예비국어교사의 중학생 논설문 평가에서 발견되는 엄격성 및 일관성의 특성

박영민*

<차 례>

Ⅰ. 서론

학교 쓰기 평가에서는 평가자로 참여하는 국어교사의 특성이 평가 결과에 영향을 미치므로 국어교사는 높은 수준의 전문성을 갖추고 있어야 한다. 평가 기준에 따라 글을 변별적으로 읽어내는 능력, 읽어낸 결과를 척도에 따라 적확한 점수로 해석해 내는 능력, 필체와 같은 외적 정보나 학생의 이름, 성(性), 출신 지역, 출신 학교 등과 같은 개인적 정보를 통제하면서 글을 정확하게 평가하는 능력 등이 그러한 전문성에 해당한다.

그런데 이와 더불어 중요하게 언급되는 쓰기 평가 전문성 중의 하나가 바로 적절한 수준으로 일관성을 유지하는 능력이다. 여기에서 말하는

* 한국교원대학교 제2대학 국어교육과 교수

일관성이란 국어교사가 학생 글을 평가할 때 채점의 엄격성을 처음부터 끝까지 동일하게 유지하는 정도를 말한다. 국어교사가 일관성을 적절한 수준으로 유지하면 채점의 엄격성이 모든 학생 글에 일관성 있게 적용되는 특성을 보인다. 한편, 바로 앞에서 언급한 엄격성은 쓰기 평가에 참여하는 국어교사가 학생 글에 점수를 매길 때 엄격하게 평가하는 정도를 뜻한다. 엄격성이 높으면 점수를 박하게, 엄격성이 낮으면 점수를 후하게 부여하는 특성을 보인다(박영민 · 최숙기, 2010 ; 박종임 · 박영민, 2012).

쓰기 평가에서 말하는 일관성은 엄격성을 일관성 있게 유지하는 정도를 뜻하므로 이 일관성을 평가자에게 요구되는 '엄격성의 일관성'이라고 부르기도 한다. 이 점에서 보면, 일관성은 엄격성을 전제로 한, 따로 분리해서 다루기 어려운 쓰기 평가의 방법적 개념이라고 할 수 있다. 학교에서 이루어지는 쓰기 평가의 상황이라면, 국어교사가 학생 글을 평가할 때 점수를 후하게 주는 정도(엄격함)나 박하게 주는 정도(관대함)를 적절하게 유지하는 정도를 일관성이라고 부르는 것이다.

일관성을 적절한 수준으로 유지하는 것이 중요한 이유는 이것이 쓰기 평가 결과의 변별도와 신뢰도를 확보할 수 있기 때문이다. 쓰기 평가 환경의 변화에 따라 일관성이 흔들린다면, 그 평가 결과를 통해서는 학생들의 쓰기 능력을 올바로 변별하기 어려울 뿐만 아니라 그 결과가 믿을 만한지를 판단하기도 어렵다. 가령, 유사한 쓰기 능력이 있다고 추정되는 철수와 영희 두 학생에 대해, 어떤 국어교사가 일관성 없이 철수의 글은 엄격하게 평가하여 박한 점수를 주고 영희의 글은 관대하게 평가하여 후한 점수를 준다면, 그 평가 결과로는 철수와 영희의 쓰기 능력을 변별하기도 어렵고 그 평가 결과를 의심 없이 수용하기도 어렵다.

쓰기 평가가 이루어지는 기간 내내 학생 글 전편에 대해 일관성을 적절한 수준으로 유지한다는 것은 매우 어려운 일이다(박종임 · 박영민, 2011). 그래서 일관성 유지를 쓰기 평가 전문성의 주요 요인으로 간주하거니와, 쓰기 평가 경험이 많은 국어교사들도 일관성을 유지하는 일이 쉽지 않다. 평가 경험만으로는 일관성 유지가 충족되지 않는다. 단지 학생 글을 평가해 본 경험이 많다고 해서 어떠한 조치 없이 엄격성을 일관성 있게 유지

하는 능력이 형성되는 것이 아니라는 뜻이다. 평가 경험이 엄격성을 일관성 있게 유지하는 데 기여하는 주요 변인이었다면, 사실 일관성 유지 능력을 쓰기 평가의 전문성으로 인정할 필요도 없었을 것이다.

이러한 상황을 고려하여, 이 연구에서는 예비국어교사를 대상으로 하여 중학생 논설문 평가에서 드러나는 엄격성과 일관성의 특성을 분석함으로써 현재 예비국어교사들이 가지고 있는 쓰기 평가와 관련된 교육적 정보를 제공하고자 한다. 예비국어교사는 양성 교육을 이수하는 과정 중에 있지만, 이들은 국어교사로 입직할 평가자들이므로 이들이 전문적인 평가자들에게 요구되는 엄격성과 일관성이 어떠한 특징을 보이는지를 검토하는 것은 의의가 있다.

이러한 목적을 위하여 이 연구에서는 예비국어교사 32명을 평가자로 선정한 후, 임의로 수집한 중학생 논설문 40편을 평가하도록 설계하고 채점 결과를 수집하였다. 이들로부터 수집한 채점 자료는 컴퓨터 프로그램인 FACETS을 활용하여 분석하였다. 이 프로그램은 다국면 Rasch 모형을 적용하여 일관성을 분석할 수 있도록 개발된 도구이다.

이 연구는 예비국어교사 32명을 표집하여 자료를 수집하고 분석하였다는 점에서 결과를 일반화하는 데 한계가 있을 수 있다. 좀 더 일반화된 결론을 얻으려면, 평가 과정을 동일하게 유지하는 데 어려움이 따르기는 하지만, 표집 크기를 확대하여 평가 결과를 수집하고 분석할 필요가 있다. 이러한 한계에도 불구하고, 이 연구 결과를 통해 얻는 정보는 국어교사 양성 과정에서 쓰기 평가 전문성 신장을 위해 교육과정에 무엇을 포함해야 하는지, 더 나아가 무엇을 더 추가하거나 중점적으로 교육해야 하는지를 결정할 때 중요한 판단의 근거가 될 수 있을 것이다.

Ⅱ. 이론적 배경

1. 평가자의 엄격성

수행평가가 여러 영역에서 활용되면서 수행평가에서 발생하는 오류에 대한 논의가 지속적으로 전개되어 왔다. 평가 이론에서는 어떤 변인에 의해 발생하는 평가 결과의 변화나 변동을 오류로 규정하고 있는데, 이를 검토해 온 선행 연구에 따르면 수행평가에서 발생하는 주요 오류는 그 주된 원인이 평가자의 특성과 관련되어 있다. 일반적으로 수행평가는 평가자의 개별적 특성이나 주관적 특성이 영향을 미치므로 수행평가의 오류는 결국 평가자에게 그 원인이 있다고 보는 것이다(Cantor & Hoover, 1986 ; Engelhard, 1992, 1994 ; Engelhard, Gordon, & Gabrielson, 1991 ; Gabrielson, Gordon & Engelhard, 1995 ; McNamara, 1996 ; Ruth & Murphy, 1988 ; 김성숙, 1995, 2001).

수행평가 중에서도 쓰기 평가처럼 언어 수행을 평가하는 경우에는 평가자가 보이는 엄격성이 평가 결과의 차이를 발생시키는 오류의 원인으로 지적되기도 했다(Engelhard, 1994 ; Engelhard & Myford, 2003 ; Lumley & McNamara, 1995). 한 피험자에 대해 어떤 평가자는 엄격하게 평가하고 어떤 평가자는 관대하게 평가하면, 평가자들이 내린 평가 결과는 차이가 클 수밖에 없고 이것이 피험자의 능력을 서로 다르게 추정하게 되는 오류의 원인이 된다고 보았던 것이다.

평가자의 엄격성은 평가자가 얼마나 평가 기준을 엄격하게 적용하여 점수를 부여하는가, 즉 점수를 얼마나 후하게 주거나 박하게 주는가를 뜻한다. 가령, 쓰기 평가에서 평가자가 엄격하면 평가 척도의 평균 이하로 채점하는 경향을 보이며, 평가자가 관대하면 평가 척도의 평균 이상으로 채점하는 경향을 보인다. 1~5점 척도를 쓴다면, 엄격한 평가자는 1~3점 사이의 점수를 부여하고 관대한 평가자는 3~5점 사이의 점수를 부여하는 것이다. 평가자가 보이는 엄격한 정도와 관대한 정도는 대립적인 특성에 해당하므로, 일반적으로는 엄격성을 대표 개념으로 하여 엄격성의 높고

낮음으로 표현한다.

언어 수행평가에서 평가자의 엄격성을 오류로 보았던 이유는 평가자의 엄격성이 높으면 피험자의 능력을 과소 추정하는 경향이 발생하고 평가자의 엄격성이 낮으면 피험자의 능력을 과대 추정하는 경향이 나타나기 때문이다. 학교에서 언어 수행평가가 고부담 시험으로 치러지는 경우라면, 예를 들어 논술문 쓰기를 승급 또는 졸업 시험으로 채택한 상황이라면, 평가자로 참여하는 교사의 엄격성에 따라 탈락자가 한 명도 없을 수도 있고 무더기로 나올 수도 있다. 원점수를 보정하지 않는 절대평가 방법을 취한다면 평가자의 엄격성은 평가 결과 및 교육적 의사결정에 매우 큰 영향을 미치는 변인이 된다고 할 수 있다.

이처럼 평가자들이 보이는 엄격성의 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 바탕으로 하여 추정하는 피험자의 능력이 신뢰할 만하다고 보기는 어려울 것이다. 이러한 맥락에서 Cronbach(1990)는 평가자가 불러일으키는 가장 민감하면서도 심각한 오류를 평가자의 엄격성 효과(severity effect)로 설명한 바 있다.

평가자의 엄격성이 차이가 나는 이유는 평가자들이 소유하고 있는 기준이나 척도에 차이가 있기 때문이다. 객관적인 평가 결과를 얻으려면 평가자마다 서로 다른 이 엄격성 수준을 조정할 필요가 있다. 지금까지 소개된 가장 효과적인 방법은 평가자 훈련이나 평가자 협의를 적용하는 것이다. 평가자들은 자기 자신이 가지고 있는 평가 기준이나 척도를 비교하고 대조하고 조율하고 조정함으로써 엄격성 수준을 비슷한 수준으로 맞출 수 있다.

그러나 평가자의 엄격성 효과를 완전히 해소하는 것은 쉬운 일이 아니다. 특정한 프로그램으로 구성된 평가자 훈련을 거치더라도 그 차이가 소거되기는커녕 여전히 뚜렷하다는 연구 결과가 보고되기도 하였다(Barrett, 2001 ; Lumley & McNamara, 1995 ; Weigle, 1998). 엄격성의 차이를 완전히 소거

하는 것이 불가능하다면, 완전 소거를 목표로 하기보다는 수용할 수 있는 적절한 수준에서 엄격성의 차이를 인정하는 것이 적절할 수도 있을 것이다.

2. 평가자의 일관성

쓰기 평가와 같은 수행평가는 평가자의 판단이 점수로 환산되어 부여되므로 평가자 신뢰도는 측정의 일관성을 설명하는 주요한 지표로 쓰인다. 평가자 신뢰도 중에서 평가자 개인의 내적 신뢰도는 한 명의 평가자가 여러 명의 피험자를 평가할 때 요구되는 일관성을 뜻하는데, 이것이 바로 평가 전문성과 밀접한 관련이 있다. 평가자들 사이의 신뢰도는 평가자들끼리 얼마나 유사한 정도로 평가하는가를 말한다면, 평가자의 내적 신뢰도, 즉 평가자 일관성은 피험자의 균일하게 읽어내는 능력과 관련이 있기 때문이다.

평가자의 일관성이란 평가자가 피험자에게 점수를 부여하는 경향이 평가회기 내내 일정하게 지속되는 현상을 말한다. 평가자는 평가를 마칠 때까지 일관성을 유지할 수 있어야 평가의 내적 신뢰도를 확보할 수 있다. 평가자가 일관성을 유지하기 위해서는 평가자 자기 자신이 내면화하고 있는 평가 기준이나 척도가 변하지 않도록 통제하면서 평가를 시행할 수 있어야 한다(박종임 · 박영민, 2011 ; 최숙기 · 박영민, 2011). 물론 이렇게 하기 위해서는 평가자가 많은 노력을 기울여야 한다. 이러한 이유에서 연구자들은 평가자 훈련의 중요한 목표는 평가자의 일관성 확보가 되어야 한다고 강조하였다(Cason & Cason, 1984 ; Lunz & Stahl, 1993).

평가자가 첫 피험자부터 마지막 피험자까지 일관성을 유지하면서 평가하였는지를 알아보기 위해서는 다국면 Rasch 분석을 적용하는 것이 효과적이다. 이 분석 방법은 일반적인 평가자 신뢰도로 간주되어 온 평가자 간 신뢰도뿐만 아니라 평가자 내 신뢰도에 대한 정보를 제공해 주기 때문이다. 다국면 Rasch 분석을 적용했을 때는 얻는 일관성이 바로 평가자 내 신뢰도에 대응한다. 다국면 Rasch 분석을 적용한 후 일관성의 수준을 확인

했을 때, 일반적으로 내적합 지수가 1.3보다 크면 부적합으로, 0.7보다 작으면 과적합으로 판정한다. 이를 바탕으로 하여 평가자가 일관성을 얼마나 적절하게 유지하였는지를 파악할 수 있다(Bond & Fox, 2001 ; McNamara, 1996).

Rasch 분석에서 부적합은 평가 결과의 일관성이 지나치게 적거나 일관성이 없는 것으로 해석하며, 과적합은 일관성이 지나치게 높은 것으로 해석한다. 일관성이 과도하게 낮다는 것은 평가자가 피험자에게 일관성이 없게 점수를 부여했다는 뜻이므로 평가자가 피험자의 능력을 올바로 추정하지 못했다고 볼 수 있다.

한편, 일관성이 지나치게 높다는 것은 평가자가 채점한 점수가 거의 변화가 없다는 것을 뜻하므로, 서로 다른 피험자를 올바로 변별하지 못했거나 서로 다른 평가 요소를 올바로 구별하지 못했다고 볼 수 있다. 예를 들어 서로 구별된 평가 요소임에도 불구하고 평가자가 모든 평가 요소에 특정 점수를 일률적으로 부여할 때, 예를 들면 모든 평가 요소의 점수를 1111, 2222, 3333처럼 부여할 때 내적합 지수가 과적합으로 산출된다(Wright, & Linacre, 1990 ; McNamara, 1996; Bond & Fox, 2001). 어떤 연구자들은 과적합이 발생하는 이러한 현상을 특정 점수에 종속된 평가 현상으로 말하기도 한다(박영민 · 최숙기, 2010 ; 최숙기 · 박영민, 2011).

Lunz & Stahl(1990)은 다국면 Rasch 분석을 통해 일정한 채점 기간 동안 평가자의 일관성이 적절하게 유지되는지를 조사한 바 있다. 연구에 참여한 3명의 평가자는 1~4일 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 평가회기 동안 평가의 일관성을 적절하게 유지하지 못하는 문제를 발견하였다. Lumley & McNamara(1995)은 말하기 평가 기간 중 평가자 일관성이 어떻게 변화하는지를 분석하였는데, 특히 평가자와 평가 시간의 상호작용 효과를 분석하여 평가자의 일관성의 변화가 평가에 미치는 영향을 분석한 바 있다.

Ⅲ. 연구 방법

1. 연구 대상

중학생들이 작성한 논설문을 평가할 때 예비국어교사들이 보이는 엄격성의 일관성을 분석하기 위하여 예비국어교사 32명을 평가자로 선정하고 연구를 수행하였다. 평가자로 선정된 예비국어교사들은 모두 연구자가 재직하는 대학에서 교사 양성 과정을 이수하는 학생들이다. 평가자로 선정된 예비국어교사 32명 중 남자는 13명(40.6%)이었고 여자는 19명(59.4%)이었다.

평가자로 위촉된 예비국어교사들은 모두 교사 양성 대학의 국어교육과 3학년에 재학 중이었으며, 이전에 중학생이 작성한 글을 직접 평가해 본 경험은 없었다. 이들은 교육 실습을 이수하지 않으며, 쓰기 평가와 관련된 이론을 부분적으로 학습한 상태였다. 그러나 쓰기 평가와 관련된 실행적 훈련은 이수하지 않았다. 이러한 상황은 국어교사 양성 대학의 보편적인 상황과 크게 다르지 않으므로 평가자로 위촉된 예비국어교사들은 평균적인 예비국어교사의 모습을 반영하고 있다고 할 수 있다. 다만, 중학생 논설문 평가를 원활하게 진행하기 위하여 평가자로 참여한 예비국어교사들에게는 간략한 평가 방법이나 절차 등에 대해 안내가 제공되었다.

2. 검사 도구

이 연구에서 예비국어교사들에게 제공한 중학생 글은 논설문인데, 그 글은 <표 1>과 같은 과제에 따라 작성되었다.

<표 1> 중학생 논설문 쓰기 과제

다음 학기부터 우리 학교는 쉬는 토요일 없이 매주 토요일마다 등교하는 방안을 검토하고 있습니다. 그래서 우리 학교의 학생, 선생님, 부모님을 대상으로 의견을 모으고 있는 중입니다. 여러분은 매주 토요일마다 등교하는 것에 대해 찬성하는지, 또는 반대하는지 하나의 입장을 선택해 주세요. 자신의 입장을 밝히고, 자신과 반대되는 주장을 펴는 학교 선생님이나 친구를 설득하는 글을 쓰세요.

이 과제에 따라 중학생들이 작성한 논설문 중에서 40편을 임의로 선정하여 평가 자료집을 제작하였으며, 평가자로 참여한 예비국어교사들은 중학생 논설문 40편을 모두 채점하였다. 이 연구는 엄격성의 일관성을 분석하는 것을 목적으로 하고 있으므로 완전 배분 채점 모형을 적용하였다.

예비국어교사들에게 제공한 중학생 논설문은 글씨 모양이나 학생의 성별 정보 등 주관적인 영향 요인을 통제하기 위하여 워드 프로세서로 입력하여 작성였다. 중학생 논설문을 입력할 때 의미의 혼동이나 오해를 초래할 수 있는 띄어쓰기는 수정하였으나 맞춤법 오류는 수정하지 않았다. 입력 및 채점의 어려움이 뒤따르지만 맞춤법 오류를 수정하지 않은 이유는 이와 관련된 내용이 채점 기준에 포함되어 있기 때문이다.

평가자로 참여한 예비국어교사들에게는 평가 기준표와 채점표를 함께 제공하였다. <표 2>는 예비국어교사들에게 제공한 평가 기준표를 제시한 것이다. 이 기준표는 Spandel & Culham(1996)의 기준을 우리나라 중학생의 논설문 평가에 적합하도록 수정한 것이다.[1] 평가 기준의 하위 요인에는 내용, 조직, 표현, 단어 선택, 형식 및 어법 등이 포함되었다. 평가 기준표에는 5점, 3점, 1점에 해당하는 평가 기준을 제시하였지만, 예비국어교사들이 1점에서 5점 사이의 점수를 자유롭게 줄 수 있도록 설계하였다.

---

1) Spandel & Culham(1996)은 학생들이 작성하는 대부분의 글에 적용할 수 있도록 하기 위하여 '아이디어와 내용'(idea and content), '조직'(organization), '어조'(voice), '단어 선택'(word choice), '문장 유창성'(sentence fluency), '쓰기 관습'(convention)을 평가 요소로 설정하였다. 그러나 이 평가 요소들은 영어를 기반으로 한 것이어서 국어로 작성한 글에는 부적합한 것이 포함되어 있다. '문장 유창성'이 그러한 예이다. 그러므로 이 연구에서는 Spandel & Culham(1996)의 평가 요소 중에서 '문장 유창성'은 삭제하였으며, '목소리'는 '표현'으로 수정하여 평가 요소를 설정하였다.

<표 2> 중학생 논설문 평가 기준표

평가 요인	평가 기준
내용	• 5점 : 주장이 명확하고 타당한 추론과 논거를 통하여 독자를 효과적으로 설득시킨다. 객관적인 사실과 개인적 경험을 통하여 주장을 효과적으로 뒷받침하고 있다. 내용이 신선하며 독창적이다. 또한 자신의 주장과 상반된 주장에 대해서도 언급하며 타당한 반론을 제시한다. • 3점 : 독자가 주장과 근거를 파악하는 데에는 문제가 없지만 주장이 피상적이고 이를 뒷받침하는 근거가 부족하기 때문에 독자를 효과적으로 설득하는 데에는 실패한다. 내용이 피상적이고 주제에 대한 새로운 시각을 거의 보여주지 못한다. 자신의 주장과 상반된 주장을 언급하고는 있지만 타당한 반론을 제시하지 못한다. • 1점 : 주장이 분명하게 드러나지 않기 때문에 글쓴이의 의도를 이해하기 어렵고 주장과 근거의 구별이 없이 산만하게 제시되어 있으며, 비슷한 내용이 계속해서 반복되고 있다.
조직	• 5점 : 자신의 주장이 더욱 의미 있게 전달될 수 있도록 주장을 뒷받침하는 세부 내용들이 적절하게 배치되어 있고, 그 순서가 논리적이며 유기적이다. 글의 서론, 본론, 결론이 잘 드러나게 조직되어 있어서 독자가 내용 전개를 이해하고 예측하는 데에 도움을 준다. • 3점 : 글의 배열이 어느 정도 논리적이기는 하지만, 주장과 이를 뒷받침하는 세부 내용들이 다소 혼란스럽게 배열되어 있어서 독자의 주의가 분산된다. 또한 글의 서론, 본론, 결론이 드러나기는 하지만 도입이나 결론의 역할을 효과적으로 하고 있지는 않다. • 1점 : 글의 배열이 전혀 유기적이지 않고, 독자가 파악할 수 있는 전개 구조가 드러나지 않아서 내용 파악을 어렵게 한다.
표현 (어조 및 태도)	• 5점 : 글쓴이의 생각이 명확하게 표현되었고, 독자가 쉽게 이해할 수 있도록 표현되었으며, 독창성이나 개성이 잘 나타난다. 글쓴이가 주제에 대하여 적극적으로 참여하고 반응하려는 의지와 주체적인 목소리가 잘 드러나고, 독자를 설득시키려는 노력이 돋보인다. • 3점 : 글쓴이의 생각이 명확하게 드러나고 독자가 이해할 수 있도록 표현되었으나, 독창성이나 개성이 잘 드러나지 않고, 주제에 대한 적극적인 참여 의식이 부족하다. • 1점 : 글쓴이의 생각이 명확하게 표현되지 않고, 내용을 기계적으로 나열하여 독창성이 드러나지 않는다. 또한 독자가 쉽게 이해하기 어려운 표현들이 많다.
단어 선택	• 5점 : 내용을 정확하고 자연스럽게 전달할 수 있는 단어가 선택되었다. • 3점 : 대체적으로 단어 선택이 내용 전달에 무리가 없으나, 부적절한 단어들이 포함되어 있다. • 1점 : 내용을 전달하는 단어가 매우 제한적이거나 부적절하여 내용 이해에 오히려 혼란을 준다.

평가 요인	평가 기준
형식 및 어법	• 5점 : 글쓴이는 표준적인 쓰기 관습(어법, 구두점, 철자, 단락 구분 등)을 잘 이해하고 있으며 독자의 가독성을 고려하여 이러한 관습을 효과적으로 사용하고 있다. • 3점 : 제한된 범위에서만 표준적인 쓰기 관습을 지키고 있다. • 1점 : 어법, 구두점, 철자 등에서 잘못된 것이 많아 내용 파악을 방해한다.

3. 연구 절차

예비국어교사 32명에게는 중학생들이 작성한 논설문 40편, 논설문 쓰기 과제 제시문, 평가 기준표, 채점표, 평가 과정 기록표를 제공하였다. 평가자로 참여하는 예비국어교사들에게 논설문 쓰기 과제를 알려줌으로써 중학생들이 어떠한 과제에 따라 논설문을 작성한 것인지를 알 수 있도록 하였다. 평가자로 참여한 예비국어교사들에게 논설문 쓰기 과제 제시문과 평가 기준표, 평가 과정 기록표 등에 대해 설명한 후, 궁금한 사항을 연구자에게 질문하도록 하였다. 이 후에는 평가자 훈련 없이 개별적으로 자유로운 장소에서 중학생 논설문을 평가하도록 하였다.

평가자 훈련을 실시하지 않은 이유는 이 연구의 목적이 예비국어교사들이 보이는 일관성의 수준을 확인하는 데 있기 때문이다. 평가자 훈련을 적용하면 예비국어교사 개개인이 가지고 있는 특성이 소거될 수 있으므로 목적에 부합하는 정보를 획득할 수 없다. 평가자 훈련이 일관성을 유지하는 데 효과가 있는지를 살피고자 한다면 평가자 훈련을 독립 변인으로 하는 별도의 연구를 설계하여 진행해야 할 것이다. 예비국어교사의 논설문 평가 자료는 2011년 5월 18일부터 5월 31일까지 수집하였다.

4. 분석 방법

중학생 논설문 평가에서 나타나는 예비국어교사의 엄격함의 일관성을 분석하기 위해 이 연구에서는 평가자로서의 예비국어교사, 예비국어교사

의 성별, 논설문의 평가 요인, 평가 대상자로서의 학생이라는 4국면을 Rasch 모형에 적용하여 분석하였다. 이를 바탕으로 하여 중학생 논설문 평가에서 성(性)이 p인 예비국어교사 j가 중학생 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 다국면 분석 모형을 정리하면 <표 3>과 같다.

<표 3> 다국면 Rasch 분석 모형

$$\log(P_{nijpk} / P_{nij(k-1)}) = B_n - D_i - C_j - U_p - F_k$$

P_{nijpk} = 성(性)이 p인 평가자(예비국어교사) j가 평가 요인 i에 대해 중학생 n에게 점수 k-1보다 하나의 등급 점수가 높은 k를 줄 확률
$P_{nijp(k-1)}$ = 평가자(국어교사) j가 평가 요인 i에 대해 학생 n에게 점수 k-1을 줄 확률
B_n = 피험자(중학생) n의 논설문 쓰기 능력
D_i = 논설문 평가 요인 i의 엄격성
C_j = 평가자(예비국어교사) j의 엄격성
U_p = 성(性)이 p인 평가자(국어교사) j의 엄격성
F_k = 평가 척도 k-1에 대한 척도 k의 엄격성

<표 3>과 같이 각 국면의 분석 결과를 logit 척도로 변환하면 중학생들의 논설문 쓰기 능력, 평가 요인의 엄격성, 예비국어교사가 보이는 엄격함의 수준에 대한 값을 얻을 수 있다. 이 값을 내림차순으로 정리함으로써 중학생 논설문 평가에서 확인되는 예비국어교사의 엄격성의 정도를 효과적으로 파악할 수 있다.

이 값 중에서 예비국어교사에 대한 모형 적합도를 나타내는 수치는 예비국어교사의 평가 점수가 중학생의 논설문 쓰기 능력을 얼마나 정확하게 추정했는지에 대한 정보를 제공해 준다. 또한, 성별 적합도 통계치는 예비국어교사를 성별로 구분했을 때 어떤 집단의 예비국어교사들이 중학생의 논설문 쓰기 능력을 일관성 있게 평가하고 있는지에 대한 정보를 제공해 준다.

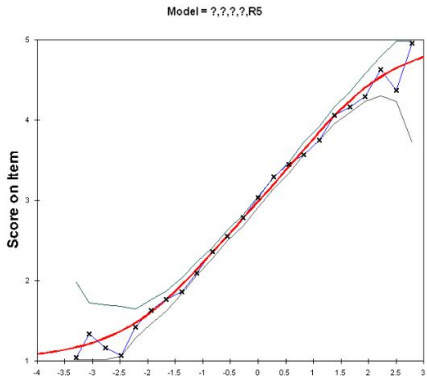
예비국어교사가 채점한 자료는 컴퓨터 프로그램 FACETS ver 3.66을 활용하여 분석하였다. 분석은 예비국어교사의 쓰기 평가 특성 중에서 엄격성의 일관성을 중심으로 수행되었으며, 평가 기준표의 하위 요인이 단

일한 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.

Ⅳ. 연구 결과 및 논의

1. 자료의 적합도

예비국어교사들이 채점한 자료가 다국면 Rasch 모형을 적용한 분석에 적합한지를 알아보기 위하여 모형 적합도 분석을 실시하였으며, 그 결과는 <그림 1>과 같다. <그림 1>은 이론적 문항 반응 곡선과 실제 평가 자료에서 나타난 반응 빈도 사이의 관계를 도식적으로 표현한 것이다. 가로축은 '학생 능력 추정치−과제 난도 추정치'를 나타내고, 세로축은 그에 따른 기대 점수를 나타낸다. 모형 적합도 분석은 95% 신뢰구간을 설정하고 관찰치 곡선이 문항 특성 곡선에 얼마나 잘 부합하는지를 판단하는 것이다.

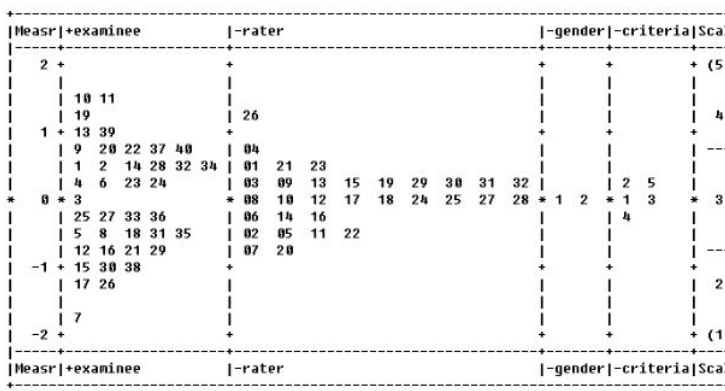


<그림 1> 중학생 논설문 채점 자료의 모형 적합도

<그림 1>과 같이 대부분의 관찰치는 신뢰구간 내에 위치하고 있으므로 예비국어교사 32명으로부터 수집한 중학생 논설문 채점 자료는 다국면 Rasch 분석 모형에 적합하다고 할 수 있다.

2. 엄격성 및 일관성 분석

중학생 논설문에 대한 예비국어교사의 평가 자료를 FACETS 프로그램으로 분석하여 얻은 logit 값에 따라 분포 현황을 제시하면 <그림 2>와 같다. <그림 2>는 각 국면의 logit 모수 추정치를 표현한 것이다. 첫째 열의 지수(Measr)는 Rasch 모형의 공통 척도(logit)를 의미하는데, 국어교사(평가자)의 엄격성, 능력, 평가 요인의 분포 위치를 판단할 때 활용된다. 둘째 열(+examinee)은 중학생 논설문 40편의 logit 분포를, 셋째 열(-rater)은 평가자인 예비국어교사 32명의 logit 분포를 보여준다. 넷째 열(-gender)은 예비국어교사의 성별 logit 분포를 보여주고, 다섯째 열(-criteria)은 평가 요인별 logit 분포를 보여준다. 마지막으로 여섯째 열(Scale)은 평가 요인의 척도별 logit 분포를 도식화한 것이다.



<그림 2> 수험자×채점자×성별×평가영역 분포도

<그림 2>는 이 연구에서 설정한 4국면, 즉 예비국어교사(평가자), 중학생(피험자), 예비국어교사의 성별, 쓰기 평가 요인을 동일 척도에 제시한 것이다. 따라서 각 국면의 logit 분포를 상대적이면서도 명시적으로 확인할 수 있다.

<그림 3>는 전체 국면의 엄격성에 대한 정보를 보여주는데, 특히 주목해야 할 것은 평가자, 성별, 경력별 국면이다. 이 국면들이 바로 중학생 논설문 평가에서 드러나는 예비국어교사의 특성을 보여주기 때문이다. 평가자 국면에서는 중학생 논설문을 평가한 예비국어교사 32명 각각의 엄격성이 어떻게 차이가 있는지를 보여주고, 성별 국면과 경력별 국면에서는 국어교사들이 성별 및 경력별로 엄격성 수준이 어떻게 다른지를 보여준다.

중학생 논설문에 대한 예비국어교사의 평가 특성을 좀 더 구체적으로 파악하기 위하여 logit 값을 표로 정리하면 <표 4>와 같다. 이 표는 각 국어교사별 logit 점수 및 표준오차, 내적합 지수, 외적합 지수를 제시한 것이며, logit 값의 크기에 따라 내림차순으로 정리한 것이다.

<표 4>에 제시된 적합도는 예비국어교사가 평가 과정에서 보인 엄격성의 일관성이 적합한 수준인지를 알려주는 지표이다. 이 적합도를 통해서 엄격성이 일관성 있게 유지되었는지를 판단할 수 있다. 일반적으로 내적합 지수가 0.7~1.3의 범위에 있을 때 일관성이 적절하게 유지된 것으로 해석한다.

<표 4> 채점자 엄격성 수준 및 적합도

평가자	엄격성 (logit)	SE	Infit MS	Outfit MS	평가자	엄격성 (logit)	SE	Infit MS	Outfit MS
1	0.41	.08	0.56	0.56	17	-0.12	.08	0.65	0.65
2	-0.60	.08	1.28	1.26	18	-0.08	.08	0.43	0.42
3	0.33	.08	0.62	0.63	19	0.33	.08	1.31	1.30
4	0.81	.08	0.86	0.85	20	-0.82	.08	1.56	1.59
5	-0.43	.08	0.89	0.88	21	0.46	.08	1.06	1.05
6	-0.33	.08	0.90	0.95	22	-0.38	.08	1.38	1.36
7	-0.65	.08	1.44	1.42	23	0.38	.08	1.3	1.29
8	0.02	.08	1.30	1.29	24	-0.08	.08	0.46	0.46

평가자	엄격성 (logit)	SE	Infit MS	Outfit MS	평가자	엄격성 (logit)	SE	Infit MS	Outfit MS
8	0.02	.08	1.30	1.29	24	-0.08	.08	0.46	0.46
9	0.19	.08	0.47	0.48	25	0.11	.08	0.81	0.80
10	-0.01	.08	0.76	0.77	26	1.25	.09	0.78	0.76
11	-0.62	.08	1.78	1.75	27	-0.08	.08	2.31	2.32
12	-0.04	.08	0.66	0.67	28	-0.03	.08	1.76	1.74
13	0.15	.08	1.08	1.08	29	0.21	.08	0.87	0.86
14	-0.16	.08	0.53	0.54	30	0.13	.08	0.55	0.55
15	0.22	.08	0.76	0.76	31	0.31	.08	0.93	0.95
16	-0.14	.08	1.25	1.26	32	0.23	.08	0.72	0.74
$\overline{X}$	logit=0.30		SE=0.08		Infit MS=1.00		Outfit MS=1.00		
s	logit=0.43		SE=0.00		Infit MS=0.45		Outfit MS=0.44		

<표 4>에 따르면, 예비국어교사의 엄격성은 -0.82 logit(SE= .08)부터 1.25 logit(SE= 0.9)까지의 범위에 분포하고 있다. 이러한 분포는 X^2=795.2 (p<.001), 분리신뢰도 R=.96으로서 통계적으로 유의하다.

평가자로 참여한 예비국어교사들을 엄격성 수준에 따라 두 집단으로 구분하면 <표 5>와 같다. 이 중 26번 예비국어교사는 중학생 논설문을 가장 엄격하게 평가한 것으로 나타났으며(1.25 logit), 20번 예비국어교사는 가장 관대하게 평가한 것으로 나타났다(-0.82 logit).

<표 5> 엄격성 수준에 따른 구분

엄격성 수준	예비국어교사 번호
엄격한 평가자	01, 03, 04, 08, 09, 13, 15, 19, 21, 23, 25, 26, 29, 30, 31, 32
관대한 평가자	02, 05, 06, 07, 10, 11, 12, 14, 16, 17, 18, 20, 22, 24, 27, 28

앞에 제시된 <표 4>에 따라 평가자로 참여한 예비국어교사의 일관성을 분석할 수 있는데, 내적합 지수가 0.7 이하이면 일관성이 과도한 '과적

합'으로, 내적합 지수가 1.3 이상이면 일관성이 없는 '부적합'으로 판정할 수 있다. 그러므로 내적합 지수가 0.7~1.3 사이에 놓이면 '적합'으로 판정할 수 있다.

적합도 분석에 따르면 예비국어교사 01, 07, 09, 12, 14, 17, 18, 20, 22, 24, 28, 30번 등 12명(37.5%)이 내적합 지수가 0.7보다 낮아 과적합 경향을 보이는 것으로 확인되었다. 이 예비국어교사들은 중학생 논설문을 평가할 때 각 평가 요인에 설정된 1~5점 척도를 변별력 있게 적용하지 못한 채 2-2-2-2-2, 3-3-3-3-3, 4-4-4-4-4처럼 특정 점수에 종속된 형태로 채점한 것으로 보인다. 평가 요소를 적용하면서 중학생 논설문을 평가할 때 1~5점 척도를 변별하지 않은 채, 논설문 전체에 대한 판단을 특정 점수로 정해 둔 다음, 그것을 일률적으로 모든 평가 요인에 적용하여 점수를 부여했을 가능성이 높다는 뜻이다.

이에 비해 예비국어교사 03, 11, 19, 27 등 4명(12.5%)은 내적합 지수가 1.3보다 높아 부적합 경향을 보이는 것으로 분석되었다. 이 예비국어교사들은 중학생 논설문을 평가할 때 쓰기 능력이 낮은 글에 대해 예상치 않게 높은 점수를 부여하였거나, 이와 반대로 쓰기 능력이 우수한 글에 대해 기대보다 낮은 점수를 부여했을 것으로 추정된다. 논설문의 질적 우열을 적절하게 변별해 내지 못할 때, 다시 말하면 평가자가 학생들의 논설문 쓰기 능력을 적절하게 읽어내지 못할 때 이러한 부적합 현상이 발생한다.

과적합 경향과 부적합 경향만을 대조해서 말한다면, 후자가 쓰기 평가 전문성이 훨씬 더 떨어지는 예에 해당한다. 부적합 경향을 보였다는 것은 평가자가 학생 글을 읽고 평가하였음에도 불구하고 학생들의 쓰기 능력을 올바로 추정해 낼 수 있는 힘이 부족하다는 의미이기 때문이다. 그러므로 부적합 경향을 보이는 예비국어교사들은 학생 글의 질적 수준을 올바로 변별해 낼 수 있는 전문성 함양 교육을 더 이수할 필요가 있다. 물론 국어교사 양성 기관에서는 이에 필요한 프로그램을 개발하고 제공해야 한다.

이외의 예비국어교사들 16명은 내적합 지수가 모두 0.7~1.3의 범위 내에 속하여 일관성을 적합한 수준으로 유지한 것으로 분석되었다. 이들의 논설문 평가 결과만이 엄격성이 일관성 있게 유지되었다고 할 수 있다.

적합한 수준의 일관성을 보인 예비국어교사는 전체 32명 중 16명(50%)인데, 이러한 결과는 예상을 벗어나는 특징적인 현상이다. 예비국어교사들은 중학생 글을 평가해 본 경험이 없었음에도 불구하고, 엄격성의 일관성을 적합하게 유지하는 경우가 많았기 때문이다. 현직국어교사의 일관성을 분석한 선행 연구에 따르면, 중학생이 작성한 글을 평가한 현직국어교사 중에서 일관성을 적절하게 유지한 비율은 56% 정도였다(박영민 · 최숙기, 2010 ; 최숙기 · 박영민, 2011). 현직국어교사들은 학생 글을 평가해 본 경험이 많을 뿐만 아니라 국어교사 양성 교육 및 입직 시험 과정을 거친 교과 전문가들이었음에도 불구하고 적절한 수준의 일관성을 보인 평가자는 과반을 조금 상회할 뿐이었다. 이에 비할 때, 학생 글을 접해 본 경험도 없고 양성 과정을 모두 마치지도 않은 예비국어교사들의 50%가 엄격성의 일관성을 적절하게 유지하고 있다는 결과는 주목을 끌지 않을 수 없다. 이러한 결과를 통해 추론컨대, 학생 글을 평가해 본 경험이 반드시 일관성을 적절하게 유지하는 능력을 반드시 보장해 주는 것은 아니라고 할 수 있다.

부적합으로 판정된 예비국어교사는 4명(12.5%)이었는데, 이러한 결과도 주목할 만하다. 선행 연구에 따르면 현직국어교사의 경우에도 68명 중 9명, 즉 13.04%가 부적합 경향을 보였기 때문이다(박영민 · 최숙기, 2009). 평가 경험이 많은 현직국어교사와 경험이 없는 예비국어교사의 부적합 경향은 매우 유사한 수준을 보이고 있다. 이번 연구에 참여한 예비국어교사들이 예외적으로 현직국어교사의 평가 전문성 수준과 유사한 능력을 지니고 있다고 할 수도 있겠지만, 선행 연구에서 내린 결론처럼 현직국어교사의 쓰기 평가 전문성이 기대보다 낮았다고 볼 수도 있다. 이에 대해서는 후속 연구가 더 이루어져야 좀 더 명확하게 알 수 있을 것이다.

3. 성별에 따른 엄격성 및 일관성 분석

평가자로 참여한 예비국어교사들이 성별에 따라 중학생 논설문에 대한 평가가 차이가 있는지를 알아보기 위하여 <표 6>과 같이 분석하였다.

외국의 선행 연구에서는 평가자의 성별에 따라 평가 결과가 차이가 있다는 결과를 보고한 예가 발견되고 있으므로(Peterson & Kennedy, 2006 ; Peterson, 1998 ; Haswell & Haswell, 1995 ; Roulis, 1995 ; Barnes, 1990), 이러한 경향이 예비국어교사들에게서도 나타나는지를 알아볼 필요가 있다. 성별에 따라 평가 차이가 있다면 이를 통제하기 위한 방안도 필요하게 될 것이다.

<표 6> 성별에 따른 엄격성 및 일관성 분석 결과

평가자 성별	엄격성(logit)	SE	Infit MS	Outfit MS
남	0.00	.02	0.96	0.96
여	0.00	.02	1.03	1.03
$\overline{X}$	0.00	0.02	0.99	0.99
s	0.00	0.00	0.05	0.04
Adj SD= .00 분리도= .00 분리신뢰도= .00				

<표 6>에 제시된 것처럼, 예비국어교사의 성별에 따라 엄격성이 차이가 있는지를 분석한 결과, X^2=.00(p<.001), 분리신뢰도 R=.00으로서 엄격성의 유의한 차이는 발견할 수 없었다. 외국의 여러 선행 연구에서는 평가자의 성별에 따른 평가의 차이를 지속적으로 지적해 왔고 박영민・최숙기(2010)에서도 여교사가 남교사에 비해 더 엄격하게 채점하는 경향이 있다는 점을 지적하였지만, 이번에 조사된 예비국어교사의 중학생 논설문 평가에서는 이와 유사한 결과를 얻을 수 없었다. 박영민・최숙기(2009)에서는 국어교사의 성별에 따른 평가 차이가 통계적으로 유의하지 않음을 보고하였는데, 이 결과와는 상반된다.

연구 결과가 일치하지 않는다는 점에서 볼 때, 평가자의 성별에 따른 평가 결과의 차이에 대해서는 좀 더 반복적인 조사가 필요하하다. 예비국어사와 현직국어교사의 집단 차이가 일부 영향을 미쳤는지, 표집에 따른 오차가 일부 영향을 미쳤는지 등에 대해서 후속적인 연구가 이루어져야 한다. 믿을 만한 수준의 반복적인 경향이 발견되지 않는다면 이에 대해서는 보편적인 경향을 예단하기 어려울 수도 있다.

4. 평가 요인 활용의 분석

예비국어교사들에게 제시된 평가 기준표에는 중학생 논설문 평가에 적합하게 조정된 5가지의 평가 요인이 설정되어 있다. 그런데 실제적인 채점 과정에서 이 5가지 평가 요인이 어떻게 활용되었는지를 확인해 볼 필요가 있다. 특정한 평가 요인이 과도하게 많이 쓰였거나 적게 쓰였을 수 있기 때문이다. 이를 통해서 중학생 논설문을 평가할 때 보이는 예비국어교사의 특성을 알 수 있으며, 설정된 평가 요인의 적절성도 파악할 수 있다. 이에 대한 분석 결과는 <표 7>과 같다.

<표 7> 평가 요인의 활용 양상

평가 요인	엄격성(logit)	SE	Infit MS	Outfit MS
내용	-0.01	.03	0.95	0.95
조직	0.18	.03	0.98	0.97
표현	-0.90	.03	0.95	0.95
단어선택	-0.34	.03	1.01	1.01
형식·어법	0.26	.03	1.13	1.12
$\overline{X}$	0.00	0.03	1.00	1.00
s	0.24	0.00	0.07	0.07
Adj SD= .00 분리도= .00 분리신뢰도= .00				

<표 7>에 따르면 예비국어교사들은 5가지의 평가 요인 중에서 '형식 및 어법'을 가장 엄격하게 적용하였으며, '표현'을 가장 관대하게 적용하였다. logit으로 볼 때 각각 0.26, -0.90을 보였다. 예비국어교사들이 '형식 및 어법'을 엄격하게 적용하는 것은 예비국어교사들이 보이는 보편적인 특성으로 추정된다. 현직국어교사와 예비국어교사의 학생 글 평가 결과를 대조한 연구에서도 예비국어교사들은 현직국어교사와 달리 글의 형식적 요건에 해당하는 '형식 및 어법'의 점수를 낮게 주는 엄격한 경향을 보였다(박영민·최숙기, 2009).

'형식 및 어법'은 객관식 문항처럼 맞은 것과 그렇지 않은 것을 비교적 선명하게 구분할 수 있다는 특징이 있는데, 예비국어교사들의 평가 결과에는 이러한 특징이 여과 없이 투영된 것으로 보인다. 예비국어교사들은 대학의 양성 과정에서 맞춤법 등과 관련된 교육을 집중적으로 이수하고 있으므로 이와 관련된 평가 요소에 민감하게 반응했을 가능성도 있다.

또한, <표 7>를 통해 볼 때 예비국어교사들이 적용한 평가 요인은 모두 적합도가 0.7~1.3 사이에 분포하고 있는데, 이를 통해 각 평가 요인들은 일관성이 적절하게 유지되었다고 판단할 수 있다. 5가지 평가 요인의 일관성이 적절하게 유지되었다는 것은 예비국어교사들이 중학생 논설문을 평가할 때 이 평가 요인이 일관성을 유지하는 데 기여하였음을 뜻한다. 그러므로 이 5가지 평가 요인은 예비국어교사들에게 일관성이 있는 평가 요인으로 기능하고 있으며, 평가자로 참여한 예비국어교사들도 이 5가지 평가 요인을 일관성 있게 적용할 수 있었다고 볼 수 있다.

5. 평가 척도 활용의 분석

일반적으로 평가 기준표에는 각각의 평가 요인별로 일정한 점수의 평가 척도가 제시되어 있다. 이 연구에서 적용한 평가 기준표에는 5점의 평가 척도가 제시되었다. 평가자들이 글을 평가할 때 활용하는 평가 척도는 조금씩 차이가 있는데, 이를 분석해 보면 평가자들이 가지고 있는 쓰기 평가 국면의 특성을 효과적으로 파악할 수 있다. 즉, 예비국어교사들이 5가지의 평가 요인에 설정되어 있는 척도를 어떻게 사용하였는가를 분석함으로써 중학생 논설문 평가에서 드러나는 예비국어교사의 평가 특성을 파악할 수 있다. 예비국어교사의 평가 척도 활용을 분석한 결과는 <표 8>에 제시하였다.

<표 8> 평가 척도의 활용 양상

평가 척도	사용 빈도	사용 비율	Outfit MS	평균 측정값
1	670	10%	1.1	-0.99
2	1370	21%	0.9	-0.63
3	2492	39%	1.0	0.02
4	1235	19%	0.9	0.54
5	633	10%	1.0	0.95

<표 8>에 따르면 예비국어교사들은 평가 척도 중에서 '3점'을 가장 많이 사용하였으며, '2점'을 그 다음 순위로 많이 사용하였다. '3점'과 '2점'의 활용 비율은 60%에 해당하는 것이어서 중학생 논설문을 평가한 예비국어교사들은 중앙 집중의 경향을 벗어나지 못한 것으로 보인다. 가장 적게 사용된 평가 척도는 '5점'이었다. 즉, 예비국어교사들이 중학생 논설문을 평가할 때 '5점'을 주는 일은 상대적으로 매우 적었다는 뜻이다.

예비국어교사들이 사용한 각각의 점수 척도들은 모두 적합도 지수가 0.7~1.3 사이에 분포하고 있어 일관성이 적절한 수준으로 유지되었다. 일관성 수준이 기준치에서 벗어하는 척도가 없었다는 것은 이 점수 척도가 일관성을 유지하는 데 기여하였음을 의미한다고 할 수 있다.

<표 8>에서 중요하게 관찰해야 하는 것은 척도 등급이 높아짐에 따라 평균 측정값이 비례하여 상승하고 있는지의 여부이다. 척도 등급이 높아짐에 따라 logit으로 제시되는 평균 측정값이 증가할 때, 평가에서 사용한 척도가 올바로 기능하고 있다고 해석할 수 있기 때문이다. 그런데 <표 8>의 평균 측정값을 보면 평가 척도가 상위 수준으로 옮겨갈수록 이에 비례하여 상승하고 있음을 확인할 수 있다. 그러므로 예비국어교사들이 적용한 평가 척도는 올바르게 기능했다는 점을 알 수 있다.

그런데 예비국어교사들이 적용한 평가 척도가 올바르게 기능하기는 했지만, 평균 측정값의 logit를 살펴볼 때 각각의 척도가 동일한 간격을 유지하지는 못한 것으로 보인다. '1점'과 '2점' 사이는 0.36 logit, '2점'과 '3점' 사이는 0.65 logit, '3점'과 '4점' 사이는 0.52 logit, '4점'과 '5점' 사이는 0.41 logit이므로, 예비국어교사들이 활용한 평가 척도는 명목적으로는

동간이지만 실제적으로는 동간을 유지하지 못하고 있다.

'2점' 척도와 '3점' 척도 사이는 다른 구간과 달리 상대적으로 큰 0.65 logit을 보이고 있는데, 이를 통해 보건대 예비국어교사들은 '2점'과 '3점'을 변별할 때 인지적 부담이 더 컸을 것으로 추측된다. '2점' 척도와 '3점' 척도 사이의 logit이 상대적으로 크다는 것은 '2점'에서 '3점'으로 올리는 것이 '3점'에서 '4점'으로 올리는 것이나 '1점'에서 '2점'으로 올리는 것보다 상대적으로 더 어려웠다는 뜻이다. 그러므로 예비국어교사들은 '3점'에 집중되는 경향을 보였음에도 불구하고, 중학생 논설문이 웬만해서는 3점을 잘 주지 않는 경향을 동시에 보이고 있다고 할 수 있다.

Ⅴ. 결론

이 연구에서는 예비국어교사 32명을 평가자로 위촉하여 중학생이 작성한 논설문을 평가하게 하고, 예비국어교사들이 보이는 평가의 엄격성 및 일관성을 분석하였다. 예비국어교사들이 보이는 쓰기 평가의 엄격성 및 일관성에 대한 분석 결과는 국어교사의 쓰기 평가 전문성 신장을 위하여 무엇을 어떻게 해야 할지에 대한 정보를 제공해 준다는 점에서 의의를 발견할 수 있다.

이 연구에서 얻은 결과는 다음과 같다.

첫째, 엄격한 평가자와 관대한 평가자는 logit 값 0을 기준으로 하여 각각 12명씩 할당되었으며, 일관성이 적합한 평가자가 32명 중 16명(50%), 부적합을 보인 평가자가 32명 중 4명(12.5%), 과적합을 보인 평가자가 32명 중 12명(37.5%)이었다. 이러한 적합-부적합-과적합 비율은 현직국어교사를 연구한 선행 연구에서도 유사하게 반복된다(박영민 · 최숙기, 2009). 예비국어교사들이 현직국어교사와 유사한 수준으로 일관성을 유지했다는 것은 이 연구에 참여한 예비국어교사들이 중학생 글을 평가하는 능력이 우수했음을 뜻한다고 해석할 수 있지만, 현직국어교사의 평가 능력이 예

비국어교사들보다 더 나은 것으로 보기 어렵다는 뜻으로도 해석할 수 있다. 그러므로 현직국어교사의 쓰기 평가 전문성을 높이기 위한 방안이 요구된다.

둘째, 예비국어교사들의 성별에 따라 엄격성이 다른지를 분석하였는데, 통계적으로 유의한 차이는 발견되지 않았다. Peterson & Kennedy(2006), Peterson(1998) 등을 비롯한 외국의 여러 선행 연구, 박영민·최숙기(2009) 등에서는 평가자 성별에 따른 차이를 지적해 왔지만, 이 연구의 분석에서는 차이가 발견되지 않았다. 이러한 결과를 지지하는 다른 선행 연구도 부분적으로 존재한다.

셋째, 예비국어교사들을 성별로 집단을 구분하였을 때, 일관성은 적합한 수준을 유지하는 것으로 분석되었다. 남자 예비국어교사들은 내적합지수가 0.96, 여자 예비국어교사들은 1.03을 보였다. 이러한 결과는 예비국어교사 각각의 평가 결과를 성별 평균으로 처리한 데에서 비롯된 것으로 보인다. 성별 평균은 예비국어교사 각각의 평가 특성을 소거하여 얻은 결과이므로 일관성이 적합한 수준에 분포하도록 하는 데 크게 기여한다.

넷째, 예비국어교사들이 중학생 논설문을 평가할 때 활용한 평가 요인을 분석한 결과, '형식 및 어법'에서 가장 엄격한 것으로 나타났다. 가장 관대한 요인은 '표현'이었다. '형식 및 어법'은 올바른지의 여부가 객관적으로 드러난다는 점, 예비국어교사들은 양성 과정에서 맞춤법과 관련된 학습을 많이 하고 있다는 점 등의 영향으로 인해 이 요인이 가장 엄격하게 평가되었던 것으로 보인다.

다섯째, 예비국어교사들의 평가 척도 활용을 분석하였는데, '2점' 및 '3점'을 준 비율이 60%로서 중앙 집중의 경향을 보인다는 사실을 확인하였다. 뿐만 아니라, 명목적으로는 동간으로 설정된 척도를 실제적으로는 동간으로 평가하지 않았는데, 특히 '2점'과 '3점' 사이의 간격이 0.65 logit으로 가장 컸다. 이를 통해서 예비국어교사들이 중학생 논설문을 평가할 때 평가 척도를 동간으로 유지하지 않았다는 점, '2점'과 '3점'을 변별하는 데 인지적 부담이 컸다는 점을 추론해 낼 수 있다.

이러한 결과를 통해 볼 때, 현직국어교사들의 쓰기 평가 전문성을 높

이기 위한 연수 교육이 강화되어야 하겠지만, 예비국어교사들의 평가 전문성을 높이기 위한 교육도 병행해서 이루어져야 할 것으로 보인다. 일관성 수준을 적합하게 유지한 예비국어교사가 50%였지만, 과적합 내지 부적합을 보임으로써 평가 전문성이 부족한 예비국어교사들도 50%에 달하기 때문이다. 예비국어교사들은 현직국어교사로 입직할 후보자들이므로 이들의 쓰기 평가 전문성을 기르는 일은 곧 학교 쓰기 평가의 전문성을 기르는 일로 이어진다. 국어교사 양성 대학에서는 예비국어교사의 쓰기 평가 전문성을 기르기 위한 방안을 모색해야 할 것이다.

또한, 예비국어교사들이 평가 척도를 동간으로 유지할 수 있는 교육과 훈련을 병행해서 시행해야 할 것으로 판단된다. 사실, 학생의 글을 채점하는 평가자의 인식적 판단으로는 동간 척도를 유지하는 것이 매우 어렵다. 평가의 국면에 따라, 평가의 상황에 따라 평가의 준거, 기준 등이 미세하게 진동하기 때문이다. 평가자가 기계가 아닐진대 이러한 현상은 매우 자연스러운 것이다. 현실적으로도 어지간해서는 '5점'을 잘 주지 않는 평가자가 있고, 웬만해서는 '1점'을 잘 부여하지 않는 평가자도 존재한다. 그러나 학생의 쓰기 능력을 적절하게 변별하고 쓰기 능력을 타당하게 추정하기 위해서는, 그리고 수량으로 얻어낸 평가 결과를 통계적으로 분석하기 위해서는 평가 척도를 동간으로 유지하는 능력을 좀 더 높은 수준으로 끌어올리지 않으면 안 된다.*

* 본 논문은 2012. 2. 29. 투고되었으며, 2012. 3. 9. 심사가 시작되어 2012. 3. 31. 심사가 종료되었음.

▣ 참고문헌

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(2010), "Rasch 모형을 활용한 국어교사의 쓰기 평가 특성 분석", 『국어교육학연구』, 37, 367-391.

박종임 · 박영민(2011), "Rasch 모형을 활용한 국어교사의 채점 일관성 변화 양상 및 원인 분석", 『우리어문연구』, 39, 301-335.

박종임 · 박영민(2012), "평가자 일관성에 따른 설명문 평가 예시문 선정의 차이 연구", 『작문연구』, 14, 301-338.

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.

최숙기 · 박영민(2011), "논설문 평가에 나타난 국어교사의 평가 특성 및 편향 분석", 『교육과정평가연구』, 14(1), 201-228.

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.

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

Barritt, L., Stock, P. L., & Clark, F. (1986). Researching practice : evaluating student essays. *College Composition and Communication*, 7, 315-327.

Bond, T. G. & Fox, C. M.(2007), Applying the Rasch Model : Fundamental Measurement in *the Human Sciences*(2nd ed.), NJ : Lawrence Erlbaum Associates, Inc.

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.

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

Cronbach, L. J.(1990), *Essential of Psychological Testing*(5th ed.), New York : Harper Collins.

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 Enterance Examination Board.

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., 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.

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.

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.

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.

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.

Lumley, T. & McNamara, T. F.(1995), Rater characteristics and rater bias : Implications for training, *Language Testing*, 12(1), 54-7.

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. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, 13, 425-444.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1994). Inter judge reliability and decision reproducibility. *Educational and Psychological Measurement*, 54, 913-925.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1996). The invariance of judge severity calibrations. In M. R. Wilson & G. Engelhard, Jr. (Eds.), *Objective Measurement Theory into Practice*, 3, 99-112. Norwood, NJ : Ablex.

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

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

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. 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.

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).

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

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.

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.

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

<초록>

예비국어교사의 중학생 논설문 평가에서 발견되는 엄격성 및 일관성의 특성

박영민

이 연구에서는 예비국어교사 32명을 중학생 논설문 평가자로 삼아 엄격성 및 일관성을 분석하였다. 이 연구에서 얻은 결과는 다음과 같다. 첫째, 엄격한 평가자와 관대한 평가자는 각각 16명씩이었으며, 일관성이 적합한 평가자가 32명 중 16명(50%), 부적합을 보인 평가자가 4명(12.5%), 과적합을 보인 평가자가 12명(37.5%)이었다. 둘째, 예비국어교사들의 성별에 따른 엄격성은 통계적으로 유의한 차이가 발견되지 않았다. 셋째, 예비국어교사들을 성별로 집단을 구분하였을 때, 일관성은 적합한 수준을 유지하는 것으로 분석되었다. 남자 예비국어교사들은 내적합 지수가 0.96, 여자 예비국어교사들은 1.03을 보였다. 넷째, 예비국어교사들이 중학생 논설문을 평가할 때 활용한 평가 요인을 분석한 결과, '형식 및 어법'에서 가장 엄격한 것으로 나타났다. 가장 관대한 평가 요인은 '표현'이었다. 다섯째, 예비국어교사들의 평가 척도 활용을 분석하였는데, '2점' 및 '3점'을 준 비율이 60%였으며, 명목적으로는 동간인 평가 척도를 실제적으로는 동간으로 평가하지 않았다. 이러한 결과를 통해 볼 때, 예비국어교사들의 평가 전문성을 높이기 위한 교육도 병행해서 이루어져야 할 것으로 판단된다.

【핵심어】 예비국어교사, 쓰기 평가, 엄격성, 일관성, Rasch 분석

<Abstract>

The Feature Analysis of Pre-service Korean Language Teacher's Scoring Severity and Consistence in Assessing Persuasive Writing

Park, Young-min

The aims of this paper is to analyze the feature of severity and consistence of 32 raters as pre-service Korean language teachers assessing persuasive wiring. The results are following. First, 16 raters are took into severe range and the other 16 raters to lenient range. Second, there aren't differences of raters severity according to their genders. Third, each gender group are showed consistency range. Forth, raters grade 'form and grammar' as the highest and grade 'expression' as the lowest. Fifth, as analyzed using of assessment scale, raters of 60% assigned to 2 and 3 points, so these show the central tendency. This result will be helper to develop programs for pre-service Korean language teacher to raise the assessment expertise.

【Key words】 pre-service Korean language teacher, writing assessment, rating severity, rating consistence, Rasch measurement model