교육과정평가연구
The Journal of Curriculum and Evaluation
2011, Vol. 14, No. 1, pp. 201～228

논설문 평가에 나타난 국어교사의 평가 특성 및 편향 분석

최 숙 기(대구대학교 특임교수)*
박 영 민(한국교원대학교 부교수)**

≪ 요 약 ≫

이 연구는 국어교사의 평가 결과를 MFRM(Multi-Faceted Rasch Measurement)로 분석함으로써 국어교사의 평가자로서의 특성과 편향을 밝히는 데 목적이 있다. 이를 위해 중학생이 쓴 논설문 35편을 국어교사 68명에게 채점하게 하고, 여기에서 얻은 결과를 적용하여 분석하였다. 국어교사는 별도의 평가자 협의를 거치지 않고 개별적으로 제시된 분석적 쓰기 평가 기준에 근거하여 5점 척도(1점에서 5점)로 학생 논설문을 평가하였다. 평가 요소는 내용, 조직, 표현, 형식 및 어법, 단어 선택이었다. 먼저 중학생 논설문 평가에 참여한 국어교사의 평가 특성을 참여한 국어교사들의 평가 엄격성과 일관성을 평가자의 성별, 경력별 특성 분석을 실시하였고, 평가자와 쓰기 평가 요인 간, 국어교사와 학생 간의 편향 유형을 살펴보기 위해 상호작용 분석을 실시하였다. 결과에 따르면, 68명의 국어교사 중 14명(20.58%)은 과적합 평가자, 10명(14.70%)은 부적합 평가자로 나타났다. 엄격성은 여교사가 남교사에 비해 높았고, 1～5년의 경력을 지닌 집단의 엄격성이 가장 높았다. 평가 요인에서 국어교사들은 단어 선택 요인에서 가장 높은 점수를 부여하였으며, 다음은 표현, 조직, 형식 및 어법, 내용의 순서로 나타났다. 국어교사 68명 중 48명의 평가에서 학생 글에 대한 편향이 유의미한 것으로 분석되었고, 최소 편향은 1개, 최대 편향은 17개로 나타났다. 국어교사와 평가 요인 간에 편향은 42개로 나타났으며, 내용 요인, 형식 및 어법 요인과 관련한 평가자 편향이 두드러졌다. 이 연구는 학생 쓰기 평가에 나타난 국어교사들의 평가 특성과 편향에 관한 정보를 제공함으로써 국어교사의 쓰기 평가와 관련한 교육이나 연구에 기여할 수 있을 것이다.

주제어 : 국어교사, 쓰기 평가, 논설문, 다국면 라쉬 모형, FACETS 프로그램, 평가자 편향

* 제1저자
** 교신저자, enrapture@knue.ac.kr

Ⅰ. 서론

이 연구는 중학생 논설문을 평가한 국어교사의 평가 특성을 분석하고, 이들이 평가자로서 어떤 편향을 보이는지를 살펴보는 데 목적이 있다. 평가자로서 국어교사는 학생 글을 평가하는 과정에서 여러 가지 요인으로부터 영향을 받는다. 채점이 이루어지는 상황, 학생 글이 반영하고 있는 특성으로부터 영향을 받을 뿐만 아니라, 심지어는 국어교사 자기 자신의 심리적 조건으로부터도 영향을 받는다.

평가에 영향을 미치는 요인들은 학생 글을 평가하는 국어교사의 엄격성이나 일관성에 변화를 초래할 수 있다. 엄격성과 일관성의 변화는 쓰기 평가에서 유지되어야 할 평가의 객관성과 신뢰성에 밀접하게 관련을 맺고 있기 때문에 주목할 필요가 있다. 그러므로 이 연구에서는 중학생 논설문 평가에서 발견되는 국어교사의 평가 특성과, 학생 변인 및 평가 범주 변인과 관련한 국어교사의 편향을 분석하고자 한다.

평가자로서 국어교사는 엄격성도 다르고 그 엄격성을 유지하는 일관성도 다르다. 어떤 국어교사는 학생 글을 매우 엄격하게 평가하는가 하면, 어떤 국어교사는 매우 관대하게 평가하기도 한다. 국어교사는 이러한 엄격성을 유지하는 일관성에서도 차이가 있다. 처음부터 끝까지 엄격한 정도를 일관성 있게 유지하면서 평가하는 국어교사가 있는가 하면, 엄격성을 일관성 있게 유지하지 못한 채 학생 글을 채점하는 국어교사도 있다. 엄격하면 학생 글을 채점한 점수가 낮고 관대하면 점수가 높은데, 이러한 엄격성의 차이는 평가자로서 기능하는 국어교사의 특성을 형성한다.

이러한 엄격성과 일관성은 중요한 의미가 있음에도 불구하고, 현재 국어교육에서 수행된 쓰기 평가에 관한 연구에서는 이와 같은 분석이 잘 다루어지지 않았다. 그래서 과학적 분석 결과에 토대를 둔 쓰기 평가의 방법 제안이나 문제의 개선이 쉽지 않았다. 국어교사들이 일반적으로 보이는 엄격성이나 일관성은 어떠한 수준인지, 국어교사의 성(性)이나 경력에 따라 엄격성과 일관성의 수준이 다른지, 평가 요소에 따른 국어교사의 엄격성이나 일관성은 차이가 있는지 등에 대한 정보가 축적되어 있지 않으므로, 쓰기 평가에서 발견되는 문제를 해결하는 데에도 어려움이 있다.

학교 쓰기 평가 장면에서는 국어교사 한 명이 학생 글 평가를 맡는 경우가 대부분이다. 즉, 국어교사 한 명이 학생 글을 평가하고, 그 결과를 그대로 반영하여 성적을 처리하는 것이다. 이런 경우에는 국어교사의 엄격성과 일관성이 매우 중요하다. 국어교사 한 명이 채점하므로 엄격성은 학생 글의 수준을 결정하는 데에 직접적으로 영향을 미치며, 일관성은 그 평가 결과를 신뢰할 수 있는지를 판단하는 데에 직접적으로 영향을 미치기 때문이다. 의욕적으로 세 명의 국어교사가 평가하여 평균을 반영한다고 하더라도 사정은 비슷하다. 세 명

중 한 국어교사가 특히 평가자의 내적 일관성이 낮다면 학생의 쓰기 능력을 잘못 추정한 점수를 반영할 가능성이 크기 때문이다. 이러한 문제를 효과적으로 개선하기 위해서는 이 연구와 같은 접근 방법이 요구된다.

이 연구에서는 학생 논설문 평가에서 발견되는 국어교사의 특성을 탐색하되, 엄격성과 일관성, 평가자의 편향에 중점을 두고 분석을 진행하고자 한다. 이를 위하여 이 연구에서는 국어교사 68명을 평가자로 위촉한 후, 임의로 수집한 중학생 논설문 35편을 채점하게 하여 채점 자료를 수집하였다. 수집된 채점 자료는 FACETS 프로그램을 활용하여 분석하였으며, 국면은 다섯 가지로 설정하였다.

이 연구는 68명의 국어교사를 대상으로 하여 자료를 수집·분석하였으므로 연구 결과를 일반화하는 데에는 다소 한계가 있다. 그러나 국어교사의 쓰기 평가 자료를 실증적으로 활용하였다는 점, 국어교사가 지닌 평가자 특성을 변인별로 탐색했다는 점, 국어교사와 학생 글, 국어교사와 평가 요인 사이에 나타한 평가자 편향을 분석하였다는 점에서 이 연구는 의의가 있다고 할 수 있다.

Ⅱ. 쓰기 평가에 영향을 미치는 평가자 특성

1. 평가자의 엄격성과 관대함

수행평가에서 발생하는 오류에 대한 논의는 여러 선행 연구에서 다루어져 왔다. 이 연구들에 따르면, 수행평가와 관련한 주목할 만한 오류는 대체로 평가자와 관련된 것으로 밝혀진 바 있다(Cantor & Hoover, 1986; Engelhard, 1992, 1994; Engelhard, Gordon & Gabrielson, 1991; Gabrielson, Gordon & Engelhard, 1995; McNamara, 1996; Ruth & Murphy, 1988; 김성숙, 1995, 2001). 특히, 언어 수행평가에서 평가자의 주 효과가 유의한지를 분석한 연구에 따르면, 평가자의 엄격성(severity) 혹은 관대함(leniency)은 평가 결과의 차이를 크게 만드는 요인이라는 점이 확인되었다(Engelhard, 1994; Engelhard & Myford, 2003; Lumley & McNamara, 1995).

평가자의 엄격성은 평가자가 얼마나 엄격하게 점수를 부여하는가, 즉 점수를 얼마나 박하게 매기는가를 뜻한다. 평가자가 엄격하면, 척도의 중간 이하로 점수를 부여하는 경향이 나타난다. 이에 반해, 평가자의 관대함은 평가자가 얼마나 점수를 후하게 매기는가를 뜻한다. 그러므로 평가자가 관대하면, 척도의 중간 이상으로 점수를 부여하는 경향이 나타난다. 평가자가 보이는 엄격성과 관대함은 대립적인 개념이므로, 일반적으로는 엄격성을 대표 개념으

로 하여 엄격성이 높고 낮음으로 표현한다.

평가자의 엄격성이 높으면, 평가자는 피험자의 능력을 과소 추정하는 경향이 나타난다. 이에 비해, 평가자의 엄격성이 낮으면, 즉 관대하면, 평가자는 피험자의 능력을 과대 추정하는 경향을 강하게 보인다. 평가자들이 가지고 있는 이 엄격성 차이는 평가 결과의 차이를 만들어 내는 주된 원인으로 작용한다. 동일한 피험자를 평가하더라도 평가자의 엄격성 수준이 어떠한가에 따라 서로 다른 평가 점수를 부여하기 때문이다. 동일한 피험자임에도 불구하고 평가자마다 서로 다른 평가 결과를 제출한다면, 이 평가 결과를 통해 추정되는 피험자의 능력이 타당하고 신뢰롭다고 보기는 어려울 것이다. 이러한 맥락에서 Cronbach(1990)는 평가자와 관련하여 발생할 수 있는 가장 민감하고 심각한 오류를 평가자의 엄격성 효과(severity effect)로 설명한 바 있다.

평가자들이 가지고 있는 기준이나 척도가 차이가 있기 때문에 평가자들의 엄격성도 차이가 발생할 수밖에 없다. 그러나 평가자 엄격성은 객관적인 평가 결과를 얻기 위해서는 조정될 필요가 있다. 평가자 훈련이나 평가자 협의는 평가자 엄격성을 조정하는 데 효과적인 것으로 알려져 있다. 하지만 평가자의 엄격성 효과를 완전히 해소하는 것이 쉬운 일은 아니다. 특정한 평가자 훈련을 거치더라도 여전히 엄격성의 차이는 뚜렷하다는 연구 결과가 보고된 바 있기 때문이다(Barrett, 2001; Lumley & McNamara, 1995; Weigle, 1998).

2. 평가자의 일관성

쓰기와 같은 수행 과제에 대한 평가는 채점자들의 판단에 의하여 점수가 부여되기 때문에 평가자 신뢰도는 측정의 일관성을 설명하는 주요한 지표로 쓰인다. 평가자 신뢰도 중에서 평가자 내 신뢰도는 일반적으로 동일 평가자가 다수의 대상을 평가할 때 나타나는 평가자의 일관성(consistency)을 의미한다(성태제, 2002). 평가자의 일관성은 평가자가 점수를 부여하는 경향이 일정하게 지속되는 현상을 말한다. 평가자는 일관성을 유지할 수 있어야 신뢰성 있게 평가할 수 있는데, 평가자가 일관성을 유지하기 위해서는 평가자 자기 자신이 내면화한 평가 기준이나 척도가 변하지 않도록 통제하고 노력해야 한다. 그래서 많은 연구자들은 평가자 훈련에서 가장 중요한 목표는 평가자 일관성 확보라고 강조하기도 했다(Cason & Cason, 1984; Lunz & Stahl, 1993). Winsemen(1949)도 쓰기 평가와 같은 수행평가에서는 평가자의 일관성이 평가 국면에서 매우 중요한 요인이라고 강조한 바 있다.

평가자가 일관성 있게 평가하였는지를 알아보는 데에는 다국면 Rasch 분석을 적용하는 것이 효과적이다. 이 분석은 일반적으로 평가자의 신뢰도로 이해되어 온 평가자 간 신뢰도뿐만 아니라, 평가자 내 신뢰도를 설명하는 평가 일관성에 대한 정보를 제공해 주기 때문이다. 일반적으로 Rasch 분석 결과, 내적합 지수가 1.3보다 크면 부적합으로, 0.7보다 작으면

과적합으로 판정하여 평가 일관성을 판정한다(Bond & Fox, 2001; McNamara, 1996). 부적합은 평가자의 평가 점수가 일관성이 없는 것으로 해석하고, 과적합은 평가 결과가 특정한 점수에 종속된, 그래서 특정 점수를 일률적으로 부여하는 것으로 해석한다(Wright, & Linacre, 1990; McNamara, 1996; Bond & Fox, 2001).

Lunz & Stahl(1990)은 다국면 Rasch 분석을 통해 일정한 채점 기간 동안 평가자의 일관성이 유지되는지를 조사한 바 있다. 연구에 참여한 3명의 평가자는 1일에서 4일 동안 실시된 쓰기 평가에서 평가의 일관성을 잘 유지하지 못하였다. 이와 같은 결과는 Myford(1991)의 연구에서도 동일하게 나타났다. Myford(1991)는 전문성에 차이가 있는 평가 집단을 대상으로 하여 한 달 간 연극을 평가하게 하였는데, 평가자들이 평가의 일관성을 유지하지 못하는 문제를 발견하였다. Lumley & McNamara(1995)은 말하기 평가에서 20달 동안의 평가 기간 중 평가자 일관성이 어떻게 변화하는지를 분석하였는데, 특히 평가자와 평가 기간(평가 시간)의 상호작용 효과를 분석하여 평가자의 일관성의 변화가 평가에 미치는 영향을 구체적으로 분석하였다.

3. 평가자의 편향성

평가자 편향(bias)은 평가자가 평가 상황에서 특정 피험자, 특정 요소 등에 대해 더 엄격하거나 더 관대하게 평가하는 특성을 말한다. 특정 대상이나 특정 요인 등에 치우친 평가를 한다고 하여 평가자 편향이라고 한다. 평가자의 편향성은 다양한 국면에서 발생할 수 있는데, 이러한 평가자 편향은 평가 결과에 큰 영향을 미친다. 가령, 평가자는 깨끗한 글씨로 쓴 글에 편향적인 평가를 할 수도 있고, 특정한 학생, 예를 들면 공부를 잘 하는 학생이 작성한 글에 편향적인 평가를 할 수도 있다. 그런데 이러한 편향이 발생하면, 쓰기 평가의 전체적인 신뢰성이 떨어질 뿐만 아니라, 평가 결과에 바탕을 둔 능력 추정이 왜곡될 수도 있다. 대부분의 평가자들은 주관적 판단에 의하여 평가를 실시하기 때문에, 평가자 편향은 객관적인 평가를 수행하는 데 장애가 되기도 한다.

평가자의 편향성은 편향이 발생하는 특정 국면과 평가자 사이의 상호작용이 유의한지를 분석함으로써 확인할 수 있다. 평가자 편향은 다국면 Rasch 분석의 상호작용 분석을 통해서 잘 드러나는데, 특히 이 분석에서는 상호작용 분석, 또는 편향 분석을 통해 평가자와 평가 요인 간의 상호작용을 확인함으로서 편향 실태를 확인할 수 있다. 다국면 Rasch 분석에서는 평가자와 특정 평가 요인, 혹은 특정 평가 대상과의 상호작용 결과를 제시하여 주므로, 이를 통해서 평가자의 편향성을 파악할 수 있다(이현숙, 2008). 예를 들어, 평가자가 학생들의 성별 요인과 유의한 상호작용이 있는지, 평가 요인 중 내용 요인과 유의한 상호작용이 있는지를 분석하여 편향성이 있는지를 파악할 수 있다.

평가자 편향 분석 연구에서 주목을 끌었던 것 중의 하나는 학생 성(性)에 대한 평가자의 편향성이다. 평가자들에게 남학생이 쓴 글과 여학생이 쓴 글을 평가하게 한 후, 학생 성에 대해 편향이 있는지를 분석한 결과에 따르면, 평가자와 학생의 성 사이에는 유의미한 편향(gender bias)이 있음이 밝혀진 바 있다(Engelhard, 2002; Engelhard & Myford, 2003; Myford & Wolfe, 2003). 학생 성에 대한 평가자의 편향 외에도, 평가자의 성이 편향을 발생시키는 경우도 있다. 학생들이 작성한 서사문과 논설문을 평가할 때 여교사는 주로 글의 규칙과 조직에 대해 중점을 두는 데 비해, 남교사는 미적인 문체에 초점을 두는 경향이 강하다(Peterson & Kennedy, 2006). 학생 글을 평가할 때, 여교사는 학생 글에 대한 논평의 양도 남교사에 비해 유의하게 더 많다(Barnes, 1990; Roulis, 1995).

학생 글을 평가할 때 평가자, 즉 국어교사가 글을 쓴 학생의 성별을 어떻게 인식하는가도 편향을 발생시키는 것으로 알려져 있다. 국어교사가 글을 작성한 학생을 남학생으로 인식하는가, 아니면 여학생으로 인식하는가가 쓰기 평가에 영향을 미치는 것이다. Peterson(1998)에 따르면, 국어교사들은 남학생이 쓴 것으로 인식한 글에 대해 수정 사항을 더 많이 지적하였고 더 많은 비평을 가한다는 점이 확인되었다. 이러한 경향은 국어교사가 학생 글을 평가할 때 동성(同性)인 학생의 글을 더 부정적으로 평가하는 동성 평가 절하 현상(same-sex depreciation)으로 이어지기도 한다(Peterson, 1998). 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 보인다는 것이다(Etaugh, Houtler, & Ptasnik, 1988; Haswell & Haswell, 1995, 1996).

Ⅲ. 연구 과정 및 연구 방법

1. 연구 대상

학생 논설문 평가에서 드러나는 국어교사의 쓰기 평가 특성, 즉 엄격성과 일관성을 분석하기 위하여 국어교사 68명을 평가자로 위촉하여 평가 결과를 수집하였다. 학생 논설문을 평가한 국어교사는 전국적인 수준에서 선정하였다. 학생 논설문 평가에서 발견되는 국어교사의 엄격성과 일관성은 국어교사의 성(性)과 경력에 따라 분석함으로써 국어교사의 평가 특성을 구체적인 수준에서 확인하고자 하였다.

평가자로 위촉된 국어교사는 남교사 12명(17.39%), 여교사 56명(82.35%)이었다. 경력은 1~5년, 6~10년, 11~20년, 20년 초과로 구분하였다. 이를 정리하면 <표 1>과 같다.

〈표 1〉 평가 참여 국어교사의 규모

성별 / 경력	남	여	계
1~5년	3	25	28
6~10년	3	23	26
11~20년	5	8	13
20년 초과	1	0	1
계	12	56	68

2. 검사 도구

이 연구에서 평가 자료를 수집할 때 적용한 중학생 논설문은 '학생들의 복장 및 두발규제에 대한 자신의 입장을 서술하라.'는 과제에 따라 작성되었으며, 학생들이 작성한 글 중에서 35편을 임의로 수집하여 평가하였다. 68명의 국어교사는 중학생 논설문 35편을 모두 평가하였고, 이들의 평가 결과를 분석 자료로 활용하였다.

학생들이 작성한 논설문은 글씨 모양이나 학생의 성(性) 등 주관적인 요인의 영향을 줄이기 위하여 워드 프로세서로 입력한 후 교사들이 평가하였다. 입력 과정에서 의미의 오해를 초래할 수 있는 띄어쓰기는 수정하였으나 맞춤법 오류는 수정하지 않았다. 맞춤법 오류가 평가 기준에 포함되어 있기 때문이다.

국어교사들에게는 평가 기준표와 채점표를 함께 제공하였다. <표 2>가 제공된 평가 기준표인데, 이는 Spandel & Culham(1996)이 제시한 평가 기준을 중학생의 논설문 평가에 적합하도록 수정한 것이다.[1] 쓰기 평가 요인은 내용, 조직, 표현, 단어 선택, 형식 및 어법 등 다섯 개다. 평가 기준표에는 1점, 3점, 5점에 해당하는 평가 기준을 제시하였으나, 1점부터 5점 사이에서 자유롭게 채점하도록 안내하였다.

1) Spandel & Culham(1996)은 '아이디어와 내용'(idea and content), '조직'(organization), '어조'(voice), '단어 선택'(word choice), '문장 유창성'(sentence fluency), '쓰기 관습'(convention)을 평가 요소로 제시하였는데, 이는 학교에서 학생들이 작성하는 대부분의 글에 적용할 수 있도록 개발된 것이다. 그러나 이 평가 요소는 영어 표현을 기반으로 한 것이어서 국어 표현에는 적합하지 않은 것들이 포함되어 있다. 그러므로 이 연구에서는 Spendel & Culham(1996)의 평가 요소 중에서 적합성이 떨어지는 '문장 유창성'을 삭제하고, '목소리'를 '표현'으로 수정하여 '내용', '조직', '표현', '단어 선택', '형식 및 어법'을 평가 요소로 구성하였다.

〈표 2〉 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(주제)과 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: 중심 내용이 잘 드러나도록 조직되었다. 내용의 순서나 구조가 독자가 이해하기 쉽도록 되어 있다. •3점: 중심 내용이 잘 드러나도록 조직되었지만, 내용의 순서나 구조가 독자가 이해하는 데 어려움이 따른다. •1점: 중심 내용이 잘 드러나도록 조직되지 않았을 뿐 더러, 내용의 순서나 구조가 독자가 이해하는 데 어려움이 따른다.
표현 (어조 및 태도)	•5점: 독창적이며 흥미롭게 표현되어 있으며, 독자가 쉽고 정확하게 이해할 수 있도록 표현되었다. 필자의 주체적인 목소리도 드러난다. •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 단어들이 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 논설문 평가 자료를 수집하기 위하여, 현직 국어교사 68명에게 중학생들이 작성한 35편의 논설문과 논설문 과제, 평가 기준표, 채점표, 개인 기록표를 제공하였다. 평가자로 참여한 국어교사들은 논설문 과제와 평가 기준표에 대한 설명을 들은 후, 연구자와 질의응답을 하는 가운데 과제 및 평가 기준을 내면화하였다. 또한, 평가 기준에 대한 공통된 이해를 위하여 과제와 평가 기준에 대한 논의의 시간을 가졌다. 그런 다음, 국어교사들은 평가자 훈련을 거치지 않고 개별적으로 학생 논설문을 평가하였다. 평가자 훈련을 실시하지 않은 이유는 이 연구의 목적이 평가자인 국어교사의 평가 특성과 편향을 분석하는 데 있기 때문이다. 평가자 훈련을 할 경우, 국어교사 개개인이 가지고 있는 평가 특성이나 편향성이 소거될 가능성이 높아진다. 국어교사의 평가 자료는 2009년 1월부터 3월까지 수집하였다. 국어교사들은 평가 기준에 따라 5점 척도로 중학생 논설문을 각각 채점하였다.

4. 분석 방법

논설문 평가에서 발견되는 국어교사의 쓰기 평가 특성과 평가자로서의 편향을 파악하기 위하여 이 연구에서는 평가자로서의 국어교사, 국어교사의 성별, 경력, 논설문 평가 요인, 평가 대상으로서의 중학생이라는 5국면을 Rasch 모형에 적용하여 분석하였다. 이를 통해, 중학생 논설문 평가에서 성(性)이 p이고 경력이 q인 국어교사(평가자) j가 중학생(피험자) n의 논설문 쓰기 평가 요인 i에 대해 평가 점수가 k-1이 아닌 k를 부여할 확률과, 그 확률을 log로 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 기준도 문항처럼 간주하여 분석하므로, 이 연구에서도 <수식 1>의 평가 요인을 각각의 문항처럼 처리하였다.

〈수식 1〉 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 p이고 경력이 q인 국어교사(평가자) j가 평가 요인 i에 대해 점수 k를 줄 확률
$P_{nijpq(k-1)}$ = 성(性)이 p이고 경력이 q인 국어교사(평가자) j가 평가 요인 i에 점수 k-1을 줄 확률
B_n = 중학생(피험자) n의 논설문 쓰기 능력
D_i = 논설문 평가 요인 i의 난도(難度)
C_j = 국어교사(평가자) j의 엄격성
U_p = 성(性)이 p인 국어교사(평가자) j의 엄격성
T_q = 경력이 q인 국어교사(평가자) j의 엄격성
F_k = 평가 척도 k-1에 대한 척도 k의 난도

<수식 1>과 같이 각 국면의 분석 결과를 logit 척도로 변환하면, 중학생 논설문 쓰기 능력(점수), 평가 요인의 난도(難度), 성과 경력에 따른 국어교사의 엄격성 수준에 대한 값을 얻을 수 있다. 이 값을 내림차순으로 정리하면 중학생 논설문 평가에서 발견되는 국어교사의 쓰기 평가 특성을 효과적으로 파악할 수 있다. 특히, 국어교사(평가자)에 대한 모형 적합도 수치는 국어교사의 평가 점수가 중학생(피험자)의 논설문 쓰기 능력을 얼마나 정확하게 추정하는지에 대한 정보를 제공해 준다. 성별 및 경력별 적합도 통계치는 국어교사를 성별과 경력에 따라 구분할 때, 어떤 집단의 국어교사들이 중학생의 논설문 쓰기 능력을 일관성 있게 평가하고 있는지에 대한 정보를 제공해 준다.

수집한 자료는 FACETS ver 3.66.1(Linacre, 2004)을 이용하여 분석하였다.[2] 분석은 국어

2) FACETS 프로그램은 다국면 Rasch 모형을 컴퓨터를 통해 분석할 수 있도록 Linacre(1996)가 개발한

교사의 쓰기 평가 특성 중에서 엄격성과 일관성을 중심으로 수행되었으며, 평가 기준표의 하위 요인이 단일한 차원을 측정하는지에 대한 적합도 분석도 수행되었다. 분석 모형에 대한 적합도는 내적합 지수와 외적합 지수를 이용하여 검증을 실시하였다(Wright & Masters, 1982). 내적합 및 외적합 지수는 기대치가 1.0인 X^2 분포를 이루는데, 지수가 1.0이면 자료가 모형에 적절하다는 것을 뜻한다.

국어교사의 쓰기 평가에 나타난 편향 분석은 평가자인 국어교사와 학생 글에 대한 상호작용 분석, 국어교사와 평가 요인에 대한 상호작용 분석을 통해 확인하였다. 국어교사의 쓰기 평가에 나타난 편향 분석은 평가자인 국어교사와 학생 글에 대한 상호작용 분석과 국어교사와 쓰기 평가 요인에 대한 상호작용 분석을 통해 확인하였다. 분석 결과로 제시되는 t 값과 자유도 정보를 통해, 편향이 유의한지 검증하였다. 편향 분석은 특정 국면들 사이에 서로 상호작용이 있는지를 알아보는 분석이다.

이 분석은 서로 다른 엄격성의 차이가 일정한 유형을 가지고 모형의 기댓값과 많은 차이가 나타날 때 그 값이 유의하면 제시된다. 일반적으로 분석의 기준은 표준화값인 Z-score로 판별한다. 이 값이 +2보다 크거나 -2보다 작으면, 그 편향의 크기가 유의미한 것으로 나타난다. 그러나 FACETS 프로그램은 이 값 대신 t 값을 제시한다. 이 값의 유의한지를 검증하기 위해서는 유의도를 계산하여 유의수준을 판별함으로써 편향의 유의미성을 파악할 수 있다(장소영·신동일, 2009)

Ⅳ. 연구 결과 및 논의

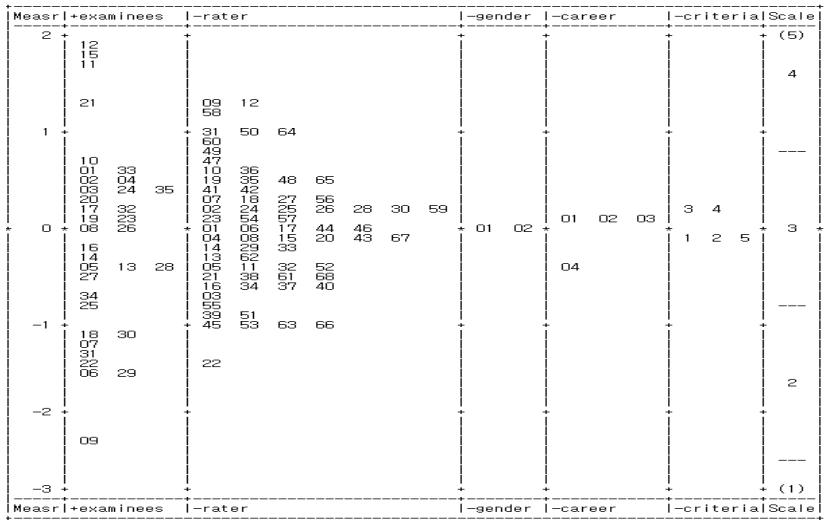
1. 국어교사의 평가자 특성

가. 엄격성 및 일관성 분석

국어교사의 쓰기 평가 자료를 FACETS 프로그램으로 분석하여 얻은 logit 값을 도식화하여 제시하면 [그림 1]과 같다. [그림 1]은 각 국면별 logit 모수 추정치를 표현한 것이다. 첫째 열의 지수(Measr)는 Rasch 모형의 공통 척도(logit)를 의미하는데, 국어교사(평가자)의 엄격성, 학생(피험자)의 능력, 평가 요인의 분포를 판단할 때 활용된다. 둘째 열(+examinees)은

것이다. 다국면 Rasch 모형은 Rasch에 의해 제시된 1모수 문항 반응 이론을 쓰기와 같은 수행평가의 채점 행위를 분석하기 위하여 개발된 것인데(지은림 외, 2000), 수식 계산이 복잡하고 어렵다는 이 모형의 문제를 해소하기 위하여 Linacre(1996)가 FACETS 프로그램을 개발한 것이다. FACETS 프로그램은 다국면 Rasch 모형 분석을 지원함으로써 평가의 여러 국면에 대한 측정치를 제공해 준다.

중학생 논설문 35편의 logit 분포를, 셋째 열(-rater)은 평가자인 국어교사 68명의 logit 분포를 보여준다. 넷째 열(-gender)은 국어교사의 성별 logit 분포를 보여주고, 다섯째 열(-career)은 국어교사의 경력별 logit 분포를 보여주며, 여섯째 열(-criteria)은 평가 요인의 logit 분포를 보여준다. 마지막으로 일곱째 열(Scale)은 평가 요인의 점수 척도별 logit 분포를 도식화한 것이다.



〔그림 1〕 학생×국어교사×성별×경력×평가 요인의 분포도

[그림 1]은 이 연구에서 설정한 5국면, 즉 중학생(피험자), 국어교사(평가자), 국어교사의 성별, 국어교사의 경력, 평가 요인을 동일 척도에 제시한 것이다. 따라서 각 국면의 logit 분포를 명시적으로 확인할 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. <그림 1>은 전체 국면의 엄격성 또는 난도에 대한 정보를 보여주는데, 특히 주목해야 할 것은 평가자, 성별, 경력별 국면이다. 이 국면들이 바로 중학생 논설문 평가에서 드러나는 국어교사의 평가 특성을 보여주기 때문이다. 평가자 국면에서는 중학생 논설문을 평가한 국어교사 68명 각각의 엄격성이 어떻게 차이가 있는지를 보여주고, 성별 국면과 경력별 국면에서는 국어교사들이 성별 및 경력별로 엄격성 수준이 어떻게 다른지를 보여준다.

〈표 3〉 국어교사의 평가 엄격성 수준과 적합도

평가자	엄격성 (logit)	SE	Infit MS	Outfit MS	평가자	엄격성 (logit)	SE	Infit MS	Outfit MS
12	1.34	0.1	0.58	0.59	53	-0.46	0.1	0.81	0.9
09	1.31	0.1	1.12	1.05	18	0.34	0.1	0.53	0.53
59	1.24	0.1	1.2	1.12	07	0.32	0.1	0.83	0.83
65	1.02	0.1	2.09	2.51	57	0.31	0.1	0.88	0.88
31	1.00	0.1	1.29	1.24	27	0.30	0.1	1.6	1.59
50	1.00	0.1	1.66	1.59	02	0.29	0.1	1.04	1.05
61	0.88	0.1	0.62	0.61	28	0.24	0.1	0.66	0.66
49	0.83	0.1	0.77	0.8	26	0.20	0.1	0.76	0.76
47	0.70	0.1	0.57	0.58	55	0.19	0.1	0.85	0.84
36	0.62	0.1	0.68	0.68	24	0.18	0.1	1.71	1.69
10	0.62	0.1	0.96	0.95	30	0.17	0.1	1.05	1.05
66	0.56	0.1	4.51	4.87	25	0.16	0.1	0.93	0.92
35	0.53	0.1	0.7	0.71	60	0.16	0.1	0.46	0.46
19	0.52	0.1	0.82	0.82	58	0.16	0.1	1.91	1.91
48	0.51	0.1	1.33	1.39	23	0.10	0.1	0.86	0.89
42	0.48	0.1	0.68	0.68	01	0.05	0.1	0.87	0.86
41	0.40	0.1	1.28	1.27	44	0.05	0.1	0.75	0.75
46	0.03	0.1	1.39	1.38	06	0.03	0.1	1.13	1.12
17	0.03	0.1	1.69	1.67	62	-0.48	0.1	0.72	0.73
08	-0.05	0.1	0.71	0.7	38	-0.52	0.1	0.91	0.91
20	-0.08	0.1	0.65	0.65	40	-0.53	0.1	0.79	0.79
15	-0.08	0.1	0.85	0.84	16	-0.55	0.1	0.48	0.48
68	-0.08	0.1	1.11	1.1	34	-0.6	0.1	1.14	1.13
04	-0.10	0.1	0.68	0.69	37	-0.61	0.1	1.02	1
43	-0.10	0.1	1.66	1.65	03	-0.62	0.1	0.75	0.75
33	-0.13	0.1	1.1	1.1	56	-0.77	0.1	0.99	0.98
14	-0.14	0.1	1.05	1.05	39	-0.88	0.1	0.68	0.69
29	-0.17	0.1	0.9	0.9	51	-0.89	0.1	0.74	0.73
63	-0.24	0.1	0.67	0.66	64	-0.95	0.1	0.39	0.4
13	-0.29	0.1	0.93	0.93	45	-0.96	0.1	0.93	0.92
11	-0.35	0.1	0.71	0.7	54	-0.96	0.1	0.73	0.72
32	-0.40	0.1	1.29	1.29	67	-0.97	0.1	0.96	0.96
05	-0.40	0.1	1	1	52	-1.22	0.1	0.76	0.75
21	-0.43	0.1	0.77	0.77	22	-1.41	0.1	0.92	0.92
$\overline{X}$	logit= .00		SE= .10			Infit MS= 1.00		Outfit MS= 1.01	
s	logit= .61		SE= .00			Infit MS= .55		Outfit MS= .59	

국어교사의 평가 특성을 좀 더 구체적으로 파악하기 위하여 수치로 정리하면 <표 3>과 같다. 이 표는 각 국어교사별 logit 점수 및 표준오차, 내적합 지수, 외적합 지수를 제시한 것이며, logit 값의 크기에 따라 내림차순으로 정리한 것이다. <표 3>에 따르면, 국어교사의 엄격성은 -1.41 logit(SE= .01)부터 1.34 logit(SE= 0.1)까지의 범위에 분포한다. 이러한 분포는 X^2= 2612.5(p < .001), 분리신뢰도 R= .97로서 통계적으로 유의한 차이가 있다.

국어교사 12는 중학생 논설문에 대해 가장 엄격하게 평가하였으며(1.34 logit), 국어교사 22는 가장 관대하게 평가한 것으로 나타났다(-1.41 logit). 적합도 분석에 의하면 국어교사 04, 12, 16, 18, 20, 28, 36, 39, 42, 47, 60, 61, 63, 64 등 14명(20.58%)이 내적합 지수가 0.7보다 낮아 과적합으로 분석되었다. 이 국어교사들은 중학생 논설문을 평가할 때 1~5점 척도를 변별력 있게 적용하지 않고, 22222, 33333, 44444처럼 특정 점수에 종속적으로 평가한 것으로 판단된다. 학생 논설문의 전체적인 인상이나 평가 결과를 평가 요인에 모두 동일하게 적용한 것으로 보인다.

이에 비해 17, 24, 27, 43, 46, 48, 50, 58, 65, 66 등 10명(14.70%)은 내적합 지수가 1.3보다 높아 부적합으로 분석되었다. 이 9명의 국어교사는 중학생 논설문을 평가할 때 쓰기 능력이 낮은 글에 대해 예상치 않게 높은 점수를 부여하였거나, 혹은 그와 반대로 쓰기 능력이 우수한 글에 대해 낮은 점수를 부여하였던 것으로 해석된다. 학생들의 논설문 쓰기 능력을 적절하게 읽어내지 못했던 것으로 보인다.

이외의 국어교사들은 내적합 지수가 모두 0.7~1.3의 범위 내에 속하여 적합한 것으로 분석되었다. 적합도를 보인 국어교사는 전체 68명 중 44명(64.70%)인데, 이들의 평가 결과만이 엄격성이 일관성 있게 적용된, 안정적인 양상을 보이고 있다고 할 수 있다. 평가자로 참여한 국어교사의 수가 제한적이지만, 이러한 결과를 바탕으로 하여 추론컨대 국어교사의 약 56% 정도만 논설문 평가에서 일관성을 유지하고 있다고 볼 수 있다.

평가자로 참여한 국어교사 중에서 56% 정도만 학생 글을 평가할 때 내적 일관성을 적절하게 유지하고 있다는 이러한 연구 결과는 국어교사의 쓰기 평가 전문성과 관련하여 주목된다. 44%의 국어교사들이 적합한 수준을 벗어났다는 것은 일반적인 기대와 달리 국어교사의 쓰기 평가 전문성이 높지 않을 가능성이 있음을 보여주기 때문이다. 특히 부적합으로 판정된 9명(13.04%)의 국어교사는 학생들의 쓰기 능력을 적절하게 추정하지 못하는 것으로 판단되므로 더욱 유의할 필요가 있다.

학생 글을 평가할 때 평가의 일관성을 유지하기 위해서는 평가자의 많은 노력이 요구된다. 쓰기 평가가 이루어지는 평가 상황도 올바로 파악해야 할 뿐만 아니라 내적 일관성을 흔드는 내적 요인과 외적 요인을 인식하고 통제할 수 있어야 한다. 그러나 44%의 국어교사들은 이러한 전문성 있는 노력을 기울이지 않았거나 기울이지 못했던 것으로 보인다. 국어교사들의 내적 일관성을 높이기 위해서는 더 많은 훈련과 노력이 이루어져야 할 것이다.

나. 국어교사의 성 및 경력에 따른 평가 특성 분석

국어교사의 성(性)과 경력은 쓰기 평가에 영향을 미치는 요인으로 알려져 있다. 국어교사의 성은 학생 성별 인식과 상호작용 효과를 보이기도 하며, 국어교사의 경력은 관대한 정도에 영향을 미친다. 그러므로 국어교사의 성과 경력에 따라 엄격성이 차이가 있는지를 분석할 필요가 있다. 우선, 국어교사의 성에 따라 엄격성이 차이가 있는지를 분석하였다. 분석 결과는 <표 4>와 같다.

〈표 4〉 국어교사의 성에 따른 엄격성 수준과 적합도

평가자 성별	엄격성(logit)	SE	Infit MS	Outfit MS
남(01)	-0.03	0.03	0.82	0.83
여(02)	0.03	0.01	1.04	1.05
$\overline{X}$	.00	.02	.93	.94
s	.05	.01	.11	.11

<표 4>에 따르면, 이에 따르면 국어교사의 성에 따른 엄격성은 남자 국어교사가 -0.03 logit, 여자 국어교사가 0.03 logit으로 분석되었다. 즉, 여자 국어교사는 남자 국어교사보다 중학생 논설문을 더 엄격하게 평가한 것으로 파악된다. 두 집단의 엄격성은 X^2= 4.6(p < .05), 분리신뢰도 R=.57로 통계적으로 유의한 차이를 보였다. 한편, 국어교사의 성에 따른 적합도 분석에서 내적합 지수가 남자 국어교사는 0.82, 여자 국어교사는 1.04을 보였다. 국어교사의 성에 따라 엄격성은 차이가 있지만, 평균으로 계산하여 볼 때 그 엄격성은 남자 교사와 여자 교사 모두 일관성 있게 유지되고 있다고 할 수 있다.

남자 교사에 비해 여자 교사가 쓰기 평가에서 더 엄격한 수준을 보인다는 사실은 선행 연구와도 유사한 점이 있다(박영민·최숙기, 2009, 2010a, 2010b). 비록 이 연구에 참여한 국어교사의 수, 글 유형 등이 제한적이지만, 쓰기 평가의 엄격성은 여자 국어교사가 남자 국어교사보다 높다는 결론을 이끌어낼 수 있다. 그러므로 쓰기 평가 전문성 신장을 위한 프로그램을 적용하거나 평가자간 신뢰도를 확보하기 위해 평가자 훈련을 실시할 경우, 이러한 국어교사의 성별 특성을 고려할 필요가 있다. 남자 국어교사들은 관대하게 평가함으로써 학생 쓰기 지도에 필요한 세부 정보를 학생 글에서 정확하게 읽어내지 못할 수 있고, 여자 국어교사들은 엄격하게 평가함으로써 학생들의 쓰기 효능감을 떨어뜨릴 수 있기 때문이다.

다음으로 국어교사의 경력에 따른 엄격성에 대한 분석 결과는 <표 5>와 같다.

〈표 5〉 국어교사의 경력에 따른 엄격성 수준과 적합도

평가자 경력	엄격성(logit)	SE	Infit MS	Outfit MS
1~5년(01)	0.08	0.02	1.11	1.13
6~10년(02)	0.12	0.02	0.90	0.92
11~20년(03)	0.16	0.03	0.98	0.98
20년 초과(04)	-0.35	0.1	0.81	0.9
$\overline{X}$	.00	.04	.95	.98
s	.20	.03	.11	.09

<표 5>에 따르면, 경력에 따른 엄격성은 -0.35 logit에서 0.16 logit의 범위에 분포하고 있다. 이러한 분포는 $X^2 = 27.6\,(p < .001)$, 분리신뢰도 R= .93로 통계적으로 유의한 차이가 있다. 그러므로 국어교사는 경력에 따라 엄격성이 유의한 차이가 있다고 할 수 있다.

경력에 따른 국어교사의 엄격성을 구체적으로 살펴보면, 1~5년 경력의 국어교사(0.08 logit)가 가장 엄격한 것으로 나타났고, 11~20년(0.16 logit), 6~10년(0.12 logit), 20년 초과(-0.35 logit)의 순서로 엄격성이 높았다. 6~10년의 경력 집단, 11~20년 경력 집단의 엄격성은 유사한 수준을 보였으며, 20년 초과의 경력을 지닌 국어교사는 가장 관대하게 평가한 것으로 나타났다. 한편, 국어교사의 경력에 따른 적합도 분석에 따르면, 네 집단 모두 내적합 지수가 0.81~1.11에 분포하여 엄격성이 일관성 있게 유지된 것으로 나타났다.

국어교사의 경력에 따른 평가의 엄격성 분석은 국어교사의 경력이 쓰기 평가의 엄격성과 밀접한 관련이 있음을 보여준다. 1~5년의 경력을 지닌 국어교사들은 엄격성이 높고 20년 초과 경력을 지닌 국어교사들은 엄격성이 낮다는 사실을 통해서 국어교사로서 경력이 증가할수록 학생 글을 관대하게 평가하는 경향이 뚜렷하다는 점을 확인할 수 있다(박영민·최숙기, 2009, 2010b). Weigle(1998)은 Rasch 모형을 활용하여 쓰기 평가 경험이 적은 평가자들이 평가 경험이 많은 평가자들보다 훨씬 더 엄격하게 평가한다는 점을 분석함으로써 이 연구와 동일한 결과를 보고한 바 있다.

이런 맥락에서 보면, 경력이 많은 국어교사는 '학생 글이 웬만하다 싶으면' 좋은 점수를 준다는 뜻인데, 이러한 현상은 경력이 증가함에 따라 국어교사의 지식 효과의 영향력이 커지면서 나타나는 것으로 이해할 수도 있고, 경력이 증가하면서 형성되는 매너리즘이나 노력의 부재로 인해 나타나는 것으로 이해할 수도 있다(Hayes & Bajzek, 2008). 그러므로 학생 글에 대한 평가를 계획하고 시행할 때에는 국어교사의 경력을 고려할 필요가 있다. 경력 차이에 따라 학생의 개별적인 점수뿐만 아니라 집단의 평균 점수가 달라질 수 있기 때문이다.

다. 평가 요인에 따른 평가 특성 분석

평가 요인에 따른 국어교사의 평가 특성을 파악하기 위하여, 논설문 평가 요인 국면의 엄격성을 분석하였다. 평가 요인의 엄격성은 평가 요소의 난도(難度)로 설명된다. 평가 요인이 얼마나 엄격한가는 곧 그 요인에 적합하게 글을 쓰는 것이 얼마나 어려운가에 대응하기 때문이다. 그러므로 이 연구에서는 평가 요인의 엄격성을 난도로 해석하여 논의하고자 한다. 평가 요인 국면에 대한 분석 결과는 <표 6>과 같다. 논설문 평가의 요인은 내용, 조직, 표현, 형식 및 어법, 단어 선택의 5가지로 구성되었으며, 각 요인별 점수 척도는 5점이다.

〈표 6〉 논설문 평가 요인에 따른 난도 수준과 적합도

평가 요인	난도(logit)	SE	Infit MS	OutfitMS
조직(01)	-0.11	0.03	0.97	0.98
표현(02)	-0.12	0.03	1.06	1.07
내용(03)	0.18	0.03	0.93	0.94
형식・어법(04)	0.17	0.03	1.05	1.06
단어 선택(05)	-0.13	0.03	1	1.01
$\overline{X}$	.00	.03	1.00	1.01
s	.14	.00	.05	.05

<표 6>에 따르면, 평가 요인에 따른 난도는 -0.13 logit에서 0.18 logit의 범위에 분포하고 있다. 평가 요인의 난도는 $X^2 = 149.5\,(p < .001)$로서 하위 요인별로 통계적으로 유의한 차이가 있는 것으로 분석되었다. <표 7>에 의하면, 국어교사들은 중학생 논설문을 평가할 때, 단어 선택 요인에서 가장 높은 점수를 부여하였고, 표현, 조직, 형식 및 어법, 내용의 순서로 높은 점수를 부여하였다. 내용 요인은 국어교사들로 가장 엄격한 평가, 즉 가장 낮은 점수를 받은 것으로 분석되었다. 한편, 평가 요인에 따른 적합도 분석에 의하면, 내적합 지수가 0.97부터 1.06의 범위에 속함으로써 5개의 평가 요인은 모두 적합한 것으로 확인되었다.

<표 7>의 분석 결과는 설명문의 예와 차이가 있다(박영민·최숙기, 2010b). 학생 글이 논설문이든 설명문이든 단어 선택 요인은 국어교사들로부터 가장 관대한 평가를 받았으나, 차순위로 관대한 평가를 받은 하위 요인은 서로 달랐다. 설명문은 형식 및 어법이었지만, 논설문은 표현이었다. 형식 및 어법 요인은 설명문 평가에서는 관대한 것으로 분석되었으나, 논설문 평가에서는 엄격한 요인으로 분석되었다. 또한, 내용 요인은 설명문 평가에서는 조직이나 표현 범주보다 점수를 받기 쉽지만, 논설문에서는 가장 점수를 받기 어려운 것으로 분석되었다. 이러한 결과는 글 유형이 무엇인가에 따라 국어교사의 평가 중점이 달라진다는 사실을 보여준다.

2. 국어교사의 평가 편향 분석

가. 국어교사와 학생 글 사이의 편향

68명의 국어교사가 35편의 중학생 논설문을 평가할 때 국어교사와 학생 글, 국어교사와 평가 요인 간에 편향적 채점 경향을 보이는지에 대한 분석은, 일반적으로 국어교사들의 쓰기 평가에서 발생할 수 있는 편향의 문제를 이해하고 개선하는 데 중요한 정보를 제공하여 줄 수 있다. 이 연구에서는 평가자로 참여한 국어교사 68명과 중학생 논설문 35편 사이에 존재하는 상호작용의 효과를 분석하였다. 분석 결과에 따르면, 2,380개의 상호작용 가운데 131개에서 유의미한 편향이 발견되었다. 이 중 81개의 편향은 국어교사가 예측 점수보다 더 높은 점수를 부여한 것(즉, 관대한 것으로)으로 나타났고, 나머지 50개의 편향은 국어교사가 예측점수보다 더 낮은 점수를 부여한 것(즉, 엄격한 것으로)으로 나타났다.

81개의 편향은 모든 국어교사들에게서 나타난 것은 아니며, 전체 68명의 국어교사들 가운데 48명이 학생 글과의 상호작용 분석에서 유의한 편향이 있는 것으로 나타났다. 학생 글에서는 35편 중 학생 글 5를 제외한 나머지 글에서 모두 편향이 나타났다. 편향이 나타난 국어교사 48명은 평균 2.78개의 편향을 보였다. 이들 중 국어교사의 최소 편향은 1개였으며, 이에 해당하는 국어교사는 3, 7, 8, 11, 13, 21, 29, 34, 38, 44, 45, 48, 53, 56, 60, 61, 62, 67이었다. 이에 반해 최대 편향은 17개로 나타났고, 이에 해당하는 국어교사는 평가자 65였다. 평가자 65는 이미 앞서 평가자 특성 분석에서 부적합을 보인 평가자로 나타나 이 국어교사의 평가 결과를 수용할 때는 매우 신중한 접근이 요구된다.

다음 <표 7>은 국어교사와 학생 글의 편향적 상호작용 분석의 결과 중 최소 편향과 최대 편향을 보인 예를 제시한 것이다.

〈표 7〉 국어교사와 학생 글의 최소 편향 및 최대 편향

평가자	피험자	Obsvd	Exp.	Bias	error	t	fit MS
3	26	24	16.2	-3.17	1.03	3.07	0.9
7	32	19	14.1	-1.57	0.56	2.82	0.3
8	35	23	15.9	-2.49	0.76	3.27	0.6
11	23	9	15.8	2.29	0.62	-3.72	0.4
13	26	21	15.2	-1.86	0.6	3.11	0.2
21	25	18	13.2	-1.55	0.56	2.79	0.4
29	3	9	16.1	2.36	0.62	-3.84	0.4
34	13	10	15	1.65	0.59	-2.8	0.6
38	19	24	16.6	-3.05	1.03	2.95	0.8
44	4	24	15.9	-3.25	1.03	3.16	0.8

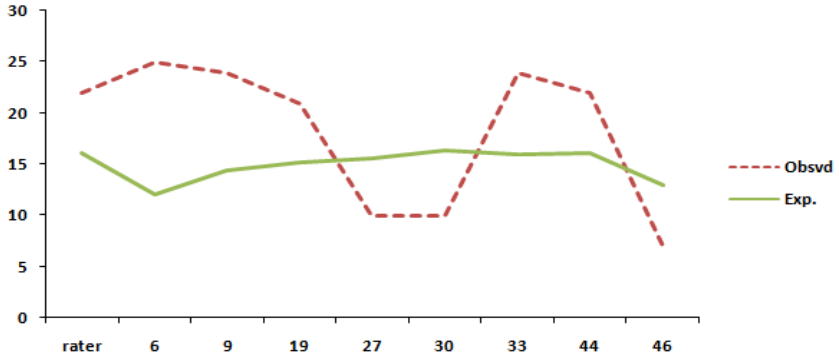
45	18	19	14.1	-1.56	0.56	2.8	0.8
53	13	10	16.1	2	0.59	-3.41	0.7
60	11	12	17.2	1.71	0.57	-2.99	0.3
61	34	19	13.8	-1.68	0.56	3.02	0.2
62	32	22	15.8	-2.05	0.65	3.14	0.4
67	6	15	10	-1.64	0.58	2.85	0.7
65	1	7	14.6	2.82	0.78	-3.63	0.8
65	2	6	14.5	3.56	1.05	-3.41	0.9
65	9	23	7	-5.75	0.76	7.56	1.7
65	10	5	14.9	4.77	1.67	-2.85	0.1
65	11	8	18.2	3.44	0.67	-5.14	0.5
65	12	7	18.8	4.14	0.78	-5.33	0.6
65	14	20	11.9	-2.59	0.57	4.56	0.5
65	17	7	13.7	2.49	0.78	-3.21	0.8
65	20	24	13.7	-3.98	1.03	3.86	0.8
65	22	14	8.7	-1.8	0.58	3.12	1.1
65	25	21	10.4	-3.45	0.6	5.79	1
65	26	6	12.8	3	1.05	-2.87	0.9
65	27	24	11.3	-4.79	1.03	4.65	0.9
65	29	16	8.5	-2.55	0.57	4.49	0.2
65	33	21	14.8	-1.99	0.6	3.33	0.4
65	34	22	10.7	-3.74	0.65	5.74	0.4
65	35	6	14.1	3.44	1.05	-3.29	0.9

<표 7>에서 평가자의 편향의 정도를 의미하는 편향치가 음수라는 것은 평가자가 예상 점수보다 관대하였음을 뜻하며, 양수는 평가자가 예상 점수보다 엄격하였음을 의미한다. 최소 편향을 보이는 국어교사들은 예상 점수보다 더 관대하게 채점한 것으로 보인다. 그러나 최대 편향의 정도를 보이는 국어교사 65의 경우, 엄격한 정도와 관대한 정도의 비율이 유사한 것으로 나타났다. 이와 같은 경향은 국어교사의 평균 정도의 편향을 보이는 국어교사 46과도 비슷하다. 그러므로 편향의 빈도나 혹은 편향의 비율에 대한 정보만을 가지고 평가자의 평가 전문성을 진단하는 것에는 한계가 있을 수 있다. 그럼에도 불구하고 일반적으로는 국어교사의 논설문 평가에 나타난 편향을 살펴보기 위해 상호작용 분석을 실시할 경우에는, 편향성이 나타난 빈도를 통해 평가자로서 국어교사가 가지고 있는 편향의 일반적 정보를 파악할 수 있다.

이와 함께, 국어교사와 학생의 글 변인, 즉 문체, 글씨, 내용 선호 양상, 성별 등의 변인의 상호작용 분석도 국어교사의 평가 편향을 분석하는 데 유효하다. 이러한 분석 결과는 쓰기

평가의 방법을 개선하는 데에도 기여할 수 있다. 국어교사의 편향이 높게 나타나는 학생 글을 판별해 내면, 국어교사의 평가 편향에 영향을 미치는 학생 글의 변인이 무엇인지를 판단할 수 있고, 따라서 이를 토대로 하여 평가자 훈련의 방법을 마련할 수 있기 때문이다.

예를 들어, 국어교사와 학생 글의 상호작용 분석 결과에 따르면, 유의미한 편향이 나타난 학생 글은 34편이므로 학생 글 당 평균 3.85개의 편향이 발견되었다. 그런데 학생 논설문 35편 중 학생 글 4(부록 참조)는 [그림 2]와 같이 총 9명의 국어교사의 편향이 발생하였다. 이는 학생 글 당 평균적인 편향 수준을 매우 상회하는 것이다. 그러므로 이 학생 글에서 발견되는 특성을 바탕으로 하여 평가자 훈련이나 평가자 협의를 실시하고 편향의 문제를 해소할 수 있다면, 학생 논설문에 대한 국어교사의 평가 신뢰도를 개선 할 수 있게 될 것이다.



〔그림 2〕 학생 글 4에 대한 국어교사들의 예측 점수와 관찰 점수

나. 국어교사와 평가 요인 사이의 편향

앞에서는 국어교사와 학생 글 사이의 상호작용을 분석하였는데, 여기에서는 국어교사와 평가 요인 간의 상호작용을 분석하여 평가 편향을 살펴보았다. 일반적으로 쓰기 평가에서 평가자의 편향을 초래하는 변인 중 하나가 바로 평가 요인이다. 평가자가 인식하고 있는 이상적인 평가 요인이 각각 다를 수 있으므로 자연스럽게 평가의 편향이 발생할 수 있다.

분석 결과에 따르면, 340개의 상호작용 가운데 42개의 편향이 유의한 것으로 나타났다. 내용 요인과 관련한 편향은 15개, 조직 요인 관련 편향은 3개, 표현 요인 관련 편향은 5개, 형식 및 어법 요인 관련 편향은 11개, 단어 선택 요인 관련 편향은 8개로 나타났으며, 평가 요인 당 평균 편향은 8.4개로 분석되었다. 그러므로 중학생의 논설문 쓰기 평가에서는 내용 요인, 형식 및 어법 요인과 관련한 국어교사의 편향이 두드러지게 나타난다고 할 수 있다.

그러나 42개의 유의한 편향을 국어교사 당 편향으로 분석하면 평균 0.62개이므로, 이 연구

에서 적용한 평가 요인별 편향은 낮은 수준인 것으로 보인다. 국어교사 68명 중, 평가 요인과의 편향이 발견되지 않은 국어교사는 43명(31.6%)이었다. 이러한 결과로부터 평가 요소와 관련한 국어교사의 평가 오류는 상대적으로 낮다고 예측할 수 있으며, 따라서 이 연구에서 적용한 평가 요소는 역으로 신뢰성을 지지해 준다고 할 수 있다.

나머지 국어교사 중, 편향이 1개로 나타난 평가자는 10명, 2개는 13명으로 나타났으며, 최대 편향은 3개이며 이에 해당하는 평가자는 2명으로 나타났다. 국어교사와 평가 요인 간의 편향 중 최소 편향과 최대 편향을 보인 예를 제시하면 <표 8>과 같다.

〈표 8〉 국어교사와 평가 요인 간의 최소 편향 및 최대 편향

평가자	평가요인	Obsvd	Exp.	Bias	error	t	fit MS
3	내용	105	114.5	0.45	0.22	-2.06	0.8
5	표현	115	104.6	-0.49	0.22	2.26	1
7	형식 어법	100	88.9	-0.53	0.22	2.46	0.5
11	조직	120	109.8	-0.48	0.22	2.2	1
18	표현	98	87.5	-0.51	0.22	2.32	0.3
18	형식 어법	77	87.7	0.54	0.23	-2.35	0.5
19	내용	79	90.6	0.58	0.23	-2.55	0.9
19	형식 어법	94	84.9	-0.44	0.22	2.02	0.6
23	내용	112	100.1	-0.56	0.22	2.59	0.6
23	단어선택	89	100.5	0.55	0.22	-2.5	0.7
27	내용	119	95.7	-1.1	0.22	5.06	0.9
27	형식 어법	99	89.8	-0.44	0.22	2.02	0.8
27	단어선택	73	96.1	1.15	0.23	-4.97	1.3
29	내용	90	105.1	0.72	0.22	-3.27	0.8
29	단어선택	119	105.5	-0.63	0.22	2.91	0.8
34	단어선택	125	113.7	-0.53	0.22	2.43	0.8
35	표현	93	83.7	-0.45	0.22	2.06	0.4
36	조직	79	89.5	0.52	0.23	-2.31	0.8
36	단어선택	102	89.8	-0.58	0.22	2.67	0.3
43	내용	94	104.3	0.49	0.22	-2.23	1.2
43	조직	115	104.4	-0.5	0.22	2.29	2
43	단어선택	92	104.7	0.6	0.22	-2.76	1.3
44	내용	88	101.3	0.64	0.22	-2.88	0.7
44	단어선택	116	101.7	-0.67	0.22	3.09	0.6
45	내용	134	122.5	-0.56	0.22	2.5	0.7
45	형식 어법	107	116.5	0.45	0.22	-2.07	0.9
47	내용	77	88.2	0.56	0.23	-2.47	0.5

47	형식 어법	92	82.5	-0.46	0.22	2.11	0.7
48	내용	79	91.6	0.63	0.23	-2.77	1.7
48	형식 어법	103	85.9	-0.82	0.22	3.78	1
55	내용	129	118.5	-0.5	0.22	2.27	0.7
55	형식 어법	97	112.5	0.73	0.22	-3.35	0.9
57	형식 어법	67	91.4	1.26	0.24	-5.25	1.7
60	내용	99	84.6	-0.7	0.22	3.2	0.5
61	형식 어법	92	106.1	0.67	0.22	-3.06	0.8
65	표현	99	84.7	-0.69	0.22	3.16	4.1
65	형식 어법	72	84.9	0.67	0.23	-2.85	4.5
66	내용	104	122.7	0.88	0.22	-4.05	0.6
66	단어선택	137	123.1	-0.68	0.23	3.01	0.9
67	내용	124	103.1	-0.98	0.22	4.5	0.6
67	표현	87	97	0.48	0.22	-2.18	0.8

<표 8>에 따르면, 최대 편향을 보인 국어교사 27은 학생 논설문을 평가할 때 국어교사들이 어떠한 편향을 보이는지를 알려주는데, 특히 국어교사 27은 평가 요인과의 편향이 평가 전체의 적합성에도 영향을 미치는 것으로 보인다는 점에서 주목된다. 국어교사 27은 평가 적합도 분석에서 부적합으로 판정되었기 때문이다.

평가 요인과의 편향은 앞서 살펴본 바와 같이, 국어교사가 개별 평가 요인에서 보일 수 있는 평가의 정확성에 대한 문제를 판별할 수 있도록 돕는 정보를 제공해 준다. 중학생 논설문 평가에 나타난 평가 요인에 대한 국어교사의 편향은 평가자의 개인적인 평가 오류에 대한 점검을 가능하게 해 준다. 따라서 특정 평가 요인에 대한 편향이 여러 국어교사들에게서 발견된다면, 이 평가 요인에 제시된 평가 기준에 대한 서술이 적절한지를 검토할 필요가 있을 것이다.

Ⅴ. 결론

이 연구의 목적은 중학생 논설문 평가에 나타난 국어교사의 평가 특성과 편향을 분석하는 데 있다. 평가자로 참여하는 국어교사 68명으로부터 평가 자료를 수집한 후, 다국면 Rasch 분석을 토대로 하여 국어교사가 논설문을 평가할 때 어떤 수준의 엄격성과 일관성을 보이는지를 확인하였으며, 국어교사의 성별 변인과 평가 요인이 논설문 평가에 어떤 편향을 보이는지를 확인하였다.

분석 결과는 다음과 같다. 첫째, 68명의 국어교사 중 14명(20.58%)이 과적합으로 분석되었다. 과적합으로 분석된 국어교사들은 평가 점수 척도를 변별력 있게 사용하지 않고 특정 점수에 종속된 경향을 보인 것으로 해석된다. 둘째, 68명의 국어교사 중 10명(14.70%)이 부적합으로 분석되었다. 부적합으로 분석된 국어교사들은 중학생들의 논설문 쓰기 능력을 적절하게 평가하지 못한 것으로 해석된다.

이상의 두 결과를 통해서 국어교사들은 부분적으로 쓰기 평가 전문성에서 문제가 있음을 추론할 수 있다. 국어교사 68명 중에서 단지 39명(56.52%)만이 내적 일관성이 적절하게 유지하면서 학생 글을 평가하고 있기 때문이다. 이러한 결과는 쓰기 평가 전문성 신장을 위한 방안을 모색할 필요가 있음을 일깨워 준다.

셋째, 성별에 따른 분석에서 일관성은 남자 국어교사와 여자 국어교사가 모두 적절하게 유지하였으나, 엄격성은 여자 국어교사가 남자 국어교사보다 더 높았다. 넷째, 경력에 따른 분석에서 20년을 초과하는 경력을 지닌 국어교사들의 엄격성이 가장 낮았다. 1~5년 경력 집단은 엄격성이 가장 높았다. 다섯째, 평가 요인에 따른 국어교사의 엄격성에 대한 분석에서, 단어 선택 요인에서 가장 높은 점수를 부여 하였고, 표현, 조직, 형식 및 어법, 내용의 순서로 점수를 각각 부여하는 것으로 나타났다.

이러한 결과를 통해서 쓰기 평가에서 발견되는 국어교사의 엄격성은 성별과 경력, 평가 요인에 따라 차이가 있다는 점을 확인할 수 있다. 그러므로 쓰기 평가 전문성 신장을 위한 프로그램을 적용하거나, 학교 평가에서 평가자 훈련을 실시할 때 이러한 특성을 고려할 필요가 있다.

여섯째, 국어교사와 학생 글 사이에서 발견된 2,380개의 상호작용 중, 131개만이 유의한 편향으로 분석되었다. 48명의 국어교사들이 학생 글과 편향을 보였고, 학생 글 5를 제외한 나머지 모든 글에서 편향이 발견되었다. 최소 편향은 1개, 최대 편향은 17개였다. 국어교사 한 사람당 편향은 2.78개로 나타났다. 일곱째, 국어교사와 평가 요인 사이에서 발견된 340개의 상호작용 중, 유의한 편향은 42개로 분석되었다. 각 평가 요인 당 평균 편향은 8.4개로 나타났으며, 국어교사 당 편향은 0.62개로 낮았다. 그러므로 이 연구에서 적용한 중학생 논설문 평가 요소는 국어교사 논설문 평가의 신뢰도를 유지하는 데 유의하다고 할 수 있다.

이와 같은 결과를 토대로 한, 쓰기 평가에 대한 함의는 다음과 같이 정리할 수 있다.

첫째, 논설문 쓰기 평가에 참여한 국어교사들은 다양한 평가자 효과(rater effect)를 보였으므로, 이를 적절하게 통제하기 위한 방안이 마련되어야 한다. 평가자 효과는 평가자의 특정 변인이 평가 결과에 영향을 미치는 것을 말한다. 이 연구의 결과에 따르면, 평가자로 참여한 국어교사들 중 일부는 엄격성이나 일관성에서 심각한 문제를 보였다. 이는 곧 평가자 오류로 이어지므로 신뢰성이 있는 평가, 전문성이 있는 평가를 위해서는 반드시 평가자 훈련을 통한 조정 방안이 마련될 필요가 있다.

둘째, 평가자 훈련을 계획할 때에는 국어교사의 평가자 특성을 고려해야 한다. 이를 위해서는 기대 점수에서 크게 벗어나는 평가를 하는 국어교사의 특성, 엄격성과 일관성에서 문제를 드러낸 국어교사의 특성 등을 분석하여 평가자 훈련에 반영할 필요가 있다. 특히, 여자 국어교사나 경력이 짧은 국어교사는 학생들의 논설문 쓰기 능력을 추정할 때 지나치게 엄격한 성향을 보이므로 이를 고려하여 평가 훈련을 계획해야 한다. 경력이 짧은 국어교사들의 평가가 엄격한 것은 지식 효과(knowledge effect)와도 관련이 깊으므로(박영민·최숙기, 2009), 이 점을 고려하여 평가자 훈련을 계획하는 것이 좋다.

셋째, 논설문 평가에서 내용 요인과 형식 및 어법 요인에 대한 평가 기준을 상세화하고 이에 대한 평가자 간의 협의를 확대할 필요가 있다. 이 연구 결과에 따르면, 논설문 평가를 할 때 국어교사들은 이 두 평가 요인에 대하여 가장 엄격한 평가 경향을 보였다. 그러면서도 동시에 이 두 평가 요인에서 평가자의 편향이 가장 빈번하게 발생하였다. 이는 논설문을 평가할 때 국어교사들이 내용과 형식 및 어법 요인에서 판정의 어려움을 겪고 있거나, 또는 편향을 불러일으키는 오류에 노출되어 있음을 암시하여 준다. 그러므로 이러한 문제를 해소하기 위하여 내용과 형식 및 어법 요인의 평가 기준을 상세화하고, 이 평가 기준을 협의하는 기회를 충분히 제공해야 할 것이다.

학생 글을 평가하는 데 영향을 미치는 요인은 매우 다양하다. 그러나 이와 같은 요인에 관한 탐색은 지금까지 제한적으로 이루어져 왔다. 특히, 국어 교육에서 이루어지고 있는 쓰기 평가 영역에서 평가자로서 국어교사의 특성에 대한 정보나 평가 전문성에 대한 정보를 제공해 주는 연구들은 매우 부족한 실정이다. 그러므로 이 연구 결과는 중학생 논설문을 평가할 때 드러나는 국어교사의 특성과 편향 유형을 이해하는 데, 그리고 이를 바탕으로 하여 쓰기 평가 방법을 개선하는 데 기여할 수 있을 것이다.

참 고 문 헌

김성숙(1995), 논술 문항 채점의 변동 요인 분석과 일반화 가능도 계수의 최적화 조건, **교육평가연구** 8(1), 35-57.

김성숙(2001), 채점의 변동 요인 분석 방법에 대한 고찰: 일반화 가능도 이론과 다국면 라쉬 모형의 적용과 재해석, **교육평가연구** 14(1), 303-325.

박영민・최숙기(2009), 현직 국어교사와 예비 국어교사의 쓰기 평가 비교 연구, **교육과정평가연구** 12(1), 123-143.

박영민・최숙기(2010a), 중학생 논설문 평가의 모평균 추정과 평가 예시문 선정, **국어교육** 131, 437-461.

박영민・최숙기(2010b), Rasch 모형을 활용한 국어교사의 쓰기 평가 특성 분석, **국어교육학연구** 37, 367-391.

성태제(2002). **현대 교육평가**. 서울: 학지사

이현숙(2008), 다국면 라쉬 모형을 적용한 논술 채점 상황에서 채점 설계 및 채점자 특성이 채점의 정확성에 미치는 효과, **교육평가연구** 21(4), 129-152.

장소영・신동일(2009), **언어교육평가 연구를 위한 FACETS 프로그램 : 기초과정편**, 글로벌콘텐츠.

지은림 외(2000), **RASCH 모형의 이론과 실제**, 교육과학사.

Barrett, S. (2001) The impact of training on rater variability. *International Education Journal*, 2 (1), 49-58

Barnes, L. L.(1990), Gender bias in teachers' written comments. In S. L. Gabriel & I. Smithson(eds.), *Gender in the classroom : Power and pedagogy,* Chicago, IL: University of Illinois Press, 140-159.

Bond, T.G. & Fox, C.M. (2001). *Applying the rasch model: Fundamental measurement in the human sciences*. Mahwah, New Jersey: Lawrence Erlbaum Associates,

Cason G.J. & Cason, C.L. (1984) A deterministic theory of clinical performance rating. *Evaluation and the Health* Professions, 7, 221-247.

Cantor, N. K., & Hoover, H. D. (1986). The reliability and validity of writing assessment: An investigation of rater, prompt within mode, and prompt between mode sources of error. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York

HarperCollins.

Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, 5, 171-191.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31, 93-112.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261-287). Mahwah, NJ: Erlbaum

Engelhard, G., Jr., Gordon, B., & Gabrielson, S. (1991). The influences of mode of discourse, experiential demand, and gender on the quality of student writing. *Research into the Teaching of English*, 26, 315-336.

Engelhard, G, & Myford, C.M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model.* New York: College Entrance Examination Board.

Gabrielson, S., Gordon, B., & Engelhard, G. Jr. (1995). The effects of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education*, 8, 273-290.

Etaugh, C. B., Houtler, B., & Ptasnik, 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.

Hayes, J. R. & Bajzek, D.(2008), Understanding and reducing the knowledge effect: Implication for writers, *Written Communication*, 25(1), 104-118.

Linacre, J. M..(1989/1993), **Many-facet Rasch measurement**, Chicago, IL : MESA Press.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12 (1), 54-71.

Lunz, M. E., & Stahl, J. A. (1993). Impact of examiners on candidate scores: An introduction to the use of multifacet Rasch model analysis for oral examinations. *Teaching and Learning in Medicine*, 5(3), 174-181.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Addison Wesley Longman Ltd.

Myford, C. M. (1991, April). *Judging acting ability: The transition from notice to expert*. Paper presented at the American Educational Research Association, Chicago IL.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. J*ournal of Applied Measurement*, 4, 386-422.

Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex Publishing.

Peterson, S. S. & Kennedy. K.(2006), Sixth-grade teachers' written comments on student writing: genre and gender influences, *Written Communication*, 23(1), 36-62.

Peterson, S.(1998), Evaluation and teachers' perceptions of gender in sixth-grade student writing, *Research in the Teaching of English*, 33(2), 181-208.

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.

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.

Weigle, S. C.(1998), Using FACETS to model rater training effects, *Language Testing*, 15(2), 263-287.

Wright, B. D. & Linacre, J. M.(1990), Measuring the impact of judge severity on examination scores, *Applied Measurement in Education*, 3, 331-345.

Wright, B.D. & Masters, G.N. (1982), *Rating Scale Analysis*, Chicago: MESA Press.

· *논문접수 : 2011년 1월 1일/ 수정본 접수 : 2011년 3월 8일/ 게재승인 : 2011년 3월 11일*

ABSTRACT

The Rater Characteristics and Raters Bias in Korean Language Teacher's Persuasive Writing Assessment

Choi Sook-Ki(Adjunct Professor, Daegu University)
Park Young-Min(Associate Professor, Korea National University of Education)

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 persuasive writing of middle school students randomly. the rating data was analyzed by the computer program FACETS (ver 3.66.1, Linacre, 2004), with FACETS analyses run on the persuasive writing. The results are following. (1) the 14 teachers(20.58%) out of 68 were founded as overfiting raters. (2) the 10 teachers(14.70%) out of 68 were founded as underfiting or misfiting raters. (3) In the rater's gender, male and female teachers showed acceptable fit that means stable rating. But the female teachers had higher severity than that of the male teacher. (4) In the rater's career, teachers who above 20 years of teaching career reported the lowest severity level, and teachers who 1~5 years of teaching career reported the highest severity level within teacher groups. (5) MFRM revealed several recurring bias patterns among rater subgroups. In rater-category and rater-examinee bias interactions, Some raters also rated more severely and more leniently than expected.

Key Words : 국어교사(Korean language teachers), 쓰기평가(writing assessment), 논설문(persuasive writing), 다국면 라쉬 모형(Many-Facets Rasch Model), FACETS 프로그램(FACETS program), 평가자 편향(rater bias)

부록: <학생 글 4>[3)]

학생들이 자유롭지 못한 것이 아니다. 공부에 자유롭기 위해 규제를 한 것이다. 아직 자신이 한 일에 대해 책임지지 못 할 나이인 중학생에게 지금까지 규제해 왔던 복장 및 두발규제를 자율화 시킨다면 또 그 속에 더 많은 규제와 문제가 일어날 것이다.

지금까지 지켜져 왔던 복장 및 두발 규제가 갑자기 사라진다면 교복을 착용했을 때처럼 단정하고, 바른 행동, 공부에 집중할 수 있을까? 지금까지 철저한 규제를 한 만큼 학생들은 한풀이라고 하는 듯이 학생 신분에 맞지 않는 옷을 입을 가능성이 높다. 그렇기 때문에 두발, 복장은 규제해야 한다.

옷은 학교에서만 입는 것이 아니다. 평소에도 충분히 꾸미고 개성을 살려 입을 수 있다. 그런데 사복을 공부하는 학교까지 꾸미고 온다면 문제가 있다. 학생의 신분에 맞는 교복을 입고, 나중에 스스로를 책임질 수 있을 만한 나이가 됐을 때 자신의 개성을 마음것 살리면 되는 것이다. 공부하려 오는 학교인 만큼 공부에 집중 할 수 있는 단정한 교복과 두발이 알맞다고 생각한다.

사복을 입게 해 달라고, 자율화 해 달라고 우리의 자유를 박탈하지 말라는 학생들의 목소리가 점점 커져가고 있다. 그러나 이런 규제를 푼다면 학생들은 학교가 원하는 데로 단정하지만 자신의 개성을 살리는 교복보다 편한 그런 쪽으로만 나아갈까? 단정함, 깔끔한 학생다움의 학교는 더 이상 바랄 수 없을지도 모른다.

미국의 복장 두발이 자유다. 그래서 우리나라 사람들도 자율화를 외치고 있다. 미국은 옷에 많은 신경을 쓰지 않는다. 청바지에 티셔츠 하나 걸친 사람들이 대부분이다. 하지만 우리나라는 외모에 엄청난 신경을 쏫고 있다. 그런데 청소년들은 어련할까. 미국이 자유인 만큼 단점도 무수히 많다. 그런한 단점들은 우리나라가 자율화를 했을 때 충분히 일어날 수 있는 일이다. 꾸미는 것이 나쁜 것이 아니다. 학생에 맞지 않는 차림세, 화장이 문제인 것이다.

지금까지 지켜왔던 규제들을 푼다면 분면 문제가 생길 것이다. 학생들의 자유박탈을 하기 위해 규제를 만든 것이 아니다. 더욱 나은 공부 환경과 단정한 학생 본분에 맞는 규제를 만든 것이다. 개성을 살리는 것에 가서 개성을 살리고 공부하는 곳에서는 공부에 집중하는 학생이 맞다고 생각한다. 교복, 머리의 규제는 질풍노도의 시기인 학생들에게 매우 적합하다고 생각한다.

3) 맞춤법 오류는 실제 학생 글에 있는 것을 그대로 옮긴 것이다.