별 분포를 볼 때, 여교사가 편향성이 있다고 일반화하기에는 한계가 있다. 여교사가 남교사보다 절대적으로 많았기 때문이다.

학생 글을 기준으로 분석하였을 때, 3번 학생 글이 Rasch 모형에 의한 예측 점수에 비해 관찰 점수가 낮았다. 이는 3번 학생 설명문이 실제 수준보다 모두 낮은 점수를 일관성 있게 받은 것을 뜻한다. 65번 국어교사와 21번 국어교사는 이 학생 글에 대해 적합도가 부적합한 것으로 분석되어 평가 결과를 전적으로 신뢰하기는 어렵지만, 국어교사들이 단어 선택

요인이나 형식 어법 요인, 조직 요인에 대해 일관성 있게 낮은 점수를 부여하였다는 점에 대해서는 주목할 필요가 있다. 이러한 편향성을 보다 정밀히 분석하기 위해서는, 3번 학생 설명문을 평가할 때, 어떠한 원인에 의하여 단어 선택, 형식 및 어법, 조직 요인이 다르게 관찰되는지를 밝혀야 할 것이다.

2. 국어교사의 성 및 경력에 따른 평가 특성 분석

선행 연구에서 지적한 바와 같이, 국어교사의 성(性)과 경력에 따라 쓰기 평가 특성이 어떠한 차이가 있는지를 분석할 필요가 있다. 국어교사의 성, 교사의 경력에 따라 쓰기 평가가 달라진다면, 쓰기 평가의 특성도 차이가 있을 것으로 기대되기 때문이다.

이에 따라 국어교사의 성에 따라 학생 설명문 평가의 엄격성이 차이가 있는지를 분석하였다. 분석 결과는 <표 6>과 같다. 이에 따르면 국어교사의 성에 따른 엄격성은 -0.11 logit(SE=0.03)에서 0.11 logit(SE=0.07)까지의 범위에 분포하였다.

<표 6> 국어교사의 성에 따른 엄격성 수준과 적합도

평가자 성별	엄격성(logit)	SE	Infit MS	Outfit MS
남	-0.11	0.07	1.01	1.04
여	0.11	0.03	0.99	0.99
$\bar{X}$	.00	.05	1.00	1.02
s	.11	.02	.01	.02

<표 6>과 같은 결과는 평가자인 국어교사들이 쓰기 평가에서 보이는 엄격성이 성에 따라 다르다는 것을 보여준다. 여교사는 학생 글을 보다 엄격하게 평가하였으며, 남교사는 상대적으로 관대하게 평가하는 것으로 나타났다. 이러한 엄격성의 차이는 $X^2=7.4(p<.001)$, 분리신뢰도 $R=.86$ 로

통계적으로 유의하였다. 남교사에 비해 여교사가 쓰기 평가에서 더 엄격한 수준을 보인다는 사실은 박영민·최숙기(2009)의 결과와 유사하다. 비록 이 연구가 평가에 참여한 국어교사의 수, 평가한 학생 글의 유형 등이 다르지만, 국어교사의 쓰기 평가는 성별로 엄격성이 다르다는 점을 지지해 준다. 한편, 국어교사의 성에 따른 모형적합도 분석에서 외적합 지수가 0.99~1.04에 분포함으로써, 국어교사의 성에 따라 엄격성이 차이가 있지만, 그 엄격성은 일관성 있게 유지되고 있다고 할 수 있다.

국어교사의 경력에 따른 쓰기 평가의 엄격성에 대한 분석은 <표 7>과 같다.

<표 7> 국어교사의 경력에 따른 엄격성 수준과 적합도

평가자 경력	엄격성(logit)	SE	Infit MS	Outfit MS
1~5년	0.20	0.05	1.04	1.06
6~10년	0.15	0.05	0.90	0.90
11~20년	0.16	0.07	1.10	1.10
20년 초과	-0.52	0.26	0.85	1.01
$\bar{X}$	.00	.11	.97	1.02
s	.30	.09	.10	.08

<표 7>에 따르면, 경력에 따른 엄격성은 $-0.52 \text{ logit}(SE=0.26)$ 에서 $0.20 \text{ logit}(SE=1.1)$ 의 범위에 분포하고 있다. 이러한 분포는 $X^2=7.4(p<.001)$, 분리신뢰도 $R=.84$ 로서 경력별로 통계적으로 유의한 차이가 있는 것으로 분석되었다. 그러므로 국어교사는 경력에 따라 엄격성이 유의한 차이가 있다고 할 수 있다.

경력에 따라 국어교사의 엄격성을 살펴보면, 1~5년 경력의 국어교사가 가장 엄격성이 높은 것으로 나타났고, 11~20년, 6~10년, 20년 초과의 순서로 엄격성이 높았다. 그러나 6~10년의 경력 집단, 11~20년 경력 집단의 엄격성은 유사한 수준을 보였다. 20년 초과의 경력을 지닌 국어교사는 가장 관대하게 평가하고 있는 것으로 나타났다. 국어교사의 경력에 따른 평가의 엄격성은 박영민·최숙기(2009)의 결과와 유사하다. 두 연구의

결과를 종합해 보면, 일반적으로 1~5년의 경력을 지닌 초임 국어교사들이 학생 글을 엄격하게 평가한다고 할 수 있다. 한편, 국어교사의 경력에 따른 모형 적합도 분석에 따르면, 외적합 지수가 0.9~1.10에 분포하여 엄격성이 일관성 있게 유지되고 있음을 보여준다.

3. 평가 요인에 따른 평가 특성 분석

설명문 평가의 하위 요인에 따른 국어교사의 평가 특성을 파악하기 위하여, 설명문 평가 하위 요인 국면의 엄격성을 분석하였다. 평가 요인의 엄격성은 일반적으로 평가 요인의 난도(難度)로 설명된다. 평가 요인이 얼마나 엄격한가는 곧 그 요인에 적합하게 글을 쓰는 것이 얼마나 어렵거나 쉬운가에 대응하기 때문이다. 따라서 이 연구에서는 평가 요인의 엄격성을 난도로 해석하여 논의하고자 한다. 평가 하위 요인 국면에 대한 분석 결과는 <표 8>과 같다. 설명문 평가의 하위 요인은 내용, 조직, 표현, 형식 및 어법, 단어 선택의 5가지로 구성되었으며, 각 요인별 점수 척도는 5점이다.

<표 8> 설명문 평가 요인에 따른 난도 수준과 적합도

평가 요인	난도 (logit)	SE	Infit MS	Outfit MS
조직	0.44	0.07	1.04	1.02
표현	0.03	0.07	0.86	0.86
내용	-0.09	0.07	0.92	0.91
형식·어법	-0.11	0.07	1.15	1.21
단어 선택	-0.27	0.07	1	1
$\bar{X}$	.00	.07	1.00	1.00
s	.24	.00	.10	.12

<표 8>에 따르면, 평가 요인에 따른 난도는 $-0.27 \text{ logit}(SE=0.07)$ 에서

0.44 logit(SE=0.07)의 범위에 분포하고 있다. 평가 요인의 난도는 $X^2=57.2$ ($p<.001$)로서 하위 요인별로 통계적으로 유의한 차이가 있는 것으로 분석되었다. <표 8>에 의하면, 국어교사는 학생 설명문을 평가할 때, 단어 선택 요인에서 가장 높은 점수를 부여하고, 형식 및 어법, 내용, 표현, 조직 요인의 순서로 높은 점수를 부여하였다. 단어 선택 요인은 가장 높은 점수를 얻는, 즉 가장 쉬운 요인으로 해석되며, 조직 요인은 학생 글이 가장 점수를 얻지 못하는, 즉 가장 어려운 요인으로 해석된다. 내용과 형식·어법 요인의 난도는 그리 크지 않은 것으로 보인다. 평가 요인에 따른 모형 적합도 분석에 의하면, 외적합 지수가 0.86부터 1.21의 범위에 속함으로써 5개의 평가 요인은 모두 모형에 적합한 것으로 확인되었다.

이러한 결과는, 15명의 평가자가 264명의 8학년 학생 글에 대한 평가 결과를 분석한 Engelhard(1994)와 차이가 있다. 이 선행 연구에서는, 쓰기 평가의 요인의 난도는 문체(style), 맞춤법(mechanic), 내용 및 조직(content/organization), 표현 용법(usage), 문장 형성(sentence formation)의 순으로 나타났기 때문이다. 세부적인 평가 요인이나 평가 대상이 된 글의 유형이 다르기는 하지만, 이러한 연구 결과를 종합해 볼 때 쓰기 평가 요인 중에서 내용 요인이나 조직 요인이 난도가 높다고 할 수 있다. 학생들의 설명문 쓰기를 지도할 때에는 이 점에 유의할 필요가 있다.

V. 결론

이 연구는 쓰기 평가에서 발견되는 국어교사의 특성을 분석하는 데 목적이 있다. 쓰기 평가를 수행하는 주체로서 국어교사가 학생 글을 평가할 때 어떠한 특성을 보이는가의 문제는 매우 중요함에도 불구하고 이에 대한 연구가 부족했다. 지금까지 이루어지는 쓰기 평가 연구는 평가자 간 신뢰도가 떨어질 때 이를 어떻게 해소할 것인가에만 관심을 두었지, 평가자인 국어교사를 대상으로 하여 어떠한 특성을 보이는지에 대해서는 관심

을 두지 않았다. 이에, 이 연구에서는 학생 글 평가에서 발견되는 국어교사의 평가 특성을, Rasch 모형을 적용하여 엄격성과 일관성에 중점을 두고 분석하였다. 그리고 이러한 엄격성과 일관성이 국어교사의 성(性), 경력에 따라 어떠한 차이가 있는지를 분석하였으며, 쓰기 평가의 하위 요인에 따라 국어교사의 평가 특성이 어떠한지에 대해서도 분석하였다.

이러한 국어교사의 평가 특성 분석을 위하여 이 연구에서는 중학생 설명문 35편을 68명의 국어교사에게 평가하게 하고, 이 중 6편의 중학생 글을 무선 선정하여 Rasch 부분 점수 모형에 따라 분석하였다. 분석은 공개된 FACETS 프로그램을 활용하였다. 분석 결과는 다음과 같다.

첫째, 68명의 국어교사가 보인 엄격성은 $-1.96 \text{ logit}(SE=.26)$ 부터 $2.48 \text{ logit}(SE=.33)$ 까지의 분포를 보였다. 국어교사의 평가자 모형 적합도 분석에 의하면 10명의 국어교사가 부적합으로, 22명의 국어교사가 과적합으로 나타났다. 이를 통해 국어교사들은 평가자로서 엄격성이나 일관성이 서로 다르다는 것을 확인할 수 있었다.

둘째, 국어교사의 성, 경력에 따른 엄격성과 일관성도 차이가 발견되었다. 일관성은 성별에 따른 차이가 없었으나 엄격성은 차이가 있었다. 여교사는 남교사에 비해 더 엄격하게 평가하는 것으로 나타났다. 경력에 따른 엄격성과 일관성의 차이도 유의하게 나타났다. 엄격성은 1~5년의 국어교사가 가장 높았다.

셋째, 설명문 쓰기 평가의 하위 요인 분석에서, 국어교사들은 학생 설명문을 평가할 때, 단어 선택 요인에서 가장 높은 점수를 부여하고, 형식 및 어법, 내용, 표현, 조직 요인의 순으로 높은 점수를 부여하는 것으로 나타났다. 조직 요인은 학생 글에서 평가 점수가 가장 낮은 요인에 해당하므로, 학생들이 설명문을 썼을 때 조직 요인에서 가장 점수를 얻지 못한다고 할 수 있다.

이러한 결과를 종합해 보면, 국어교사의 쓰기 평가 특성에 해당하는 엄격성과 일관성은 국어교사 개인별, 성별, 경력별로 차이가 있는 것으로 보인다. 평가자 요인으로서 국어교사의 엄격성의 차이는 평가자간 신뢰도에 영향을 미친다는 점에 주목할 필요가 있다. 엄격성이 차이가 있다는

것은 곧 동일한 학생 글에 대해 국어교사의 평가 점수가 차이가 있음을 의미하기 때문이다. 이러한 차이는 학생의 설명문 쓰기 능력을 추정하는데에도 부정적인 영향을 미칠 수 있다.

그러므로 쓰기 평가에 대한 국어교사의 전문성을 높이기 위해서는 학생 글에 대한 엄격성과 일관성이 적절하게 유지될 수 있도록 하는 장치가 마련되어야 한다. 이러한 장치를 마련하기 위해서는 국어교사의 엄격성과 일관성에서 문제가 발생하는 원인이 무엇인지를 밝히는 연구가 더 심층적으로 이루어져야 할 것이다.

또한, 국어교사의 쓰기 평가 특성을 분석한 결과를 볼 때, 중학생 설명문 평가에 참여한 국어교사 68명 중 36명(52.94%)만이 적합한 수준을 보였다는 점도 주목할 필요가 있다. 평가 참여자의 47%는 적합한 수준에서 쓰기 평가를 수행하지 못한 것으로 해석할 수 있기 때문이다. 다소 제한적이기는 하지만, 이러한 결과는 국어교사의 쓰기 평가 전문성을 높일 수 있는 프로그램의 구성과 운영이 필요함을 일깨워준다. 후속적인 연구에서는 이에 대한 방안이 더 논의되고 탐구되어야 할 것이다.*

* 본 논문은 2010. 2. 28. 투고되었으며, 2010. 3. 12. 심사가 시작되어 2010. 3. 24. 심사가 종료되었음.

■ 참고문헌

- 가은아(2008), “국어교사의 쓰기 효능감 및 쓰기 신념이 학생 쓰기 평가에 미치는 영향”, 석사학위 논문, 한국교원대학교.
- 김성숙(1995), “논술 문항 채점의 변동 요인 분석과 일반화 가능성도 계수의 최적화 조건”, 『교육평가연구』 8(1), pp.35~57.
- 김성숙(2001), “채점의 변동 요인 분석 방법에 대한 고찰: 일반화 가능성도 이론과 다국면 라쉬 모형의 적용과 재해석”, 『교육평가연구』 14(1), pp.303~325.
- 민태범(1991), “채점자의 인지 양식이 논술 고사 채점에 미치는 영향”, 석사학위논문, 한남대학교.
- 박영목(1999), “작문 능력 평가 방법과 절차”, 『국어교육』 99, pp.1~29.
- 박영목(2008), “쓰기 평가 연구의 주요 과제”, 『작문연구』 6, pp.9~37.
- 박영민(2009a), “논술 평가자의 주관적 경험 분석”, 『우리어문연구』 33, pp.493~523.
- 박영민(2009b), “평가 예시문을 활용한 쓰기 평가 개선 방안”, 『청람어문교육』 39, pp.111~133.
- 박영민 · 최숙기(2009), “현직 국어교사와 예비 국어교사의 쓰기 평가 비교 연구”, 『교육과정평가연구』 12(1), pp.123~143.
- 박영민 · 최숙기(2010a), “국어교사의 설명문 평가에 대한 모평균 추정과 평가 예시문 선정”, 『우리어문연구』 36, pp.293~326.
- 박영민 · 최숙기(2010b), “중학생 논설문 모평균 추정과 평가 예시문 선정”, 『국어교육』 131, pp.437~461.
- 이성영(2005), “국어과 교사의 쓰기 영역 평가 전문성 기준과 모형”, 『국어교육』 117, pp.363~376.
- 이현숙(2008), “다국면 라쉬 모형을 적용한 논술 채점 상황에서 채점 설계 및 채점자 특성이 채점의 정확성에 미치는 효과”, 『교육평가연구』 21(4), pp.129~152.
- 지은림 외(2000), 『RASCH 모형의 이론과 실제』, 교육과학사.
- Barnes, L. L.(1990), Gender bias in teachers'written comments. In S. L. Gabriel & L. Smithson(eds.), *Gender in the classroom : Power and pedagogy*, Chicago, IL : University of Illinois Press, 140-159.
- Engelhard, G. Jr.(1994), Examining rater errors in the assessment of written composition with a many-faceted Rasch model, *Journal of Educational Measurement*, 31(2), 93-112.

- Etaugh, C. B., Houtler, B., & Prasnik, P.(1988), Evaluating competence of women and men : Effects of experimenter gender and group gender composition, *Psychology of Women Quarterly*, 12, 191-200.
- Haswell, J. & Haswell, R. H.(1995), Gendership and the miswriting of students, *College Composition & Communication*, 46(2), 223-254.
- Haswell, R. H. & Haswell, J. T.(1996), Gender bias and critique of student writing, *Assessing Writing*, 3, 31-83.
- Huot, B.(1990), The literature of direct writing assessment : Major concerns and prevailing trends, *Review of Educational Research*, 60(2), 237-263.
- Huot, B.(1996), Toward a new theory of writing assessment, *College Composition & Communication*, 47(4), 549-566.
- Linacre, J. M.(1989/1993), *Many-facet Rasch measurement*, Chicago, IL : MESA Press.
- Penny, J., Johnson, R. L., & Gordon, B.(2000), The effect of rating augmentation on inter-rater reliability : An empirical study of a holistic rubric, *Assessing Writing*, 7, 143-164.
- Peterson, S.(1998), Evaluation and teachers' perceptions of gender in sixth-grade student writing, *Research in the Teaching of English*, 33(2), 181-208.
- Peterson, S. S. & Kennedy. K.(2006), Sixth-grade teachers' written comments on student writing: genre and gender influences, *Written Communication*, 23(1), 36-62.
- Rich, J.(1975), Effects of children's physical attractiveness on teachers' evaluations, *Journal of Educational Psychology*, 67, 599-609.
- Roulis, E.(1995), Gendered voice in composing, gendered voice in evaluating : Gender and the assessment of writing quality, In D. L. Rubin (eds.), *Composing social identity in written language*, Hillsdale, NJ : Lawrence Erlbaum Associates, 151-183.
- Runder, L. M.(1992), Reducing errors due to the use of judges, *Practical Assessment, Research & Evaluation*, 3(3).
- Spandel, V. & Culham, R.(1996), Writing Assessment, In R. E. Blum and J. A. Alter(eds.), *A Handbook for Student Performance Assessment in an Era of Restructuring*, ASCD.
- White, M. J & Brunning, R.(2005), Implicit writing beliefs and their relation to writing quality, *Contemporary Educational Psychology*, 30(2), 166-189.
- Wright, B. D. & Linacre, J. M.(1990), Measuring the impact of judge severity on examination scores, *Applied Measurement in Education*, 3, 331-345.

<초록>

Rasch 모형을 활용한 국어교사의 쓰기 평가 특성 분석

박영민 · 최숙기

이 연구는 쓰기 평가에서 발견되는 국어교사의 특성을 분석하는 데 목적이 있다. 국어교사의 평가 특성 분석을 위하여 이 연구에서는 중학생 설명문 35편을 68 명의 국어교사에게 평가하게 하고, 이 중 6편의 중학생 글을 무선 선정하여 Rasch 부분 점수 모형에 따라 분석하였다. 분석은 공개된 FACETS 프로그램을 활용하였다. 분석 결과는 다음과 같다. 첫째, 68명의 국어교사가 보인 평가자로서의 엄격성은 $-1.96 \text{ logit}(SE=.26)$ 부터 $2.48 \text{ logit}(SE=.33)$ 까지의 분포를 보였고, 평가에 참여한 국어교사들은 평가자로서의 엄격성이나 일관성이 서로 차이가 있었다. 둘째, 국어교사의 성(性), 경력에 따라 쓰기 평가의 엄격성과 일관성도 차이가 있었다. 일관성은 성별에 따른 차이가 없었으나 엄격성은 차이가 있었다. 여교사는 남교사에 비해 학생 설명문을 더 엄격하게 평가하는 것으로 나타났다. 경력에 따른 엄격성과 일관성도 유의한 차이가 발견되었다. 1~5년의 국어교사가 가장 엄격하게 학생 글을 평가하였다. 셋째, 쓰기 평가의 하위 영역 분석에서, 국어교사들은 학생 설명문을 평가할 때, 단어 선택 요인에서 가장 높은 점수를 부여하고, 조직 요인에서 가장 낮은 점수를 부여하는 것으로 나타났다. 이러한 국어교사의 쓰기 평가 특성은 국어교사의 쓰기 평가 전문성 신장 프로그램을 마련하는 데 기여할 수 있을 것으로 생각된다.

【핵심어】 국어교사, 쓰기 평가, 설명문 평가, Rasch 모형, FACETS 프로그램

<Abstract>

The Feature Analysis of Korean Language Teacher's Writing Assessment based on Rasch Model

Park, Young-min · Choi, Sook-ki

The aims of this paper is to analyze the features of Korean language teacher's writing assessment based on Many-Facets Rasch model. For that, the 68 Korean language teachers assessed 35 expository writing of middle school students. Then, 6 expository writing was selected randomly and analyzed using free FACETS program. The results are following. First, the range of Korean language teachers' severity was from -1.96 logit(SE=.26) to 2.48 logit(SE=.33). It was showed that severity and consistence of Korean language teachers engaged writing assessment were different. Second, there were differences of Korean language teachers' severity and consistence according to their gender and career. The female teachers(0.11 logit, SE=0.07) were more severer than male teachers(-0.11 logit, SE=0.03). The range of severity as teachers' career was from -0.52 logit(SE=0.26) to 0.16 logit(SE=1.1). The severity of teacher experienced 1~5 years was more higher than another teacher. Third, as analyzed subordinate constructs of expository writing assessment, Korean language teachers assessed word choice, convention, content, craft, organization highly in order of difficulty. This result will be a helper to develop programs for expertises in students' writing assessment of Korean language teachers.

[Key words] Korean language teachers, writing assessment, expository writing, Many-Facets Rasch model, FACETS program