교육과정평가연구
The Journal of Curriculum and Evaluation
2022, Vol. 25, No. 3, pp. 149~171
DOI: https://doi.org/10.29221/jce.2022.25.3.149

# 국어교사들의 쓰기 평가에서 나타난 엄격성 및 일관성 차이 분석: 논술 평가를 중심으로

정다운 (한국교원대학교 박사과정)*

**요약**

이 연구는 현직 국어 교사 52명을 논술 평가자로 위촉하여 채점하게 한 후 나타난 결과를 바탕으로 엄격성 및 일관성을 분석하는 데 목적이 있다. 이 연구에서 확인된 결과는 다음과 같다. 첫째, 국어 교사들의 엄격성과 일관성의 양상을 분석한 결과, 적합한 평가자 22명(42.31%), 부적합 평가자 11명(21.15%), 과적합 평가자 19명(36.54%)로 나타났다. 둘째, 국어 교사들의 엄격성을 분석한 결과, 평가자의 특성을 반영한 국면에 따라 차이를 보였다. 구체적으로 성별, 경력, 평가 요인에 따른 엄격성은 통계적으로 유의한 차이를 보였으나 평가 방법에 따른 엄격성은 통계적으로 유의한 차이가 확인되지 않았다. 성별에 따른 엄격성은 여교사가 남교사보다 엄격하게 평가하는 것으로 나타났다. 경력에 따른 엄격성은 1년 이상 5년 미만의 경력을 가진 집단이 가장 엄격하게 평가하는 것으로 확인되었다. 평가 요인에서는 형식 및 어법 요인에서 가장 엄격하게 평가하는 것으로 나타났다. 반면 평가 방법에 따른 엄격성은 오류 분석형, 사고논술형, 키워드 제시형의 평가 방법을 사용하는 집단이 동일한 엄격성으로 무표형의 평가 방법을 사용하는 집단에 비해 엄격하게 평가하는 것으로 확인되었다. 셋째, 국어 교사들의 일관성을 분석한 결과, 판별 기준에 따라 차이가 나타났다. 내적합 지수를 기준으로 할 때 성별, 경력, 평가 방법, 평가 요인의 모든 국면에서 적합한 수준을 유지하는 것으로 나타났으나 내적합 표준화 값을 기준으로 할 때 국면 및 하위 요인에 따라 다양한 일관성의 양상이 나타났다. 이러한 결과를 바탕으로 할 때 논술 평가에서도 국어 교사들의 쓰기 평가 전문성을 높이는 방안이 필요할 것으로 보인다.

주제어 : 국어 교사, 쓰기 평가, 논술 평가, 엄격성, 일관성, 다국면 Rasch 모형

* 제1저자 및 교신저자, immanuel123@naver.com

# I. 서 론

현대 사회에서 언어 평가는 학습자가 실제로 얼마나 잘 구사하는지를 평가하는 직접 평가 방식으로 변화하고 있다(McNamara, 1996). 실제 학습자의 쓰기 수행 활동이 포함된 쓰기 결과물은 수행 평가라고 간주할 수 있다(Weigle, 2002).

수행 평가 방식에서는 평가 척도, 평가자, 과제 유형, 쓰기 과제의 지시문, 쓰기 텍스트 등의 많은 요인들의 피험자의 점수에 복합적으로 영향을 미친다(Bachman, 1990; Hamp-Lyons, 1990; Kroll, 1998; McNamara, 1996; Weigle, 2002). 직접 평가 방식의 쓰기 평가에서는 평가자의 내·외적 특성이 평가 결과에 크게 영향을 미치기 때문에 평가 결과에 대한 신뢰성에 대한 문제가 제기되고 있다. 평가 결과에 대한 신뢰성이 확보되지 않으면 피험자의 능력을 정확하게 읽어내지 못하기 때문에 본래 평가가 가진 목적으로 활용하기 어렵다.

이러한 점에서 쓰기 평가의 신뢰성을 높이기 위한 방안으로 쓰기 평가 전문성이 제시되고 있다. 쓰기 평가 전문성은 학생 글에 담긴 수준을 일관되게 읽어 내고 적합한 점수 척도를 변별하여 부여하는 능력을 말한다. 구체적으로 학교 현장의 쓰기 평가에서 국어 교사에게 필요한 쓰기 평가 전문성은 평가 요인이나 수준에 따라 글을 변별하는 능력과 적절한 평가 척도를 사용하는 능력, 외부 요인이나 내부 요인을 적절히 통제하여 글에 나타난 작문 능력을 정확하게 읽어 내는 것을 의미한다. 쓰기 평가 과정에서 국어 교사에게 요구되는 평가 전문성은 '쓰기 평가 수행'의 관점, 알아야 할 '지식'의 관점, 자세와 관련된 '태도'의 관점으로 구분한다(박영민, 2012).

쓰기 평가 전문성에서 중요하게 지적되는 것이 엄격성(rating severity)과 일관성(rating consistence)이다. 쓰기 평가에서 일관성은 평가에서 엄격한 정도를 일관되게 유지하는 것을 의미한다는 점에서 엄격성의 일관성이라고도 한다(박영민, 2012). 평가 과정에서 유사한 쓰기 능력을 가진 학생에게 동일한 엄격성을 유지하여 점수를 부여할 수 있는 능력이라고 볼 수 있기 때문이다. 일관성을 유지하는 것은 평가 결과의 신뢰도와 연계된다는 점에서 중요하게 언급된다.

한 평가자가 채점을 시작해서 끝날 때까지 엄격성과 일관성을 유지해야만 유사한 쓰기 능력을 가진 학생 글에 대해 유사한 점수를 부여할 수 있다. 평가 과정에서 엄격성과 일관성이 변동이 된다면 유사한 쓰기 능력을 가진 학생 글이 어느 시기에 채점이 되느냐에 따라 점수 차이가 나타날 수 있다. 그러나 국어 교사가 쓰기 평가를 하는 동안 다양한 영향 요인의 개입되기 때문에 평가를 시작해서 끝날 때까지 엄격성과 일관성을 적절하게 유지한다는 것은 쉽지 않다(박종임·박영민, 2011). 엄격성이나 일관성은 평가 경험이 많고 평가자 교육을 많이 받았다고 해서 엄격성이나 일관성이 더 잘 유지하는 것으로 볼 수 없다. 이는 평가와 관련된 양적인 조치만으로 자연스럽게 엄격성이나 일관성이 변화하는 것이 아니며 쓰기 평가 과정에 평가자에게 영향을 미치는 요인들이 많기 때문에 엄격성과 일관성을 유지하기 위한 다양한 조치들이 필요하다.

이러한 상황을 고려하여 이 연구는 논술 평가를 중심으로 국어 교사들의 쓰기 평가에서 나타난 엄격성과 일관성의 차이를 분석하는 데 목적을 두었다. 이를 위해 현직 국어 교사들의 성별, 교육 경력, 평

가 방법, 평가 요인을 중심으로 엄격성과 일관성이 어떠한 차이를 나타내는지 파악하였다. 이를 통해 현직 국어 교사들의 엄격성 및 일관성에 대한 기초 자료를 제공하고자 한다.

성별에 따른 평가 결과에서 현직 교사는 남교사에 비해 여교사가 학생이 쓴 글에 많은 논평을 남기는 것으로 나타났으나 차이가 유의하지 않았다(Barnes, 1990; Roulis, 1995; Peterson & Kennedy, 2006). 한편 예비교사는 남교사와 여교사의 성별에 따른 평가 결과의 차이가 없는 것으로 나타났으나(박영민, 2012a), 현직 교사는 여교사가 남교사에 비해 엄격하게 평가한 것으로 나타났다(최숙기·박영민, 2009, 2010a, 2010b, 2011).

교육 경력에 따른 평가 결과에서 1년~5년 경력을 가진 국어 교사가 가장 엄격하게 평가를 하고 11년~20년 경력을 가진 집단, 6년~10년 경력을 가진 집단, 20년 초과의 순으로 나타났다(최숙기·박영민, 2011). 경력에 따른 평균과 표준 편차의 차이가 확인되었으며 6~10년 경력을 가진 집단에서 가장 높은 점수를 부여한 것으로 나타났다(박영민·최숙기, 2009).

학교급에 따른 평가 결과에서 중학교가 고등학교 교사에 비해 적합한 평가자의 비율이 높은 것으로 확인되었다. 이는 중학교 교사가 평가 기회가 더 많은 것과 관련된다고 밝혀지면서 학교급보다는 평가 경험이 평가 결과가 연계되는 것으로 나타났다(장은주, 2015).

따라서 성별, 교육 경력, 학교급 요인에 따른 평가 결과가 달라진다는 것이 확인되었다. 이러한 점에서 논술 평가 결과도 성별, 교육 경력, 학교급에 따른 차이를 구체적으로 확인할 필요가 있다. 이러한 연구의 목적을 달성하기 위해 이 연구에서는 현직 국어 교사 52명을 대상으로 고등학교 1학년 논술 텍스트[1]30편에 대한 평가 결과를 수집하여 분석하였다.

이 연구는 현직 국어 교사 52명만을 대상으로 수집된 결과를 바탕으로 분석되었다는 점과 평가지에 나타난 표지만을 근거로 교사의 평가 방법을 파악했다는 점에서 일반화하기에는 한계가 있다. 향후 보다 많은 표집 집단을 형성하고 평가의 영향 요인들의 개입을 최대한 통제한 후 결과를 확인하는 후속 연구와 면담, 관찰 등을 통해 교사들의 구체적인 평가 방법을 파악하여 결과를 확인하는 후속 연구가 추가적으로 필요할 것이다.

그럼에도 불구하고 이 연구는 지금까지 쓰기 평가에서 주로 논의된 문종에서 나아가 논술이라는 문종을 특화하고 세부적인 관점에서 평가 결과를 분석하였다는 점에서  차별성을 가진다. 지은림(2008)에서도 논술고사를 대상으로 채점자의 특성에 대해 논의하였다. 지은림(2008)에서는 대입 선발을 목적으로 하는 대단위 평가로 대학 교수들을 평가자를 전공에 따라 전문가와 비전문가 집단을 평가자로 설정하여 채점하게 한 후 거시적인 관점에서 평가 결과를 비교·분석하였다. 그러나 이 연구에서는 중등학교 학교 교육 현장에서 수업의 일환으로 논술을 적용하고 현직 국어교사들을 평가자로 설정하여 채점하게 한 후 평가 결과에 대한 세부적인 관점에서 분석하고 평가지 표지에 나타난 평가 방법까지 분석하였다. 현재 고등학교 작문교육에서 논증적 글쓰기가 학습 내용으로 포함되어 있고 범교과적으로는 논술형 평가와 함께 논술의 중요성이 더욱 커짐에 따라 논술이라는 문종을 적용하여 현직 국어교사의 엄격성 및 일관성 차이를 논의한 이 연구의 결과는 중요한 기초 자료가 될 것으로 판단된다.

---

1) 여기서 논술 텍스트는 다양한 수준의 논술 능력을 가진 학생들을 대상으로 한다는 점에서 논술 글이라고 하기에도 적합하지 않고 논술 문장이라고 하기에도 적합하지 않다는 점에서 논술 텍스트라는 말을 적용하였다.

# II. 이론적 배경

## 1. 평가자의 엄격성

쓰기 평가에서 평가의 오류나 평가의 신뢰성과 관련된 부분이 지속적으로 논의되고 있다. 언어 수행에 대한 평가가 이루어지는 쓰기 평가는 엄격성이 평가 결과에 영향을 미치는 요인으로 지적되고 있다(Engelhard, 1994; Engelhard & Myford, 2003; Lumley & McNamara, 1995). 서로 다른 엄격성을 가진 평가자가 부여한 점수는 동일한 피험자의 능력에 대해 다른 평가 결과를 초래하는 요인이 되기 때문에 엄격성을 평가 결과의 오류를 유발하는 요인으로 보기도 한다(박영민, 2012). 평가자의 엄격성에 의해 발생하는 효과는 가장 민감하면서도 심각한 오류라는 점에서 평가자의 엄격성 효과(severity effect)라고 불린다(Cronbach, 1990).

평가자의 엄격성은 점수를 부여하는 경향이 엄격한지 관대한지를 의미한다(최숙기·박영민, 2011). 평가자의 엄격성은 평가 척도 중간 이하의 점수를 부여하는 엄격성(severity)과 평가 척도 중간 이상의 점수를 부여하는 관대성(leniency)으로 분류된다(Kneeland, 1929; Ford, 1931). 엄격성이 높은 평가자는 피험자의 능력을 과소추정하며 관대성이 높은 평가자는 피험자의 능력을 과대 추정하는 결과를 초래한다(설현수, 2010). 예를 들면, 쓰기 능력이 유사한 피험자가 있을 때 엄격한 평가자는 평균 척도 이하의 낮은 점수를, 관대한 평가자는 평균 척도 이상의 높은 점수를 부여함으로써 피험자의 능력을 완전히 다르게 평가를 하게 되는 것이다.

관대한 평가자가 평가 기준에 관계없이 일관되게 관대한 평가를 하는 것은 평가자 개인의 고유하고 안정적인 인성과 관련된 것이라고 주장하기도 한다(Guilford, 1954). 평가자는 평가자가 내부에 가지고 있는 주관적인 인식을 통해 자신의 고유한 전략과 기준을 적용하여 평가한다. 그러나 평가자마다 내면적으로 가진 평가 기준이나 평가 척도에 대한 인식이 다르기 때문에 평가자의 엄격성에 차이가 나타나게 된다. 이러한 점을 해결하기 위해 평가자의 엄격성을 평가자 훈련이나 협의 등을 통해 조정하거나 기준이나 척도를 비슷한 수준으로 맞추려는 노력을 한다. 그러나 평가자의 엄격성은 완전히 없어지지 않는 것으로 제시된다(박영민, 2012). 이러한 점에서 엄격성의 차이를 인정하는 가운데 쓰기 평가를 수행할 필요가 있을 것이다.

지금까지 많은 선행 연구들은 FACETS 프로그램에 다국면 Rasch 모형을 적용하여 엄격성을 측정하였다(박찬흥, 2018;  박영민, 2012a; 박영민·최숙기, 2010b). 그러나 설현수(2010)에서 단순 총점, 표준점수, FACETS 점수를 통해 엄격성의 차이를 분석한 연구 결과를 보면, 평가자의 엄격성 정도와 관계없이 일정하게 일치를 보이는 것으로 확인되었다. 즉 평가자의 엄격성을 조정하기 위해 피험자의 능력 값을 산출하거나 표준화 공식을 통해 피험자의 능력 점수를 재산출하거나 엄격성을 모형화해서 피험자의 능력 점수를 산출하는 방법이 활용되고 있으나 이를 통해 엄격성이 통제되지 않는 것으로 밝혀졌다. 다국면 Rasch 모형은 평가자 간의 엄격성을 조정한 상태에서 능력 추정치를 확인할 수 있는 유일한 모형이라고 밝히고 있다.

## 2. 평가자의 일관성

평가자의 주관성이 평가 결과에 영향을 미칠 수 있다는 점에서 평가 결과의 신뢰성과 함께 제기되는 문제가 일관성이다. 한 명의 평가자가 한 집단 전체 학생에 대해, 전체 평가자는 개별 학생에 대해 일관성을 유지할 필요가 있다. 그러나 일관성은 평가 내용의 범위, 문항의 수, 문항 난이도, 평가자의 수, 평가자 훈련, 평가에 소요되는 시간, 평가 방법 등의 다양한 요인의 영향을 받기 때문에(송영미 외, 2009) 평가자 간 일관성을 유지하는 것이 쉽지 않다(김경희·송미영, 2001).

평가자의 일관성은 평가자가 부여하는 점수가 일정한 경향을 보이며 유지되는 현상을 말한다(최숙기·박영민, 2011). 쓰기 평가에서 일관성은 평가자가 피험자의 능력을 일관되게 일어내는 능력과 관련되어 있다는 점에서 평가자의 내적 신뢰도로 언급된다. 예를 들면, 일관성을 유지하는 적합한 평가자는 평가가 시작할 때부터 끝날 때까지 유사한 수준의 피험자에게 동일한 점수 척도를 부여한다. 그러나 일관성의 변동이 심하거나 일관되지 못하는 부적합, 과적합 평가자는 쓰기 능력에 비해 과도하게 낮은 점수나 과도하게 높은 점수를 부여하는 등 쓰기 능력을 정확하게 읽어내지 못하기 때문에 평가 결과에 대한 신뢰성에서 문제가 될 수 있다. 이에 따라 연수, 상세한 채점 기준의 마련, 전문성 신장 등을 통해 일관성 확보를 위해 노력하고 있다.

일관성은 평가자의 내적 신뢰도 이전에 평가자가 부여한 점수의 경향을 파악할 수 있는 지표이다. 쓰기 평가에서 평가자의 일관성의 차이가 나타나는 것은 많은 영향 요인의 개입에 따른 것이다. 그 중에서도 평가자가 내면에 인식하고 있는 평가 기준이나 척도가 변화하는 것이 평가의 일관성의 차이를 초래하는 요인이 된다(박종임·박영민, 2011; 최숙기·박영민, 2011).

다국면 Rasch 모형은 평가자의 일관성에 관한 정보를 제공한다(안수현·김정숙, 2017). 다국면 Rasch 모형에서 내적합 지수, 외적합 지수, 내적합 표준화 값, 외적합 표준화 값을 통해 일관성 유형을 적합, 과적합, 부적합으로 분류할 수 있는 수치들을 제공한다. 평가자가 일관성을 유지되면 적합, 평가자가 일관성이 매우 높으면 과적합, 평가자가 일관성이 매우 낮거나 일관성이 없으면 부적합으로 분류한다. 일관성을 분류하는 학자들마다 다양하게 제기되고 있으나 측정된 값을 통해 해석되는 결과는 적합, 과적합, 부적합으로 공통점을 가진다. 일반적으로 부적합이 과적합보다 평가의 신뢰도 부분에서 더 문제가 될 수 있다고 지적된다(Myford & Wolfe, 2003).

쓰기 평가에서 일관성 지수는 신뢰성 판단의 중요한 요인이 된다(Linacre, 1994; 지은림, 2008). 적합한 평가자는 엄격성을 일정하게 유지하며 평가를 진행하였기 때문에 신뢰할 수 있지만 과적합이나 부적합한 평가자는 엄격성을 일정하게 유지하지 못하고 채점하였다는 점에서 신뢰하지 못한다고 보는 것이다(장미·박영민, 2021).

평가자의 엄격성과 일관성은 평가자들의 주관이 반영되고 모든 평가자의 엄격성과 일관성을 같은 수준으로 유지할 수 없다는 한계가 있다. 이에 따라 평가자의 신뢰도에 관한 연구들은 엄격성과 일관성의 우연에 따른 일치보다는 신뢰성이 높게 산출된 결과를 보고하고 있다(Kortez, 1993). 따라서 쓰기 평가에서도 개별 평가자의 엄격성과 일관성을 적정한 수준에서 인정하고 최적의 신뢰도를 나타낸 결과 값을 통해 평가 결과를 해석할 필요가 있을 것이다.

# III. 연구 방법

## 1. 연구 대상

쓰기 평가에서 나타난 엄격성 및 일관성의 차이를 분석하기 위해 현직 국어 교사 52명을 대상으로 쓰기 평가 및 조사를 실시하였다. 연구 대상을 구체화하면 〈표 1〉과 같다.

〈표 1〉 연구 대상의 현황

| 구분 | | 성별 | | 계 |
|---|---|---|---|---|
| | | 남 | 여 | |
| 학교급 | 중학교 | 3 | 21 | 24 |
| | 고등학교 | 13 | 15 | 28 |
| 교육 경력 | 초임(1년 미만) | 1 | 1 | 2 |
| | 1년 이상 ~ 5년 미만 | 4 | 8 | 12 |
| | 5년 이상 ~ 10년 미만 | 0 | 3 | 3 |
| | 10년 이상 ~ 15년 미만 | 2 | 3 | 5 |
| | 15년 이상 ~ 20년 미만 | 1 | 8 | 9 |
| | 20년 이상 | 8 | 13 | 21 |
| 합계 | | 16(30.77) | 36(69.23) | 52 |

〈표 1〉의 연구 대상을 구체적으로 살펴보면 남교사 16명(30.77%), 여교사 36명(69.23%)를 차지하고 있는 것으로 나타났다. 상대적으로 여교사의 비율이 남교사에 비해 높게 나타났다. 학교급에 따르면 중학교 교사가 24명(46.15%), 고등학교 교사가 28명(53.85%)으로 나타났다. 교육 경력별로 구분하면 초임(1년 미만)이 2명(3.85%), 1년 이상 5년 미만이 12명(23.08%), 5년 이상 10년 미만이 3명(5.77%), 10년 이상 15년 미만이 5명(9.62%), 15년 이상 20년 미만이 9명(17.31%), 20년 이상이 21명(40.39%)으로 나타났다.

이상의 내용을 종합해 볼 때 위촉된 국어 교사들은 여교사, 고등학교, 교육 경력이 20년 이상인 교사의 비율이 높게 나타났다. 이러한 특성을 종합하여 엄격성 및 일관성 분석에 적용하였다.

## 2. 검사 도구

이 연구에서 평가 자료는 고등학교 1학년 학생이 작성한 논술 텍스트이다. 논술 텍스트는 '3가지 도움 자료를 읽고 세계의 환경 문제 중 하나를 선정하여 제시하고 문제점을 해결하기 위한 방안을 서술하시오.'라는 쓰기 과제와 3가지 자료 텍스트에 의해 작성된 결과물이다. 〈 자료 1〉은 세계 위험보고서 2020에 제시된 '2020. 위험의 발생가능성 Top과 파급력 Top 10'이라는 보고서이다. 〈자료 2〉는 리서치페이퍼에 제시된 '세계 최대 위기는 기상이변'라는 신문기사이다. 〈자료 3〉은 기후변화행동연구

소와 UN General Assmbly(2015)에 제시된 '2020년 주요 국제 환경 사안과 지속가능발전목표'에 대한 분석 자료이다. 다양한 자료 텍스트로 구성한 것은 2015 개정 국어과 교육과정에서 필자가 논증에 적절한 근거를 제시하도록 다양한 자료 텍스트를 제공하도록 서술하고 있기 때문이다(권영민 외 22인, 2018).

평가 자료는 쓰기 과제와 3가지 자료 텍스트에 따라 148명의 고등학생들이 작성한 논술 텍스트 중에서 작성된 분량을 기준으로 층화임의추출 방법에 의해 500자 이상이 되는 집단에서 30편을 임의 추출하여 구성하였다. 평가 자료는 2020년 2학기 1차 고사 직후 5일간 수업시간에 정기고사의 평가 목적이 아닌 세부능력 특기사항 작성을 위한 일환으로 진행되었으며 학생들에게는 1000자의 원고지 양식을 제공하였다. 이에 따라 원고지 기준 500자 이상 작성한 집단과 500자 미만의 집단으로 층을 나누어 500자 이상의 집단에서 30편을 임의추출하였다. 학생들이 작성한 분량을 기준으로 추출한 이유는 평가자들에게 분량이 평가 과정에서  후광효과로 작용하는 것을 사전에 막기 위한 조치를 하였다.

이 연구는 평가 방법에 따른 평가 특성을 분석한다는 점에서 평가자로 위촉된 국어 교사들에게는 완전 교차 설계 방식을 적용하였다. 또한 평가자로 위촉된 국어교사들에게는 평가 기준표 1부와 평가 자료 30편이 포함된 채점표가 제공되었다. 평가 기준표는 〈표 2〉와 같다.

〈표 2〉 고등학생 논술 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.<br>• 3점: 글의 중심 내용(주제)이 다소 명확하지 못하다. 글 전체의 중심 내용(주제)과 부합하지 않은 세부 내용이 들어 있다.<br>• 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.<br>• 3점: 중심 내용이 잘 드러나도록 조직되었지만, 내용의 순서나 구조가 독자가 이해하는 데 어려움이 따른다.<br>• 1점: 중심 내용이 잘 드러나도록 조직되지 않았을 뿐 더러, 내용의 순서나 구조가 독자가 이해하는 데 어려움이 따른다. |
| 표현<br>(어조 및 태도) | • 5점: 독창적이며 흥미롭게 표현되어 있으며, 독자가 쉽고 정확하게 이해할 수 있도록 표현되었다. 필자의 주체적인 목소리도 드러난다.<br>• 3점: 독자가 쉽게 이해할 수 있도록 표현되었으나, 독창성이나 흥미는 다소 떨어진다. 필자의 주체적인 목소리도 잘 드러나지 않는다.<br>• 1점: 내용만을 기계적으로 나열하여 글의 생동감이 떨어지며 흥미를 주지 못 한다. 독자가 쉽게 이해할 수 있도록 표현되지 않았다. |
| 단어 선택 | • 5점: 내용을 정확히, 흥미롭게, 자연스럽게 전달할 수 있는 단어가 선택되었다.<br>• 3점: 대체적으로 단어 선택이 내용 전달에 무리가 없으나 부적절한 단어들이 포함되어 있다.<br>• 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)<br>• 3점: 제한된 범위에서만 글의 표준적 형식이 확인된다.<br>• 1점: 철자, 구두점, 문법에서 잘못된 것이 많아 내용을 파악하며 읽는 것이 어렵다. |

이 기준표는 최숙기·박영민(2011)에서 제시된 것으로 Spandel & Culham(1996)의 평가 기준표를 우리나라 학생들의 평가에 적합하도록 수정한 것이다. 평가 요인은 내용, 조직, 표현(어조 및 태도), 단어 선택, 형식 및 어법이 포함되고 평가 척도는 5점, 3점, 1점만 평가 기준을 제시하고 2점과 4점 척도에 대해서는 자유롭게 평가하도록 하는 방식으로 설계되었다.

## 3. 연구 절차

이 연구에서 현직 국어 교사 52명에게 고등학생들이 작성한 논술 30편, 쓰기 과제 및 자료 텍스트 3편이 포함된 쓰기 과제 제시문, 평가 기준 및 채점표가 제공되었다. 평가자로 위촉된 현직 국어교사들에게 논술의 쓰기 과제 제시문 전체를 제공함으로써 학생들이 어떠한 과제에 따라 논술을 작성하였는지 파악할 수 있도록 하였다. 평가자로 위촉된 현직 국어 교사들에게 논술 쓰기 과제 제시문, 평가 기준표 및 채점표 등을 우편, 학교 및 학교 외의 장소 등을 통해 직·간접적인 방법을 통해 평가 과제를 전달하였다. 이 연구의 목적이 쓰기 평가의 특성을 분석하는 데 있는 만큼 평가자 훈련은 따로 실시하지 않았으나 1년 미만의 평가자들은 평가에 어려움을 겪을 수 있다는 점에서 모든 평가자들에게 평가자 훈련이 가능하다는 점을 안내하고 진행하였다.

평가자 훈련을 실시하지 않은 것은 평가자 훈련이 평가자가 본연에 가진 일관성의 변화에 영향을 미칠 수 있기 때문이다. 현직 국어교사들의 논술 평가 자료는 2021년 12월 7일부터 2022년 3월 13일 동안 수집하였다. 논술 평가 자료는 COVID-19의 상황이 지속되었기에 국어 교사의 상황이나 요청에 따라 직접 또는 간접의 전달 방법을 병행하였다. 현직 국어 교사들이 손으로 하는 종이 채점에 익숙하다는 점에서 종이 채점을 시행하여 평가 결과를 수집하였다. 수집한 논술 평가지에 나타난 표지를 바탕으로 평가자들의 평가 방법을 파악하고 코딩을 한 후 FACETS 프로그램을 활용하여 분석하였다.

## 4. 분석 방법

이 연구에서는 고등학생의 논술 평가에서 나타난 현직 국어교사들의 엄격성 및 일관성 차이를 분석하기 위해 평가자(교사), 피험자(학생), 성별, 평가 방법, 평가 요인의 5국면을 다국면 Rasch 모형을 통해 분석하였다. 구체적으로 제시하면 [그림 1]과 같다.

[그림 1]에 따르면 이 연구에서 다국면 Rasch 모형을 적용하면 성별이 c이고 경력이 d이고 평가 방법이 o인 평가자 r이 고등학생 s가 쓴 논술 텍스트의 평가 요인 e에 대한 평가 점수가 f-1이 아닌 f를 점수로 부여할 확률과 그 확률을 log로 변환한 값을 확인할 수 있다(Linacre, 1989). 각 국면의 분석 결과를 logit 척도로 변환한 값은 고등학생들의 논술 능력, 평가 요인의 엄격성의 값도 함께 산출할 수 있다. 구체적으로 살펴보면, 피험자 s가 텍스트 t에서 점수 $P_{sercdof(f-1)}$을 획득할 확률에 대한 점수 $P_{sercdof}$를 획득할 확률의 log로 변환한 값은 피험자 s의 능력($D_e$)에서 쓰기과제 e의 난이도($D_e$)를 빼고 평가자 r의 엄격성($C_r$)와 어떠한 부분 점수(f-1)에서 다음 높은 점수(f)로 올리는 데 걸리는 어려움($G_f$)을 뺀 것을 의미한다. 즉, 피험자의 능력, 쓰기과제의 난이도, 평가자의 엄격성 등 다양한 요인들을 국면으로 설정하여 평가자의 능력이나 쓰기과제의 난이도에 미치는 영향도 분석하여 통제할 수 있다.

$$\log( P_{sercdof} / P_{sercdof(f-1)} ) = E_s - D_e - C_r - U_c - T_d - S_o - G_f$$

$E_s$ : 고등학생(피험자) s의 논술 텍스트 점수
$D_e$ : 논술 텍스트 평가 요인 e의 난도(難度)
$C_r$ : 국어교사(평가자) r의 엄격성
$U_c$ : 성(性)이 c인 국어교사(평가자) r의 엄격성
$T_d$ : 경력이 d인 국어교사(평가자) r의 엄격성
$S_o$ : 평가 방법이 o인 국어교사(평가자) r의 엄격성
$G_f$ : 평가 척도 f-1에 대한 척도 f의 난도

$P_{sercdof}$ : 성별이 c이고 경력이 d이고 평가 방법이 o인 평가자 r이 텍스트 t에 대해 평가 요인 e에 근거하여 점수 f를 부여할 확률
$P_{sercdof(f-1)}$ : 성별이 c이고 경력이 d이고 평가 방법이 o인 평가자 r이 텍스트 t에 대해 평가 요인 e에 근거하여 점수 f-1을 부여할 확률

[그림 1] 논술 평가에 대한 다국면 Rasch 모형

엄격성은 원자료 데이터 7,800개(텍스트 30명×평가자 52명×평가요인 5개)를 코딩하고 성별, 교육 경력, 학교급, 평가 방법에 대한 데이터도 함께 코딩하여 FACETS 프로그램에 투입한다. 측정 단면 분포도에서 첫 번째 칸의 측정 척도(scale of measurement)에 나타난 Measr와 측정치(measurement report)에서 나타난 Measure은 logit 값을 의미한다. 이들 값에서 +값일수록 평가 점수의 엄격성을, -값일수록 평가 점수의 관대성을 의미한다. 이들 logit 값은 원자료 점수 데이터가 낮은 점수로 채점되었는지, 높은 점수로 채점되었는지를 나타낸다. +값에서 가장 큰 수로 나타나는 텍스트와 집단이 가장 엄격하게 채점한 것이며, -값에서 가장 작은 수로 나타나는 텍스트와 집단이 가장 관대하게 채점한 것으로 해석한다. 내적합 지수, 외적합 지수, 내적합 표준화 값과 외적합 표준화 값은 일관성을 판별하는 지표로, 유사한 쓰기 능력을 가진 평가자가 유사한 점수를 받았는지, 채점의 시작부터 끝까지 동일한 평가자에게 일관성 있는 평가 척도를 사용하고 쓰기 능력을 적절하게 변별할 수 있는지를 나타낸다.

현직 국어교사들의 쓰기 평가 특성은 FACETS Version 3.83.6 프로그램을 활용하여 쓰기 평가 결과를 분석함으로써 확인하였다. FACETS 프로그램에 의한 다국면 Rasch 모형은 언어 수행 평가 분석에서 주로 이용되며 적합, 과적합, 부적합과 같은 적합도, 분리신뢰도를 통해 평가자를 분석할 수 있다. 또한, 평가 결과에 대한 종합적인 분석 뿐만 아니라 피험자의 능력, 텍스트, 항목, 평가자 등의 다양한 국면에서 상세한 정보를 제공해 준다(Linacre, 1996, McNamara 1996, 이영식 1998). 이에 따라 다국면 Rasch 모형을 적용한 FACETS 프로그램 논술 평가의 세부적인 관점까지 분석하려는 이 연구에 목적에 부합한다고 판단하였다.

현직 국어 교사들의 쓰기 평가 특성은 엄격성과 일관성의 관계를 중심으로 평가 기준에 대한 모형 적합도를 확인하였다. 분석 모형에 대한 적합도는 일반적으로 통용되는 내적합 지수(Infit MnSq) 및 학자들 간의 의견 차이를 보이지 않는 내적합 표준화 값(Infit ZStd)에서 공통적으로 판별되는 적합도를 적용하였다. 이 연구에서 일관성 및 엄격성의 판별 기준으로 내적합 표준화 값(Infit ZStd)을 복합적으로 설정한 것은 내적합 지수(Infit MnSq)나 외적합 지수(Outfit MnSq)를 통한 적합도 기준이 학자들마다 다르기 때문이다. 예를 들면, 내적합 표준화 값은 +2 logit 이상일 때 부적합, -2 logit 이하

일 때 과적합, -2와 +2 사이의 범위에 있을 때는 적합으로 판별한다(장소영·신동일, 2009). 그러나 내적합 지수 및 외적합 지수는 0.6 이상 1.5 이하의 범위로 보기도 하고(Lunz et al., 1990), 0.6 이상 1.4 이하의 범위로 보기도 하며(Wright et al., 1994), 0.7 또는 0.75에서 1.3 이내의 좁은 범위를 주장하기도 한다(지은림 외 역, 2003; Bond & Fox, 2001). 지은림 외 역(2003), Bond & Fox(2001)의 주장에 따르면, 내적합 지수나 외적합 지수는 0.75보다 작을 때 과적합, 1.3보다 클 때 부적합으로 판별한다는 것이다. 정확하게 구분하기 위해 내적합 표준화 값(Infit ZStd)에 따른 판별과 Bond & Fox(2001)의 내적합 지수(Infit MnSq)에 따른 판별에서 적합도 양상이 다르게 나타나는 경우에는 이를 별도로 구분하여 제시하였다.

## 5. 분석 기준

이 연구에서 평가 방법을 분석하기 위해 평가자의 평가 결과지에 나타난 표지를 바탕으로 사고논술형(thinking essay), 키워드 제시형(keyword), 오류 분석형(error analys), 무표형(nonexistence)으로 구분하였다. 사고 구술은 인간의 심리적 기제를 밝혀 내기 위해 자신의 생각을 보고하는 방법에 기반한 것이다(Kucan & Beck 1997). 이에 근거하여 사고 논술형은 평가지에 쓰기 평가에서 자신의 생각이나 견해, 글이나 평가에 대한 이유를 글로 나타내며 평가한 경우이다. 키워드 제시형은 주제나 화제가 되는 용어에 특정한 표시를 하며 평가한 경우이다. 오류 분석형은 띄어쓰기, 맞춤법, 문단 나누기 등의 학생 글에서 오류를 찾아 표시하며 평가한 경우이다. 무표형은 아무런 표시가 없이 평가가 이루어진 경우이다. 이러한 기준으로 평가지에 나타난 표지를 확인하여 분류하고 구분하였다.

교사들로부터 수집한 자료에서 평가 방법을 결정하는 것은 주관이 반영될 수 있다는 점에 기반하여 국어 교사 1인과 연구자가 협의하여 결정하였다. 평가 방법은 30개의 텍스트 중에서 다수 나타난 표지 유형을 기준으로 하였으며 분류 기준이 모호하거나 중복하여 나타나는 경우에 대해서는 협의를 거쳐 최종 결정하였다. 예를 들어 텍스트에 전체적인 부분에 표지를 나타내지 않고 극히 일부만을 표시한 경우나 표지를 모호하게 나타내어 분류가 모호한 경우 등에 대해서는 4가지 유형 분류를 두고 국어 교사 1인과 협의하고 타당도 및 신뢰도를 검증하였다.

평가자들의 평가지에 나타난 표지를 기준으로 유표형(existence)과 무표형(nonexistence) 집단으로 구분하였다. 유표형은 사고논술형, 키워드 제시형, 오류 분석형으로 세분화하고 무표형으로 구분하였다. 그 결과는 〈표 3〉과 같다.

〈표 3〉 평가 방법에 따른 동일 집단 분포

| 군집 \ 기준 | | 평가 방법 | 합계(N) |
|---|---|---|---|
| 유표형 | 사고논술형 | **r8***, **r40***, **r44*** | 3 |
| | 키워드 제시형 | r7, **r8***, **r20***, r30, r33, r34, **r41*** | 7 |
| | 오류 분석형 | r5, r6, r9, r10, r14, r15, r16, r19, **r20***, r29, r37, r38, **r40***, **r41***, **r44***, r51 | 16 |
| 무표형 | | r1, r2, r3, r4, r11, r12, r13, r17, r18, r21, r22, r23, r24, r25, r26, r27, r28, r31, r32, r35, r36, r39, r42, r43, r45, r46, r47, r48, r49, r50, r52 | 31 |

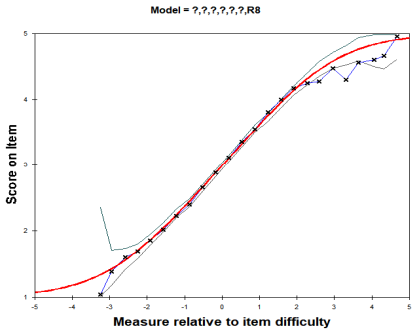
* 표는 동일 점수로 두 집단에서 중복하여 포함함.

〈표 3〉에 따르면, 평가 방법은 평가지에 나타난 표지를 기준으로 분석한 결과 4개의 집단으로 분류되었다. 사고논술형 집단은 3명, 키워드 제시형 집단은 7명, 오류 분석형 집단은 16명, 무표형 집단은 31명으로 분류되었다. 결과적으로 평가 방법을 기준으로 사고논술형, 키워드 제시형, 오류 분석형, 무표형 4개의 집단으로 구분하였다. 〈표 3〉에 따라 분류한 평가자들을 대상으로 집단별 차이를 확인한 결과, 사고논술형, 키워드 제시형, 오류 분석형, 무표형 집단의 결과가 유의한 차이를 보이는 것으로 확인되었다. 집단 간의 문항(평가 요인)에 따른 유의한 차이가 있는지 검증하기 위하여 일원배치 분산분석을 실시한 결과, 조직, 표현, 단어 선택의 평가 요인에서는  유의한 차이가 있는 것으로 나타났으나 내용, 형식 및 어법의 평가 요인에서는 유의한 차이가 없는 것으로 나타났다.

# IV. 연구 결과 및 논의

## 1. 분석 자료의 적합도

현직 국어 교사들의 평가 결과가 다국면 Rasch 모형에 적합한지를 확인하기 위해 모형 적합도를 분석하였다. 평가자들의 평가 방법에 따른 차이를 확인하기 위해 국어 교사들을 통해 수집한 평가지에 나타난 표지 형태를 적용하여 살펴보았다. 그 결과는 [그림 2]와 같다.

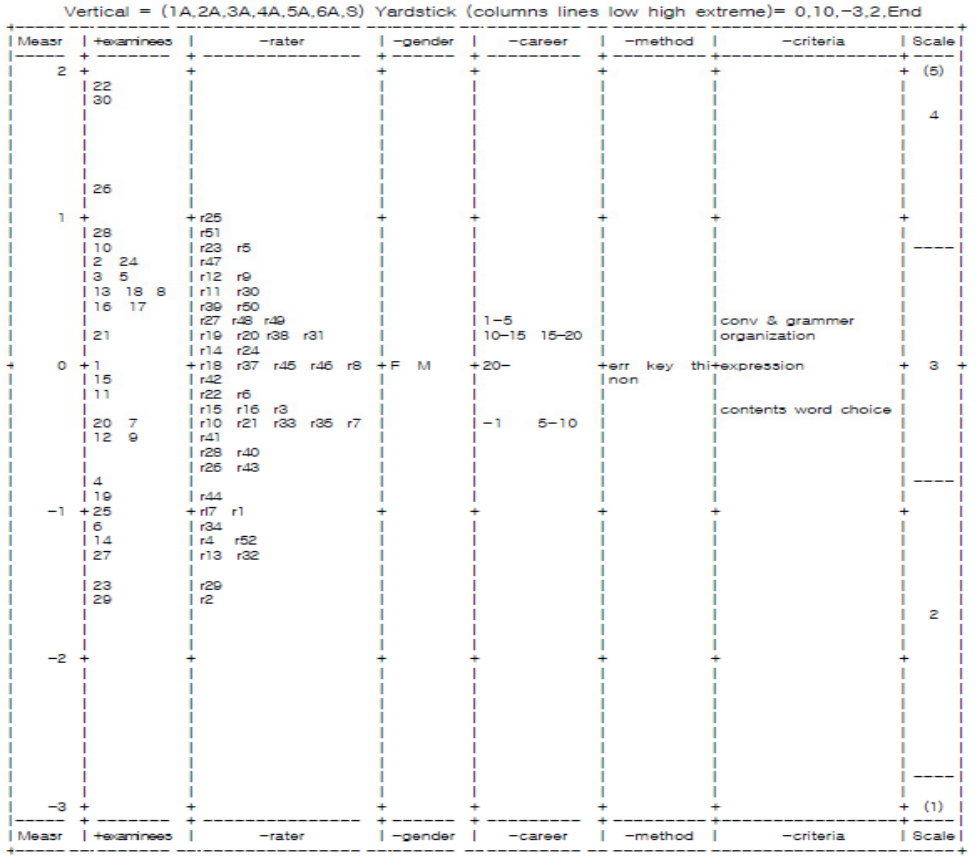



[그림 2] 고등학생 논술 텍스트 평가 자료의 모형 적합도

[그림 2]의 모형 적합도는 측정 모형에 기반하여 산출한 확률값인 관찰 점수로, 관찰 점수가 대체로 신뢰 구간에 있다는 것을 확인할 수 있다. 그러나 부분적으로 난도 추정치 3~4의 범주에서 95% 신뢰 구간 밖에 소수의 관찰 점수들이 존재한다는 것을 확인할 수 있다. 대부분의 관찰 점수가 신뢰 구간 안에 있다는 점에서 다국면 Rasch 모형에 따른 분석에 적합하다고 볼 수 있다.

## 2. 엄격성 및 일관성 분석

현직 국어 교사들의 평가 자료를 분석하기 위해 FACETS 프로그램을 활용하였다. 평가자의 주요 국면에 따라 평가자들의 채점 결과에 대한 logit 분포를 도식화해 보면 [그림 3]과 같다.




[그림 3] 평가 방법 적용에 따른 측정 단면 분포도

[그림 3]은 각 국면에 따른 logit 값에 따른 모수 추정치를 나타낸 것이다. 이는 평가 방법에 따른 각 국면의 분포가 어떻게 이루어졌는지 양상을 파악하는 토대가 된다. 즉 평가자와 관련된 국면에서는 교사의 엄격성 수준을 파악할 수 있다.

[그림 3]의 측정 단면 분포도에 따르면, 현직 국어 교사들의 각 국면에 따른 logit 모수 추정 수치는 다음과 같은 결과를 나타낸다. 첫째, 평가자 국면에서 국어 교사들의 엄격성은 대체로 채점의 일관성을 유지하여 엄격하게 평가한 것으로 나타났다. 상대적으로 25번 국어 교사는 엄격하게 평가를 한 것으로 나타났으며 2번 국어 교사는 관대하게 평가를 한 것으로 나타났다. 둘째, 성별 국면에서 국어 교사들은 남교사와 여교사가 동일한 엄격성을 나타내는 것으로 확인되었다. 셋째, 경력별 국면에서 국어

교사들은 '1년 이상 5년 미만'의 경력을 가진 교사가 가장 엄격한 평가를 하였으며 '1년 미만', '5년 이상 10년 미만'의 경력을 가진 교사가 가장 관대하게 평가를 한 것으로 나타났다. 넷째, 평가 방법 국면에서 국어 교사들은 '오류 분석형', '키워드 제시형', '사고 논술형'을 포함하는 유표형이 무표형에 비해 다소 엄격하게 평가하는 것으로 나타났다. 그러나 측정 단면 분포도 상에서는 유표형과 무표형 엄격성의 차이가 미미한 것으로 나타났다. 다섯째, 평가 요인 국면에서 국어 교사들은 '형식 및 어법'에 대한 점수가 가장 엄격한 평가를 하였으며 '내용'과 '단어 선택'은 가장 관대한 평가를 한 것으로 나타났다. 학생들의 관점에서 보면 논술에서 '형식 및 어법'에서 가장 점수를 받기 어려운 요인이며 '내용'과 '단어 선택'의 요인이 가장 점수를 받기 쉬운 요인라고 할 수 있다. 여섯째, 평가 척도 국면에서 국어 교사들의 평가 척도는 대체로 평가 결과가 2등급, 3등급, 4등급이 되도록 평가한 것으로 나타났다. 국어 교사들은 상대적으로 3등급이 가장 높은 빈도로 나타났으며 2등급, 4등급 순으로 많이 나타났다. 평균 척도가 평균 이하로 예측되어 대체로 엄격하게 평가한 것으로 판단된다.

다음으로, 평가자의 엄격성과 일관성은 평가자에 대한 신뢰도 뿐만 아니라 평가 체계에 대한 신뢰도를 좌우하는 중요한 요인이다. 일관성 있는 평가 자료는 평가자 간 점수 차이를 조정이 가능하지만 일관성이 없는 평가 자료는 조정이 불가능하기에 제외시키는 것이 적절하다고 본다(박종임, 2013). 이에 따라 현직 국어 교사들의 엄격성 및 일관성에 대해 구체적으로 확인하기 위해 평가자의 엄격성 수준과 적합도를 확인하였다. 그 결과는 〈표 4〉와 같다.

〈표 4〉는 현직 국어 교사들이 평가 과정에서 나타난 엄격성과 일관성이 적절한 수준에 있는지를 나타내는 지표이다(박영민, 2012). 이 지표는 엄격성이 일관성을 유지하고 있는지를 파악할 수 있는 토대가 된다. 〈표 4〉에 따르면 평가자의 엄격성은 .97 logit(SE=.11)부터 -2.57 logit(SE=.13)까지 분포하는 것으로 나타난다. 이러한 분포는 평가자 사이에 엄격성이 차이가 있다는 것을 의미한다. 전체 엄격성의 차이에 대한 분리 신뢰도 분석 결과에서 카이제곱($X^2$)= 2283.3(p=.000), 분리 지수(Separation)=7.01, 분리 신뢰도(R)=.98로 통계적으로 유의하였다. 또한 쓰기 평가에서 평가자 r25번은 가장 엄격하게 점수를 부여하였으며(.97 logit), 평가자 r36번은 가장 관대하게 점수를 부여하였다(=-2.57 logit).

이들을 내적합 지수와 내적합 표준화 값을 기준으로 적합도를 분석하면 평가자 r11번, r48번, r19번, r20번, r42번, r16번, r43번, r44번, r1번, r34번(총 10명, logit 내림차순)은 내적합 표준화 값이 +2 이상, 내적합 지수 1.3보다 큰 부적합(misft)으로 볼 수 있다. 한편 r14번(총 1명)은 내적합 표준화 값으로는 부적합에 속하나 내적합 지수 기준으로는 적합으로 분류되었다. 평가자 r14번을 포함하여 총 11명의 평가자들은 쓰기 평가를 할 때 쓰기 능력이 높은 평가 대상자들에게 낮은 점수를 부여하거나 쓰기 능력이 낮은 평가 대상자들에게 높은 점수를 부여하였다는 것이다.

이와는 반대로, 평가자 r5번, r24번, r8번. r37번, r45번, r6번, r3번, r10번, r33번, r29번(총 10명, logit 내림차순)는 내적합 표준화 값이 -2 이하, 내적합 지수 0.75보다 작은 과적합(overfit)으로 볼 수 있다. 한편 r50번, r49번, r31번, r35번, r7번, r41번, r40번, r17번, r4번(총 9명, logit 내림차순)은 내적합 표준화 값으로는 과적합에 속하나 내적합 지수 기준으로는 적합으로 분류되었다. 평가자 9명을 포함하여 총 19명의 평가자들은 쓰기 평가를 할 때 쓰기 능력에 대한 정확한 변별을 보이지 않고 획일적이고 편향적인 양상의 점수를 부여하였다는 것이다. 점수 척도가 5점이라고 했을 때 잘 쓴 것으로

판단되는 논술 텍스트는 모두 최고점, 못 쓴 것으로 판단되는 텍스트는 모두 최저점을 부여하였다는 것이다. 중간 단위의 점수는 부여하지 않고 양 극단의 점수로 평가하는 양상을 보였다는 것을 의미한다. 그 외 평가자들의 내적합 표준화 값은 -2 이상, +2 이하, 내적합 지수의 값은 0.75 이상 1.3 이상의 범위 내에 속하여 대체로 엄격성에서 일관성을 유지했다고 볼 수 있다. 따라서 평가자 전체 52명 중 22명(42.31%)의 평가자는 쓰기 평가 점수를 부여할 때 엄격하게 일관성 있게 안정된 평가 양상을 보였다고 할 수 있다. 이러한 결과는 박영민(2012)에서 예비교사를 대상으로 일관성을 분석한 결과에서 확인된 비율과 유사하다.

〈표 4〉 평가자의 엄격성 수준과 적합도

| 평가자 | logit | SE | 평균 자승 잔차 | | | | 평가자 | logit | SE | 평균 자승 잔차 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Infit MS | Infit ZStd | Outfit MS | Outfit ZStd | | | | Infit MS | Infit ZStd | Outfit MS | Outfit ZStd |
| r25 | .97 | .11 | 1.02 | .2 | 1.02 | .1 | r22 | -.23 | .10 | .85 | -1.4 | .84 | -1.5 |
| r51 | .93 | .11 | .99 | .0 | .98 | -.1 | r6 | -.24 | .11 | .46 | -6.2 | .46 | -6.1 |
| r23 | .84 | .11 | .90 | -.9 | .91 | -.7 | r16 | -.25 | .10 | 1.32 | 2.7 | 1.31 | 2.6 |
| r5 | .75 | .11 | .47 | -5.9 | .50 | -5.6 | r3 | -.29 | .11 | .74 | -2.5 | .74 | -2.5 |
| r47 | .73 | .11 | 1.14 | 1.2 | 1.11 | 1.0 | r15 | -.33 | .11 | 1.23 | 1.9 | 1.21 | 1.8 |
| r9 | .64 | .11 | .93 | -.6 | .91 | -.7 | r35 | -.38 | .11 | .72 | -2.8 | .70 | -2.9 |
| r12 | .60 | .11 | .81 | -1.8 | .80 | -1.9 | r7 | -.41 | .07 | .60 | -5.9 | .60 | -6.1 |
| r30 | .48 | .11 | .97 | -.2 | .96 | -.3 | r10 | -.42 | .11 | .49 | -5.7 | .50 | -5.6 |
| r11 | .47 | .11 | 1.86 | 6.2 | 1.83 | 6.1 | r21 | -.43 | .11 | .92 | -.7 | .91 | -.8 |
| r50 | .45 | .11 | .69 | -3.0 | .70 | -2.9 | r33 | -.44 | .11 | .49 | -5.7 | .49 | -5.8 |
| r39 | .42 | .11 | 1.15 | 1.3 | 1.14 | 1.2 | r41 | -.49 | .08 | .78 | -2.9 | .81 | -2.5 |
| r49 | .32 | .11 | .69 | -3.0 | .70 | -3.2 | r28 | -.57 | .11 | 1.17 | 1.5 | 1.17 | 1.4 |
| r48 | .31 | .11 | 1.76 | 5.7 | 1.74 | 5.5 | r40 | -.59 | .07 | .77 | -3.2 | .76 | -3.3 |
| r27 | .29 | .11 | .94 | -.5 | .92 | -.6 | r43 | -.69 | .11 | 1.32 | 2.6 | 1.36 | 2.9 |
| r31 | .24 | .11 | .75 | -2.4 | .73 | -2.7 | r26 | -.74 | .11 | .90 | -.8 | .94 | -.4 |
| r19 | .21 | .11 | 1.35 | 2.9 | 1.34 | 2.8 | r44 | -.85 | .08 | 1.99 | 9.0 | 1.94 | 9.0 |
| r38 | .16 | .11 | .95 | -.4 | .94 | -.5 | r1 | -.95 | .11 | 1.54 | 4.2 | 1.74 | 5.5 |
| r20 | .16 | .07 | 1.82 | 8.5 | 1.82 | 8.5 | r17 | -1.01 | .11 | .71 | -2.8 | .71 | -2.8 |
| r24 | .14 | .11 | .50 | -5.5 | .49 | -5.7 | r34 | -1.12 | .11 | 1.33 | 2.7 | 1.39 | 3.1 |
| r14 | .08 | .11 | 1.25 | 2.1 | 1.24 | 2.0 | r4 | -1.23 | .11 | .76 | -2.3 | .78 | -2.1 |
| r8 | .03 | .11 | .67 | -4.7 | .67 | -4.7 | r52 | -1.24 | .11 | .87 | -1.2 | .87 | -1.2 |
| r37 | .02 | .11 | .60 | -4.2 | .60 | -4.2 | r32 | -1.25 | .11 | .93 | -.6 | .93 | -.6 |
| r45 | -.03 | .11 | .58 | -4.5 | .58 | -4.5 | r13 | -1.34 | .11 | .80 | -1.9 | .80 | -1.8 |
| r46 | -.03 | .11 | 1.04 | .4 | 1.02 | .2 | r29 | -1.54 | .11 | .62 | -3.9 | .64 | -3.7 |
| r8 | -.03 | .07 | 1.20 | 1.7 | 1.19 | 1.6 | r2 | -1.59 | .11 | .97 | -.1 | 1.06 | .5 |
| r42 | -.10 | .11 | 1.62 | 4.8 | 1.59 | 4.6 | r36 | -2.57 | .13 | 1.14 | 1.1 | 1.25 | 1.7 |
| $\overline{X}$ | logit =-.23 | | S.E =.10 | | Infit MS =.98 | | Infit ZStd=.6 | | | Outfit MS =.99 | | Outfit ZStd=-.5 | |
| s | logit =.74 | | S.E =.01 | | Infit MS =.38 | | Infit ZStd=3.6 | | | Outfit MS =.38 | | Outfit ZStd=3.7 | |

## 3. 성별에 따른 엄격성 및 일관성 분석

성별에 따른 쓰기 평가 엄격성 및 일관성에 차이가 있는지 분석하였다. 이는 평가자의 성별 차이가 쓰기 평가의 특성 차이로 연계될 수 있고 쓰기 평가의 결과에 영향을 미칠 수 있기 때문이다.

성별에 따른 엄격성 수준과 적합도를 파악하였다. 그 결과는 〈표 5〉와 같다.

〈표 5〉 성별에 따른 엄격성 수준과 적합도

| 평가자 성별 | logit | SE | Infit MS | Infit ZStd | Outfit MS | Outfit ZStd |
|---|---|---|---|---|---|---|
| 여(F) | .03 | .02 | 1.06 | 3.3 | .93 | 3.3 |
| 남(M) | -.03 | .02 | .87 | -6.0 | 1.15 | -5.3 |
| $\overline{X}$ | .00 | .02 | .96 | -1.3 | .97 | -1.0 |
| s | .05 | .01 | .15 | 6.7 | .13 | 6.2 |

성별에 따른 일관성은 내적합 지수를 기준으로 볼 때 여교사들은 1.06, 남교사들은 .87으로 나타나 적합한 수준의 범위에 속하는 것으로 확인되었다. 내적합 표준화 값을 기준으로 볼 때 여교사들은 부적합 양상을, 여교사들은 과적합 양상을 보이는 것으로 확인되었다. 〈표 5〉에 따르면 평가자들은 쓰기 평가에서 성별에 따라 엄격성이 다르다는 것을 의미한다. 성별에 따른 엄격성은 -.03 logit(SE=.02)부터 .03 logit(SE=.02)까지의 범위에서 나타났다. 여교사는 평가 대상자의 논술 텍스트를 보다 엄격하게 평가하였으며 상대적으로 남교사는 관대하게 평가한 것으로 나타났다. 평가자의 성별에 따른 모형 적합도 분석에서 외적합 지수가 .93~1.15에 분포하는 것을 확인할 수 있다. 이는 평가자의 성별에 따른 엄격성에는 차이가 있어도 엄격성이 일관성을 유지하고 있다는 것을 의미한다.

성별에 따른 엄격성의 차이에 대한 분리 신뢰도 분석 결과에서 카이제곱($X^2$)= 4.9(p=.03), 분리 지수(Separation)= 1.99, 분리 신뢰도(R)=.80으로 통계적으로 유의하였다. 성별에 따른 엄격성은 타당도와 신뢰도가 높은 수준으로 판단할 수 있다.

## 4. 경력에 따른 엄격성 및 일관성 분석

경력에 따른 엄격성 수준과 적합도를 파악하였다. 그 결과는 〈표 6〉과 같다.

〈표 6〉 경력에 따른 엄격성 수준과 적합도

| 평가자 교직 경력 | logit | SE | Infit MS | Infit ZStd | Outfit MS | Outfit ZStd |
|---|---|---|---|---|---|---|
| 1년 이상~5년 미만(1-5) | .35 | .03 | .89 | -3.7 | .88 | -3.9 |
| 15년 이상~20년 미만(15-20) | .24 | .04 | 1.01 | .3 | 1.00 | .01 |
| 10년 이상~15년 미만(10-15) | .18 | .04 | 1.25 | 5.1 | 1.25 | 5.1 |
| 20년 이상(20-) | .02 | .02 | .95 | -2.2 | .95 | -2.0 |
| 1년 미만(-1) | -.39 | .08 | 1.22 | 2.6 | 1.32 | 3.7 |
| 5년 이상~10년 미만(5-10) | -.40 | .05 | 1.05 | .8 | 1.09 | 1.6 |
| $\overline{X}$ | .00 | .04 | 1.06 | .5 | 1.08 | .8 |
| s | .32 | .02 | .14 | 3.2 | .17 | 3.5 |

경력에 따른 일관성은 적합한 수준에 있는 것으로 나타났다. 내적합 지수를 기준으로 볼 때 최소 수치를 보이는 1년 이상 5년 미만의 경력을 가진 교사들이 .89, 최대 수치를 보이는 10년 이상 15년 미만의 경력을 가진 교사들이 1.25로 나타나 적합한 수준의 범위에 있는 것으로 확인되었다. 내적합 표준화 값을 기준으로 볼 때 1년 이상 5년 미만의 경력을 가진 교사들과 20년 이상의 경력을 가진 교사들은 과적합의 양상을, 10년 이상 15년 미만의 경력을 가진 교사들과 1년 미만의 경력을 가진 교사들은 부적합의 양상을 보이는 것으로 확인되었다. 15년 이상 20년 미만의 경력을 가진 교사들과 5년 이상 10년 미만의 경력을 가진 교사들에게서만 적합한 양상이 확인되었다. 〈표 6〉에 따르면 평가자들은 쓰기 평가에서 경력에 따라 엄격성이 다르다는 것을 의미한다. 경력에 따른 엄격성은 -.40 logit(SE=.05)부터 .35 logit(SE=.03)까지의 범위에서 나타났다. 평가자의 경력에 따른 모형 적합도 분석에서 외적합 지수가 .88~1.32에 분포하는 것을 확인할 수 있다. 이는 평가자의 경력에 따라 엄격성이 일관성을 유지하고 있다는 것을 의미한다.

경력에 따른 엄격성의 차이에 대한 분리 신뢰도 분석 결과에서 카이제곱($X^2$)= 225.1(p=.00), 분리지수(Separation)= 6.77, 분리 신뢰도(R)=.98로 통계적으로 유의하였다. 따라서 경력에 따른 엄격성은 타당도와 신뢰도가 높은 수준으로 판단할 수 있다.

구체적으로 경력에 따른 평가자의 엄격성 차이를 파악해 보면 1년 이상~5년 미만의 경력을 갖는 평가자가 가장 엄격성이 높게 나타났고 15년 이상~20년 미만, 10년 이상~15년 미만, 20년 이상, 1년 미만, 5년 이상~10년 미만 순으로 엄격성이 높았다. 한편 5년 이상~10년 미만의 경력을 가진 평가자들은 가장 관대하게 평가를 한 것으로 나타났다. 이상의 내용을 종합해 보면 대체로 경력이 낮은 집단과 경력이 많은 집단의 엄격성이 혼재되어 나타나는 것을 확인할 수 있다.

## 5. 평가 방법에 따른 엄격성 및 일관성 분석

평가 방법을 적용한 주요 국면에 따른 엄격성 수준과 적합도를 파악하였다. 그 결과는 〈표 7〉과 같다.

〈표 7〉 평가 방법에 따른 엄격성 수준과 적합도

| 평가자 평가 방법 | logit | SE | Infit MS | Infit ZStd | Outfit MS | Outfit ZStd |
|---|---|---|---|---|---|---|
| 오류 분석형(error) | .02 | .03 | 1.01 | .4 | 1.01 | .1 |
| 사고논술형(think) | .02 | .06 | 1.13 | 2.0 | 1.12 | 1.8 |
| 키워드 제시형(key) | .02 | .04 | .95 | −1.2 | .96 | −.9 |
| 무표형(non) | −.05 | .02 | .98 | −1.0 | .99 | −.4 |
| $\overline{X}$ | .00 | .04 | 1.02 | .0 | 1.02 | .2 |
| s | .03 | .02 | .08 | 1.6 | .07 | 1.2 |

평가 방법에 따른 일관성은 내적합 지수를 기준으로 볼 때 최소 수치를 보이는 키워드 제시형의 평가 방법을 사용한 교사들이 .95, 최대 수치를 보이는 사고논술형 평가 방법을 사용한 교사들이 1.13으

로 나타나 적합한 수준의 범위에 있는 것으로 확인되었다. 내적합 표준화 값을 기준으로 볼 때 키워드 제시형, 무표형의 평가 방법을 사용한 교사들은 과적합의 양상을, 사고논술형의 평가 방법을 사용한 교사들은 부적합의 양상을 보이는 것으로 확인되었다. 오류 분석형의 평가 방법을 사용한 교사들에게서만 적합한 양상이 확인되었다. 〈표 7〉에 따르면 평가자들은 평가 방법에 따라 엄격성이 다르다는 것을 의미한다. 평가 방법에 따른 엄격성은 -.05 logit(SE=.02)부터 .02 logit(SE=.06)까지의 범위에서 나타났다. 평가 방법에 따른 모형 적합도 분석에서 외적합 지수가 0.96~1.12에 분포하는 것을 확인할 수 있다. 이는 평가자의 평가 방법에 따른 엄격성이 일관성을 유지하고 있다는 것을 의미한다.

평가 방법에 따른 엄격성의 차이에 대한 분리 신뢰도 분석 결과에서 카이제곱($X^2$)=6.1(p=.10), 분리 지수(Separation)=1.90, 분리 신뢰도(R)=.81로 통계적으로 유의한 차이가 없는 것으로 나타났다. 따라서 평가 방법에 따른 엄격성은 타당도와 신뢰도가 높은 수준으로 판단할 수 있다.

이상의 내용을 종합해 보면, 유표형이 무표형에 비해 엄격성이 높게 나타나며 유표형 간에는 동일한 엄격성이 나타났다. 이는 유표형일수록 엄격하게 평가하고 무표형일수록 관대하게 평가한다는 것을 의미한다.

## 6. 평가 요인에 따른 엄격성 및 일관성 분석

평가 요인에 따른 엄격성 수준과 적합도를 파악하였다. 그 결과는 〈표 8〉과 같다.

〈표 8〉 평가 요인에 따른 엄격성 수준과 적합도

| 평가 요인 | logit | SE | Infit MS | Infit ZStd | Outfit MS | Outfit ZStd |
|---|---|---|---|---|---|---|
| 형식 및 어법(conv & grammar) | .33 | .03 | .94 | -1.8 | .94 | -1.8 |
| 조직(organization) | .20 | .03 | 1.14 | 4.0 | 1.13 | 3.8 |
| 표현(expression) | .03 | .03 | .99 | -.4 | .99 | -.2 |
| 내용(contents) | -.26 | .03 | 1.05 | 1.6 | 1.08 | 2.3 |
| 단어 선택(word choice) | -.30 | .03 | .84 | -5.0 | .86 | -4.6 |
| $\overline{X}$ | .00 | .03 | .99 | -.3 | 1.00 | -.11 |
| s | .27 | .00 | .11 | 3.4 | .11 | 3.4 |

평가 요인에 따른 일관성은 내적합 지수를 기준으로 볼 때 최소 수치를 보이는 단어 선택의 평가 요인이 .84, 최대 수치를 보이는 조직의 평가 요인이 1.14로 나타나 적합한 수준의 범위에 있는 것으로 확인되었다. 내적합 표준화 값을 기준으로 볼 때 단어 선택 요인은 과적합의 양상을, 조직 요인은 부적합의 양상을 보이는 것으로 확인되었다. 형식 및 어법, 표현, 내용 요인은 적합한 양상이 확인되었다. 〈표 8〉에 따르면 평가자들은 쓰기 평가에서 평가 요인에 따른 난도는 -.30 logit(SE=.03)부터 .33 logit(SE=.03)까지의 범위에서 나타났다. 평가 요인에 따른 모형 적합도 분석에서 외적합 지수가 0.86~1.13에 분포하는 것을 확인할 수 있다. 이는 평가 요인에 따른 엄격성이 일관성을 유지하고 있다는 것과 평가 요인의 난도 차이는 있지만 모든 평가 요인이 모형에 적합하다는 것을 의미한다.

평가 요인에 따른 엄격성의 차이에 대한 분리 신뢰도 분석 결과에서 카이제곱($X^2$)= 318.5(p=.00), 분리 지수(Separation)= 8.87, 분리 신뢰도(R)=.99로 통계적으로 유의하였다. 따라서 평가 요인에 따른 엄격성은 타당도와 신뢰도가 높은 수준으로 판단할 수 있다.

이상의 내용을 종합해 보면, 형식 및 어법에 대한 엄격성이 가장 높고 조직, 표현, 내용, 단어 선택 순으로 나타났다. 대체로 형식 및 어법의 평가 요인에 대해서는 엄격하게 평가하고 단어 선택에 대해서는 관대하게 평가한다는 것을 의미한다.

# V. 결 론

이 연구는 국어 교사들의 쓰기 평가에서 나타난 엄격성 및 일관성의 차이를 분석하는 데 있다. 이를 위해 현직 국어 교사 52명을 대상으로 고등학생이 작성한 논술 텍스트를 평가하게 한 후 엄격성 및 일관성을 확인하였다. 현직 국어 교사들이 논술이라는 문종을 통해 나타난 쓰기 평가의 엄격성 및 일관성에 대한 결과는 국어 교사들의 쓰기 평가 전문성 신장을 위한 방향과 정보를 제공한다는 점에서 의미가 있다. 특히 논술이라는 문종을 활용하고 평가 방법의 차이까지 논의하였다는 점에서 논술 평가에 대한 정보와 개선 방안에 대한 자료를 제공한다.

이 연구에서 확인된 결과는 다음과 같다.

첫째, 엄격한 평가자와 관대한 평가자를 내적합 지수 및 내적합 표준화 값을 기준으로 분석한 결과에서 적합한 평가자가 22명(42.31%), 부적합 평가자가 11명(21.15%), 과적합 평가자가 19명(36.54%)으로 확인되었다. 이는 일반적인 쓰기 평가를 대상으로 하는 선행 연구에서도 유사한 비율을 보인다는 것을 확인할 수 있다(장은주, 2015; 박영민·최숙기, 2009). 19명의 과적합 평가자는 특정한 점수의 집중 경향을 보이며, 11명의 부적합 평가자는 고등학생들의 논술 능력을 적절하게 변별하지 못한다는 것을 의미한다. 부적합과 과적합 평가자들은 엄격성과 일관성을 제대로 유지하지 못했다는 점에서 쓰기 평가 전문성 신장을 위한 방안이 필요하다는 것을 의미한다.

둘째, 국어 교사들의 엄격성을 분석한 결과에서 평가자의 특성을 반영한 국면에 따라 차이를 보였다. 먼저, 성별에 따른 엄격성은 통계적으로 유의한 차이가 있는 것으로 확인되었다. 성별에 따른 엄격성은 유의한 차이가 있는 것으로 지적되기도 하고(박영민·최숙기, 2009; Peterson & Kennedy, 2006; Peterson, 1998) 유의한 차이가  없는 것으로 나타나기도 한다(박영민, 2012). 이 연구에서는 여교사가 남교사보다 엄격하게 평가하며 유의한 차이가 있는 것으로 나타났다. 다음으로, 경력에 따른 엄격성은 통계적으로 유의한 차이가 있는 통계적으로 유의한 차이가 있는 것으로 확인되었다. 이 연구에서는 1년 이상 5년 미만의 경력을 가진 집단이 가장 엄격하게 평가하고 5년 이상 10년 미만의 경력을 가진 집단이 가장 관대하게 평가하며 유의한 차이가 있는 것으로 나타났다. 이어서 평가 방법에 따른 엄격성은 통계적으로 유의한 차이가 없는 것으로 확인되었다. 이 연구에서는 오류 분석형, 사고논술형, 키워드 제시형의 평가 방법을 사용한 집단에서 동일한 엄격성을 보이며 엄격하게 평가하고 무표형이 다소 낮

은 엄격성을 보이며 관대하게 평가한 것으로 확인되었으나 유의한 차이는 없는 것으로 나타났다. 마지막으로, 평가 요인에 따른 엄격성은 통계적으로 유의한 차이가 있는 것으로 확인되었다. 이 연구에서는 형식 및 어법 요인에서 가장 엄격하게 평가를 하고 단어 선택 요인에서 가장 관대하게 평가하며 유의한 차이가 있는 것으로 나타났다. 이러한 결과는 평가자 특성에 따라 엄격성이 차이가 나타났다는 것을 시사하며 평가 방법에 따른 엄격성의 차이가 없다는 것은 쓰기 평가에서 교사들의 개별 평가 방법을 인정하는 것이 엄격성의 차이를 줄이는 방안이 될 수 있다는 것을 시사한다고 볼 수 있다.

셋째, 국어 교사들의 일관성을 분석한 결과에서 내적합 지수를 기준으로 할 때 성별, 경력, 평가 방법, 평가 요인의 모든 국면에서 적합한 수준을 유지하는 것으로 확인되었다. 이러한 결과는 평가 결과를 집단별 평균을 통해 산출한 결과를 바탕으로 비교하였다는 점에서 개별 평가자의 특성이 반영되지 않아 나타난 결과로 볼 수 있다. 그러나 내적합 표준화 값을 기준으로 할 때 다양한 일관성의 양상이 확인되었다.

이러한 결과를 종합해 볼 때 이 연구에 위촉된 현직 국어 교사들 중에서 적합한 평가자가 42.31%에 불과하다는 것은 나머지 57.69%의 평가자가 과적합 및 부적합의 평가 양상을 보인다는 것으로 해석할 수 있다. 이 결과는 예비교사들을 대상으로 일관성을 분석한 결과에서 확인된 비율과 매우 유사하다(박영민, 2012). 이는 평가 경력과 평가 전문성을 가진 현직 국어 교사들에게서 평가 전문성이 부족한 예비국어교사들과 유사한 비율을 보인다는 것은 현직 국어 교사들에게도 평가 전문성 신장을 위한 방안이 필요하다는 것을 시사한다.

지금까지 많은 연구들에서 쓰기 평가 전문성을 높이는 방안으로 평가 연수 및 평가 강좌를 활용하는 방안이 제시되고 있다. 이러한 점을 고려할 때 논술이라는 문종의 평가에서도 쓰기 평가 전문성을 신장할 수 있는 평가 연수나 평가 강좌를 활용한 평가 방안이 모색될 필요가 있을 것이다. 또한 성별, 경력, 평가 요인에서 엄격성의 유의한 차이가 있는 것으로 확인된다는 것은 이들 요인에 의해 평가 결과의 차이가 나타날 수 있다는 것을 의미한다. 이러한 점에서 쓰기 평가 과정에서 평가자의 특성으로 성별, 경력, 평가 요인에 대해 고려할 필요가 있다는 것을 시사한다. 그러나 평가 방법에서 엄격성의 유의한 차이가 없다는 것은 개별 평가자들의 평가 방법을 존중하는 것이 평가 결과의 차이를 줄이는 하나의 방법이 될 수 있다는 것을 시사한다.

## 참고문헌

김경희 · 송미영(2001), 채점척도에 따른 채점자의 일관성과 피험자 능력 추정의 정확성 비교. **교육평가연구** 14(1), 327-347.

박영민(2012a), 예비국어교사의 중학생 논설문 평가에서 발견되는 엄격성 및 일관성의 특성, **국어교육학연구** 43, 253-283.

박영민(2012b), 쓰기 평가 지식 측정을 위한 검사 도구 개발 연구, **청람어문교육** 45, 29-46.

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 -중학생 설명문 쓰기 평가를 중심으로- , **국어교육학연구** 37, 367-391.

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

박찬흥(2018), 예비 국어교사의 설명문 분석적 채점에서 나타나는 특성에 대한 연구, **작문연구** (39), 349-369.

설현수(2010), 평정자 간의 엄격성 차이 정도가 피험자 총점 산출 방법에 미치는 영향: 원점수, 표준점수, Facet점수 비교, **교육평가연구** 23(1), 125-147.

송영미·김수진·김희경·남명호(2009), 온라인 시스템을 활용한 대규모 서답형 평가의 채점 일관성, **교육평가연구** 22(3), 827-846.

안수현·김정숙(2017), 한국어능력시험(TOPIK) 쓰기 평가의 채점 특성 연구, **한국어교육** 28(1), 173-196.

이영식(1998), 영어 작문 평가의 채점 신뢰도에 대한 분석. **영어교육** 53(1), 179-200.

장미·박영민(2021), 엄격성의 일관성 유형에 따른 국어교사의 논설문 평가 특성 분석, **학습자중심교과교육연구** 21(11), 277-288.

장소영, 신동일(2009), **언어교육평가 연구를 위한 FACETS 프로그램: 기초 과정편**, 서울: 글로벌콘텐츠.

장은주(2015), 채점의 일관성 유형에 따른 국어교사의 쓰기 평가 특성 분석, 박사학위논문, 한국교원대학교.

지은림 (2008), 논술고사의 신뢰성에 영향을 미치는 채점자 특성 분석. **교육평가연구** 21(2), 97-113.

지은림·백순근·설현수·채선희 역(2003), **문항반응이론의 이론과 실제 : 외국어 수행평가를 중심으로,** 서울: 서현사, [McNamara, T. F.(1996), Measuring Second Language Performance. New York: Longman].

최숙기·박영민(2011), 논설문 평가에 나타난 국어교사의 평가 특성 및 편향 분석, **교육평가연구** 14(1), 201-228.

Bachman, L. F. (1990), *Fundamental considerations in language testing*, New York, NY: Oxford University Press,

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, NJ: Lawrence Erlbaum associates.

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

Ford, A. (1931). Neutralizing inequalities in ratings. *Personnel Journal*, 9, 466-469.

Guilford, J. P. (1954). *Psychometric methods (2nd ed.)*. New York: McGraw Hill.

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

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: University of Chicago Press. (Original work published in 1960).

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.

Hamp-Lyons, L. (1990), Second language writing : Assessment issues, In B. Kroll(Ed.), *Second language writing : Research insights for the classroom* (pp.69-87), Cambridge : Cambridge University Press.

Kneeland, N. (1929). That lenient tendency in rating. *Personnel Journal*, 7, 356-366.

Kortez, D. (1993). New report on Vermont portfolio project documents challenges. *National Councilon Measurement in Education*, 1(4), 1-2.

Kroll, B. (1998), *Assessing writing abilities.* Annual Review of Applied Linguistics, 18, 219-240.

Kucan, L. & Beck, I. l. (1997), Thinking Aloud and Reading Comprehension Research: Inquiry, Instruction, and Social Interaction. *Review of Educational Research*,

67(3), 271-299.

Linacre, J. M. (1989, 1994). *Many-faceted Rasch Measurement*. Chicago: MESA Press.

Linacre, J. M. (1996). *A user's guide to Facets: Rasch measurement computer program(Version 3.0)*. Chicago, IL: MESA Press.

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

Lunz, M, E. & Stahl, J. A. (1990), Judge consistency and severity across grading periods, Evaluation and the Health Professions 13(4), 425-444.

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

Myford, C. M. & Wolfe, E. W. (2003), Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part Ⅰ. *Journal of Applied Measurement* 4(4), 386-422.

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.

Rasch, G. (1960). Probabilistic Models for some intelligence and attainment tests. Copenhagen: Danish Insttitute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.

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. (2002), *Assessing writing*, Cambridge : Cambridge University Press.

Wright, B. D., Linacre, J. M., Gustafson, J. E. & Martin-Lof, P. (1994), Reasonable mean-square fit values, *Rasch Measurement Transactions,* 8(3), 369-386.

· 논문접수 : 2022.07.05. / 수정본접수 : 2022.07.29. / 게재승인 : 2022.08.10.

## ABSTRACT

# The Differences Analysis of Korean Language Teacher's Scoring Severity and Consistence in Assessing Writing : Focusing on Essay Assessment

**Jeong Da-Un**
Ph. D. student, Korea National University of Education

The purpose of this study is to analyze the rating severity and rating consistence based on the results obtained after appointing 52 incumbent Korean language teachers as essay evaluators and scoring them. The results confirmed in this study are as follows. First, as a result of analyzing the rating consistence of the strictness of Korean language teachers, there were 22 (42.31%) suitable evaluators, 11 non-conforming evaluators (21.15%), and 19 (36.54%) overfit evaluators. Second, as a result of analyzing the strictness of the Korean language teachers, there was a difference according to the phase in which the characteristics of the evaluator were reflected. Specifically, there was a statistically significant difference in rating severity according to gender, career, and assessment factors, but no statistically significant difference was found in rating severity according to assessment method. Strictness according to gender was found to be evaluated more strictly by female teachers than by male teachers. It was confirmed that the rating severity according to experience was evaluated most strictly by the group with more than 1 year and less than 5 years of experience. In the assessment factors, it was found that the formal and usage factors were evaluated most strictly. On the other hand, it was confirmed that the group using the error analysis type, thinking essay type, and keyword presentation type assessment method evaluated the severity according to the assessment method more rating severityously than the group using the non-formal assessment method with the same severity. Third, as a result of analyzing the rating consistence of Korean language teachers, there were differences according to the discrimination criteria. Based on the intrinsic fit index, it was found to maintain an appropriate level in all aspects of gender, career, assessment method, and assessment factors. However, based on the standardized values of the internal fit, various patterns of rating consistence were observed according to phases and sub-factors. Based on these results, it seems that it is necessary to increase the professionalism of Korean language teachers in writing assessment in essay assessment.

***Key Words:*** *Korean language teacher, writing assessment, essay assessment, rating severity, rating consistence, Rasch measurement model*