

# 쓰기 평가에서의 평가자 효과는 어떻게 나타나는가?

— 평가자의 엄격성과 평가 준거 해석을 중심으로

장성민 인하대학교 국어교육과 부교수



- I. 서론
- II. 이론적 배경
- III. 연구 방법
- IV. 연구 결과
- V. 논의 및 제언
- VI. 연구의 기여와 향후 과제

## I. 서론

쓰기 평가는 필자가 산출한 글을 바탕으로 내용, 조직, 표현 등 복합적인 쓰기 능력을 판단하는 과정이다. 쓰기 수행은 하나의 정답으로 환원하기 어렵고 다양한 질적 특성이 상호작용하여 나타나므로, 평가자의 해석과 판단은 쓰기 평가에서 필수적인 역할을 한다(장성민, 2026: 229). 평가자 효과(rater effects)는 언어 수행 평가에서 평가자와 관련하여 나타나는 체계적인 점수 변산을 의미하며(Eckes, 2005; Eckes & Jin, 2021), 쓰기 능력의 측정과 평가 절차의 타당성을 저해할 수 있다. 따라서 쓰기 평가에서는 평가 대상 글의 수행 수준뿐 아니라 평가자가 어떠한 판단을 거쳐 점수를 부여하는지도 함께 분석할 필요가 있다.

기존 연구에서는 주로 다국면 라쉬 모형을 활용하여 평가자의 엄격성과 채점 일관성을 분석하고, 이를 토대로 쓰기 평가의 신뢰성과 타당성을 검토해 왔다(예: 박영민·최숙기, 2010ㄴ; 백유진·가은아, 2021). 그러나 엄격성과 일관성은 평가자의 전반적인 채점 성향을 보여 줄 뿐, 평가자가 평가 준거를 어떠한 관점에서 이해하고 텍스트에 적용하였는지를 직접 설명하지 못한다.

쓰기 평가는 평가 준거와 텍스트의 특성을 연결하는 판단 과정이므로 동일한 루브릭을 사용하더라도 준거의 의미와 적용 범위에 대한 해석에 따라 준거별 점수가 달라질 수 있다(Heidari, Ghanbari, & Abbasi, 2022: 2). 따라서 평가자 효과를 이해하기 위해서는 측정학적 특성과 함께 평가 준거의 해석 양상을 살펴볼 필요가 있다.

평가 준거의 해석과 적용은 평가자가 지닌 다양한 특성과도 관련될 가능성이 있다. 선행 연구에서는 성별, 평가 경험, 쓰기 지식과 신념, 전공 분야 및 교육 경험 등이 평가자의 판단과 관련되는지를 검토해 왔으나, 그 결과는 일관되지 않았다(박영민·최숙기, 2009, 2010ㄱ, 2010ㄴ; 백유진·가은아, 2021; 송지언, 2016; Eckes, 2005; Heidari et al., 2022; Soler, 2000). 특히 담당 학교급은 평가자가 지속적으로 경험해 온 교수·학습 환경과 학습자 수행에 대한 기대를 반영하는 전문적 배경 요인이 될 수 있으나, 학교급에 따른 차이가 평가 준거별 판단과 어떻게 관련되는지는 충분히 밝혀지지 않았다.

평가자 효과를 구체적으로 이해하기 위해서는 분석 수준도 함께 고려할 필요가 있다. 평가자 효과는 평가 과제, 평가 준거 및 평가 맥락에 따라 다르게 나타날 수 있으므로(Johnson & Lim, 2009; Leckie & Baird, 2011; Palermo, 2022), 총점만으로는 그 구체적인 양상을 충분히 파악하기 어렵다. 총점에서 드러나지 않은 집단 간 차이가 특정 평가 준거에서 나타날 수 있으며, 준거별 점수 차이가 확인되더라도 그 판단 근거는 점수 자료만으로 설명하기 어렵다. 따라서 총점과 준거별 점수를 함께 분석하고, 차이가 확인된 준거의 해석 과정을 질적으로 탐색할 필요가 있다.

이에 이 연구에서는 평가 대상 글, 평가 준거 및 평가자의 측정 특성을 확인하고, 평가자의 성별과 담당 학교급에 따른 평가자 효과를 전반적인 엄격성과 총점 및 평가 준거별 채점 양상의 차원에서 분석하고자 한다. 또한 학교급 간 차이가 확인된 평가 준거를 대상으로 후속 면담을 실시하여 평가자들이 평가 준거를 어떻게 해석하고 텍스트에 적용하였는지를 탐색한다. 이를 통해 평가자의 엄격성으로 대표되는 측정학적 평가자 효과를 확인하고, 이를 평가 준

거 해석 과정과 연결하여 설명함으로써, 평가자의 개인적·전문적 배경과 쓰기 평가 판단 사이의 관련성을 보다 심층적으로 밝히고자 한다.

이 연구의 연구 문제는 다음과 같다.

연구 문제 1. 평가자 효과를 분석하기 위한 평가 대상 글, 평가 준거 및 평가자의 측정 특성은 어떠한가?

- 1-1. 평가 대상 글의 수행 수준은 어떠한가?
- 1-2. 평가 준거의 측정 특성(난이도, 변별도)은 어떠한가?
- 1-3. 평가자의 채점 특성(엄격성, 일관성)은 어떠한가?

연구 문제 2. 평가자의 성별과 담당 학교급에 따라 평가자 효과는 어떻게 나타나는가?

- 2-1. 평가자의 성별과 담당 학교급에 따라 전반적인 엄격성과 총점·평가 준거별 채점 양상에는 어떠한 차이가 나타나는가?
- 2-2. 담당 학교급에 따라 차이가 나타난 평가 준거는 어떠한 판단 근거를 바탕으로 해석되고 적용되는가?

## II. 이론적 배경

### 1. 평가자 효과

평가자 효과(rater effects)는 언어 수행 평가에서 평가자와 관련하여 나타나는 다양한 형태의 측정 오차와 편향을 포괄하는 개념이다(Eckes & Jin, 2021: 66). 이는 단순히 평가자 간 점수 차이가 존재한다는 사실이 아니라, 평가 대상자의 실제 수행 수준과는 무관하게 평가자의 채점 성향이나 판단 방식이 점수에 지속적으로 반영되는 현상을 의미한다. 이러한 차이는 평가자가 측정하고자 하는 구성 개념(construct)에 포함되지 않는 ‘체계적(sys-

tematic)’ 변산으로 간주되며(Eckes, 2005: 197), 평가 대상자의 능력 추정과 평가 결과의 해석을 왜곡하여 평가의 타당성을 저해할 수 있다. 따라서 수행 평가에서는 평가 대상자의 수행뿐 아니라 평가자가 점수를 부여하는 과정에서 나타나는 체계적인 특성도 함께 검토할 필요가 있다.

평가자 효과는 동일한 수행이라도 어떤 평가자가 채점하느냐에 따라 서로 다른 평가 결과가 나타날 수 있다는 점에서 중요하다. Eckes(2005)는 평가자마다 점수를 부여하는 성향이 서로 다르며, 이러한 차이가 평가 대상자의 능력 추정에 영향을 미칠 수 있음을 확인하였다. 또한 평가자들이 전반적으로는 비교적 일관된 양상을 보이더라도, 평가 준거에 따라 엄격성의 수준이나 평정 척도의 사용 방식이 달라질 수 있으며, 평가자와 평가 준거의 상호작용에서도 모형의 기대와 다른 반응 양상이 나타날 수 있음을 보고하였다(Eckes, 2005: 216-217). 이는 평가자의 채점 특성을 총점 수준에서만 아니라 평가 준거별 양상까지 함께 살펴볼 필요가 있음을 시사한다. 평가자 효과는 평가자의 전반적인 채점 성향뿐 아니라 평정 척도와 평가 준거를 사용하는 방식에서도 다양한 유형으로 나타날 수 있다.<sup>1)</sup>

쓰기 평가 연구에서는 이러한 평가자 효과를 분석하기 위하여 다국면 라쉬 모형(Many-Facet Rasch Model, MFRM)을 비롯한 심리측정 모형이 널리 활용되어 왔다. 다국면 라쉬 모형은 평가자의 엄격성을 추정하여 평가자 간 채점 성향의 차이를 비교하고, 내적합과 외적합 통계량을 통해 채점 결과가 모형에서 기대되는 반응 양상에 부합하는 정도를 검토하는 데 유용하다.

---

1) 평가자 효과의 대표적인 유형으로는 ‘엄격성(severity)’과 ‘관대성(leniency)’, ‘중앙화 경향(centrality)’과 ‘극단화 경향(extremity)’, ‘후광 효과(halo effect)’ 등이 있다. 엄격성과 관대성은 평가자가 다른 평가자 집단이나 기준 평정에 비해 지속적으로 낮거나 높은 점수를 부여하는 경향을 의미하며(Eckes & Jin, 2021: 67), 중앙화 경향은 평정 척도의 중간 범주를 주로 사용하는 현상, 극단화 경향은 양극단 범주를 상대적으로 자주 사용하는 현상을 의미한다(Eckes & Jin, 2021: 66). 후광 효과는 평가자가 각 평가 준거를 충분히 구별하지 않고 전반적인 인상을 여러 준거에 동일하게 반영하여 준거별 점수 간 상관이 과도하게 높아지는 현상을 말한다(Leckie & Baird, 2011: 400).

그러나 이러한 측정학적 분석은 평가자 효과의 존재와 정도를 확인할 수 있을 뿐, 그러한 차이가 어떠한 판단 과정을 거쳐 형성되었는지를 직접 설명하지는 못한다(Li, Liao, & Jia, 2025: 2).

최근의 평가자 효과 연구는 이러한 한계를 보완하기 위해 평가자 효과를 점수에서 나타나는 측정학적 특성으로만 이해하기보다, 평가자가 수행을 어떻게 이해하고 평가 준거를 어떻게 해석하여 적용하는지에 관심을 확대하고 있다. 평가자의 경험(예: 훈련된 평가자 vs. 초보 평가자), 평가 성향(예: 형식 중심 vs. 내용 중심), 언어적 배경(예: 피평가자의 언어적 특성에 대한 친숙성) 등 다양한 요인이 평가 준거의 해석과 적용 방식에 영향을 미칠 수 있으며(Yan, 2014: 503), 이러한 판단 과정은 동일한 평가 루브릭을 사용하더라도 서로 다른 평가 결과가 나타나는 이유를 설명하는 중요한 단서가 된다.

## 2. 평가자의 엄격성과 평가 준거 해석

평가자 효과 연구에서 가장 핵심적으로 다루어져 온 특성은 평가자의 엄격성이다. 엄격성은 평가자가 다른 평가자에 비해 전반적으로 낮은 점수를 부여하는 안정적인 채점 성향을, 관대성은 상대적으로 높은 점수를 부여하는 채점 성향을 의미한다(Eckes & Jin, 2021: 67). 이러한 특성은 동일한 수행에 대해서도 평가자 간 점수 차이가 발생하는 이유를 설명하는 중요한 요인으로 인식되어 왔으며, 다국면 라쉬 모형을 활용한 연구에서 대표적인 분석 대상으로 다루어져 왔다.

그러나 평가자 효과를 평가자의 엄격성만으로 충분히 설명하기는 어렵다. 동일한 수준의 엄격성을 보이는 평가자라도 평가 준거를 어떻게 이해하고 적용하는지에 따라 서로 다른 평가 판단을 내릴 수 있기 때문이다. Heidari et al.(2022: 2)은 평가자의 채점을 단순한 점수 부여 과정이 아니라 학생의 수행을 평가 준거와 연결하여 의미를 구성하는 인지적 판단 과정으로 설명하였다. 평가자는 글을 읽으면서 다양한 텍스트 특성 가운데 어떤 요소를 특정 평가

준거의 증거로 볼 것인지 판단하고, 여러 수행 특성을 종합하여 최종 점수를 결정한다. 따라서 동일한 평가 루브릭을 사용하더라도 평가자가 평가 준거의 의미와 적용 범위를 어떻게 이해하는지에 따라 평가 결과는 달라질 수 있다.

쓰기 평가는 일반적으로 ‘내용, 조직, 표현’과 같은 복합적인 구성 개념을 토대로 이루어진다. 논증적 글쓰기의 내용 영역에는 주장, 이유와 근거, 반론 고려 등이, 조직 영역에는 글과 문단의 구조 및 전개가, 표현 영역에는 문장과 어휘의 적절성 및 문법적 정확성 등이 포함된다(민병곤·남가영·김선희 외, 2023; 박혜영·이상하·이명진 외, 2021; 장성민, 2026). 이러한 평가 준거는 실제 글에서 서로 긴밀하게 연결되어 나타나며, 대부분 질적 판단을 요구한다. 따라서 평가자는 루브릭을 기계적으로 적용하기보다 자신의 전문적 판단을 바탕으로 평가 준거의 의미를 해석하고 이를 텍스트의 특성과 연결하여 점수를 부여하게 된다. 이 과정에서 어떤 텍스트 특성을 중요하게 인식하는지, 여러 수행 특성을 어떻게 종합하는지에 따라 동일한 루브릭에서도 평가자 간 차이가 나타날 수 있다.

최근의 평가자 효과 연구는 이러한 평가 준거 해석 과정을 중요한 연구 대상으로 다루고 있다. Heidari et al.(2022)은 사고구술과 면담을 통해 동일한 루브릭을 사용한 평가자들도 텍스트에서 주목하는 특성과 평가 준거를 적용하는 방식이 서로 다름을 확인하였다. 국내에서도 송지언(2016)은 평가자의 총평과 채점 근거를 분석한 결과, 동일한 글을 평가하면서도 평가자마다 중요하게 인식하는 텍스트 특성과 평가 근거가 다르게 나타남을 보고하였다. 이러한 연구들은 평가자 효과를 이해하기 위해서는 평가자의 전반적인 엄격성과 함께 평가 준거를 어떻게 해석하고 적용하는지도 함께 살펴볼 필요가 있음을 시사한다.

### 3. 평가자 배경 요인과 평가자 효과

평가자 효과를 이해하기 위해서는 평가자의 채점 성향뿐 아니라 그러한 판단과 관련된 배경 요인도 함께 고려할 필요가 있다. 평가자의 성별, 전공

분야, 언어적 배경, 교육 및 평가 경험 등 다양한 특성이 평가자의 판단과 관련될 수 있는 요인으로 제시되어 왔으며, 최근에는 평가자 효과를 개인의 특성뿐 아니라 평가자가 속한 교육적 맥락과 전문성의 산물로 이해하려는 관점이 확대되고 있다. 다만 이러한 배경 요인이 평가자 효과와 어떠한 관련성을 가지는지에 대해서는 연구마다 일관된 결과가 나타나지 않았다.

성별과 언어적 배경 등 인구학적 특성을 중심으로 한 연구에서도 일관된 결론은 확인되지 않았다. Soler(2000)는 성별과 근무 기관에 따른 평가자 차이를 분석하였으나 뚜렷한 차이를 확인하지 못하였으며, Johnson & Lim(2009) 역시 영어 원어민과 비원어민 평가자 간에 언어적 배경에 따른 체계적인 편향은 거의 나타나지 않는다고 보고하였다. 국내 연구에서도 성별에 따른 평가자 효과는 연구마다 상이한 결과가 보고되었다(박영민·최숙기, 2010ㄱ, 2010ㄴ; 정미경, 2011). 이러한 결과는 성별과 같은 인구학적 특성만으로 평가자 효과를 설명하기 어렵다는 점을 시사한다.

평가자의 교육 경험과 전문적 배경 역시 평가자 효과와의 관련성이 지속적으로 검토되어 왔으나, 그 결과는 일관되지 않았다. Leckie & Baird(2011)는 평가자의 경력이 반드시 높은 평가 정확성으로 이어지지 않으며 평가자마다 엄격성의 차이가 나타날 수 있음을 보고하였고, Palermo(2022)는 일반적인 학력이나 경력보다 채점자 자격 검사에서의 수행과 지속적인 채점 모니터링이 평가 정확성을 더 잘 설명한다고 보고하였다. 반면 국내 연구에서는 쓰기 지도 경험과 평가 경험이 평가 준거 적용 방식과 관련될 수 있다는 결과가 제시되었으며(박영민·최숙기, 2009, 2010ㄴ; 백유진·가은아, 2021), 송지연(2016)은 동일한 글을 평가하더라도 평가자의 전문적 배경에 따라 중요하게 인식하는 텍스트 특성과 평가 근거가 달라질 수 있음을 확인하였다. 이러한 연구들은 평가자 효과를 단순한 경력이나 전문성의 수준으로 설명하기보다, 평가자가 축적해 온 교육 경험과 실제 평가 맥락을 함께 고려할 필요가 있음을 보여준다.

이러한 관점에서 담당 학교급은 평가자의 교육 경험을 반영하는 중요한

배경 요인으로 가정할 수 있다. 학교급에 따라 교육과정의 목표, 학습자의 발달 수준, 쓰기 교육의 내용과 평가의 목적이 달라지며, 평가자는 특정 학교급의 학생을 지속적으로 지도하고 평가하는 과정에서 해당 수준의 수행에 대한 기대와 판단 기준을 형성하게 된다. 이러한 교육 경험은 동일한 글을 평가하는 상황에서도 평가 준거를 해석하고 적용하는 방식과 관련될 가능성이 있다.

이처럼 평가자 효과는 성별과 같은 인구학적 특성뿐 아니라 담당 학교급과 같은 교육적 맥락과도 관련될 가능성이 있다. 또한 이러한 관련성은 평가자의 전반적인 엄격성뿐 아니라 평가 준거를 해석하고 적용하는 방식에서도 나타날 수 있다. 따라서 평가자 효과를 보다 심층적으로 이해하기 위해서는 평가자의 배경 요인과 함께 엄격성 및 평가 준거 해석을 통합적으로 살펴볼 필요가 있다.

### III. 연구 방법

#### 1. 참여자

이 연구에는 중등 교사 및 대학교 글쓰기 교강사 등 총 60명이 참여하였다. 중등 교사는 모두 현직 국어교사로 구성하였으며, 대학교 집단은 대학에서 글쓰기 과목을 담당한 경험이 있거나 국어를 전공한 교강사로 선정하였다. 학교급에 따른 채점 특성을 비교하기 위하여 중학교, 고등학교, 대학교 소속 평가자를 각각 20명씩 표집하였으며, 학교급 내 성별 분포의 영향을 통제하기 위해 각 집단을 남성 10명, 여성 10명으로 구성하였다.<sup>2)</sup> 이에 따라 전체 참여자는 남성 30명, 여성 30명으로 균형을 이루었다. 연령은 30대 이하 24명, 40대 21명, 50대 이상 15명이었다.

---

2) 참여자가 현재 담당하는 교육 단계에서 일정한 교육 경험을 형성하였다고 볼 수 있도록, 최근 1년 이상 동일한 학교급에서 학생을 지도하거나 글쓰기 과목을 담당한 평가자를 표집하였다.

참여자는 학교급별 표본 수를 확보하기 위하여 연구자 네트워크를 통해 모집하였다. 이 연구는 기관 단위의 조직적 표집이 아닌 개인 단위의 참여자를 대상으로 하였기 때문에 확률 표집을 적용하기는 어려웠다. 그러나 참여자를 모집하는 과정에서 개인의 채점 성향이나 평가 특성을 기준으로 선발하지 않았으며, 연구자는 참여자의 엄격성이나 평가 준거 해석 특성을 사전에 알 수 없는 상태에서 모집을 진행하였다. 따라서 이 연구의 표집은 편의 표집의 성격을 가지지만, 특정한 채점 성향을 가진 평가자가 의도적으로 포함될 가능성은 제한적이라고 판단하였다.

이 연구에서 초등학교 교사는 연구 대상에서 제외하였다. 이는 이 연구에서 사용한 평가 대상 글이 약 1,000자 내외의 논증적 글쓰기 과제로 구성되어 있어, 이러한 과제를 실제 수업에서 활용하고 평가하는 경험이 초등학교 교육 맥락에서는 실제적이지 않다고 판단하였기 때문이다. 학교급 간 비교의 타당성을 확보하기 위해 평가 과제와 실제 쓰기 평가 경험이 비교적 유사한 중학교, 고등학교 및 대학교를 연구 대상으로 선정하였다.

한편 이 연구에서는 참여자에게 별도의 평가자 훈련을 실시하지 않았다. 평가자 훈련은 일반적으로 평가 준거에 대한 공통된 이해를 형성하고 채점의 일관성을 높이기 위한 절차이지만, 이 연구의 목적은 평가자의 자연스러운 엄격성과 평가 준거 해석의 특성을 분석하는 데 있다. 사전 훈련을 실시할 경우 참여자 간 평가 준거에 대한 해석이 수렴되어 학교급에 따른 고유한 채점 특성이 약화될 가능성이 있다. 이에 따라 참여자에게는 평가 과제와 평가 루브릭만 동일하게 제공하고, 별도의 협의나 채점 기준 보정 없이 독립적으로 채점을 수행하도록 하였다.

## 2. 평가 자료

이 연구에서는 국립국어원이 구축하여 “모두의 말뭉치”에 공개한 2023년 글쓰기 원시 자료 및 채점 자료 말뭉치(민병곤 외, 2023)에서 평가 대상

글을 선정하였다. 이 말뭉치는 전국 9개 권역 국공립대학교 학생들이 2023년에 작성한 약 1,000자 내외의 논증적 글쓰기 자료와 이에 대한 채점 결과를 포함하고 있으며, 국내에서 구축·공개된 대표적인 대규모 논증적 글쓰기 말뭉치 가운데 하나이다.

해당 말뭉치에는 총 9개 문항에 대해 작성된 9,371편의 글이 포함되어 있다. 이 연구에서는 이 가운데 794편으로 구성된 Q1 문항의 응답을 분석 대상으로 하였다. Q1 문항은 ‘로봇세 도입에 대한 자신의 의견을 논리적으로 제시하라’는 과제로, 작성 조건으로 ① 서론·본론·결론을 갖춘 완결된 글을 작성할 것(제목은 쓰지 말 것), ② 분량은 1,000자 내외, ③ 작성 시간은 90분임을 제시하고 있다. 이 문항은 일정 수준 이상의 분량과 논증 구조를 요구하는 과제로서 내용, 조직, 표현 등 다양한 쓰기 평가 준거를 적용하기에 적합하다고 판단하였다.

평가 대상 글은 총 20편을 선정하였다. 평가 대상 글의 수는 평가자의 채점 부담을 고려하는 동시에 다양한 수행 수준에서 안정적인 채점 특성을 관찰할 수 있도록 결정하였다. 글을 선정하는 과정에서는 특정 장르적 특성(예: 서사적 논증)이나 형식적 결함(예: 문단 구분이 없는 글)으로 인해 평가자 간 해석 차이가 과도하게 발생하는 사례를 제외하고, 일반적인 논증적 글쓰기의 특성을 갖춘 전형적인 글을 위주로 선별하였다. 이는 평가 대상 글의 특수성보다 평가자의 채점 특성이 보다 분명하게 드러나도록 하기 위한 것이다.

또한 평가 대상 글이 특정 수행 수준에 편중되지 않도록 말뭉치에 제시된 평균 점수를 기준으로 수행 수준을 네 구간으로 구분하고, 각 구간에서 5편씩 총 20편을 선정하였다. 총점을 기준으로 평균 점수 3.5점 이상, 3.0점 이상 3.5점 미만, 2.5점 이상 3.0점 미만, 2.5점 미만의 글을 각각 5편씩 포함하여 다양한 수행 수준을 균형 있게 반영하고자 하였다.

평가 루브릭은 민병곤 외(2023)의 평가 준거를 바탕으로 일부 수정하여 사용하였다. 최종 루브릭은 ‘내용’ 영역 3개 준거, ‘조직’ 영역 2개 준거, ‘표현’ 영역 2개 준거로 구성된 총 7개의 평가 준거를 포함하였다. 이러한 구성

은 국립국어원의 말뭉치와 한국지능정보사회진흥원이 구축·공개한 대규모 글쓰기 데이터셋, 그리고 한국교육과정평가원(박혜영 외, 2021)의 평가 준거 체계와의 대응성을 고려하여, 후속 연구에서 관련 자료의 활용 가능성을 높이기 위한 목적에서 이루어졌다(〈표 1〉 참조).

평가 루브릭은 민병곤 외(2023: 49-50)의 형식을 따랐으며, 각 평가 준거마다 두 개의 하위 준거를 설정하고, 각 점수 수준에서는 두 하위 준거의 수행 수준 조합을 종합적으로 반영한 기술문을 제시하였다(〈표 2〉 참조). 이를 통해 평가자가 각 준거를 동일한 형식과 절차에 따라 적용할 수 있도록 하였다.

〈표 1〉 평가 준거 설정

|    | 국립국어원(민병곤 외, 2023; 장성민 외, 2024) | 한국지능정보사회진흥원                | 한국교육과정평가원(박혜영 외, 2021)              | 현재 연구     |
|----|---------------------------------|----------------------------|-------------------------------------|-----------|
| 내용 | •주장                             | •주제의 명료성                   | •주장 및 근거의 명료성과 타당성<br>•주장 및 근거의 구체성 | •주장       |
|    | •이유/근거의 적절성<br>•이유/근거의 충분성      | •근거의 적절성                   |                                     | •이유와 근거   |
|    | •다른 입장에 대한 고려                   | •사고의 창의성                   |                                     | •반론 고려    |
|    | •문제 상황 제시                       |                            |                                     |           |
|    |                                 | •프롬프트 독해력                  | •과제 수행의 충실성<br>(※ '과제' 영역으로 별도 설정)  |           |
| 조직 | •글 전체 조직                        | •문단 간 구조의 적절성<br>•구조의 일관성  | •글의 통일성                             | •글 전체 조직  |
|    | •문단 내 조직                        | •문단 내 구조의 적절성              | •글의 체계성                             | •문단 내 조직  |
|    |                                 | •분량의 적절성                   |                                     |           |
| 표현 | •문장과 어휘                         | •단어 사용의 적절성<br>•문장 표현의 적절성 | •어휘 및 문장의 적절성                       | •문장과 어휘   |
|    | •어문 규범과 관습                      | •문법의 적절성                   | •어법의 정확성                            | •어문 규범 준수 |

〈표 2〉 평가 루브리

| 영역 / 준거  | 내 용                                                                                                                                                      |                                                                                                                                                                                      |                                                                                                                                                               | 조 직                                                                                                                                                                            |                                                                                                                                                                     | 표 현                                                                                                                                                                     |                                   |
|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
|          | 내용 1<br>수상                                                                                                                                               | 내용 2<br>이유와 근거                                                                                                                                                                       | 내용 3<br>반론 고려                                                                                                                                                 | 조직 1<br>클 전제 조직                                                                                                                                                                | 조직 2<br>문단 내 조직                                                                                                                                                     | 표현 1<br>문장과 어휘                                                                                                                                                          | 표현 2<br>어문 규범 준수                  |
| 하위<br>준거 | <ul style="list-style-type: none"><li>• <b>논제 포함성</b><br/>: 주제가 논제에 부합하는가?</li><li>• <b>명제성 및 일관성</b><br/>: 주제가 명확하게 제시되고 글 전체에 걸쳐 일관되게 유지되는가?</li></ul> | <ul style="list-style-type: none"><li>• <b>이론의 직접성 및 충분성</b><br/>: 주장을 뒷받침하는 이유가 직접하며 충분한 수로 제시되는가?</li><li>• <b>근거의 구체성 및 타당성</b><br/>: 각 이유를 뒷받침하는 근거가 구체적이며 논리적으로 타당한가?</li></ul> | <ul style="list-style-type: none"><li>• <b>대안적 입장의 탐구</b><br/>: 다른 입장에 대한 접근이 이루어졌는가?</li><li>• <b>반응의 직접성</b><br/>: 다른 입장에 대해 논리적이고 타당하게 대응하고 있는가?</li></ul> | <ul style="list-style-type: none"><li>• <b>문단 구성의 직접성</b><br/>: 하나의 중심 내용을 유지하면서 이를 충분히 전개하고 있는가? (내용, 형식 모두 포함)</li><li>• <b>문단 배열의 용이성</b><br/>: 문단 배열(전개)이 체계적인가?</li></ul> | <ul style="list-style-type: none"><li>• <b>문장 표현의 직접성</b><br/>: 문장 길이가 적절하고 그 표현이 자연스러운가? (예: 주술호용)</li><li>• <b>어휘 사용의 직접성</b><br/>: 어휘 사용이 의미 전달에 적합한가?</li></ul> | <ul style="list-style-type: none"><li>• <b>태어났을 및 표기</b><br/>: 맞춤법과 표기 규칙을 정확하게 준수하였는가? (오탈자 여부 포함)</li><li>• <b>타어났기 및 문장 부호</b><br/>: 띄어쓰기와 문장 부호 사용이 적절한가?</li></ul> |                                   |
| 5        | 논제에 정합하는 주장이 포함되어 제시되며, 글 전체에 걸쳐 일관되고 분명하게 유지됨.                                                                                                          | 두 가지 이상의 직접한 이유를 제시하고 각 이유에 대해 구체적이고 타당한 근거를 들어 충분히 논증함.                                                                                                                             | 다른 입장을 깊이 있게 탐구하고, 이에 대해 논리적이고 타당하게 대응함.                                                                                                                      | 서론, 본론, 결론이 내용과 형식 측면에서 명확히 구분되며, 전체 전개가 매우 논리적이고 체계적임.                                                                                                                        | 각 문단이 하나의 중심 내용을 일관되게 유지하면서 이를 충분히 전개하고 있음. 문장 간 연결이 매우 자연스럽고 논리적임.                                                                                                 | 문장이 명확하고 자연스럽게 표현되며, 의미에 적합한 어휘가 정확하게 사용됨.                                                                                                                              | 맞춤법, 표기, 띄어쓰기 문장 부호를 매우 정확하게 사용함. |
| 4        | 논제에 부합하는 주장이 비교적 명확하게 제시되며, 전반적으로 일관되게 유지됨.                                                                                                              | 두 가지 이상의 직접한 이유를 제시하고 각 이유에 대해 비교적 타당한 근거를 들어 논증함.                                                                                                                                   | 다른 입장을 제시하고 이에 대해 비교적 직접하게 대응함.                                                                                                                               | 서론, 본론, 결론의 구조가 비교적 분명하며, 전체 전개가 대체로 논리적이고 안정적인 구조로 이루어짐.                                                                                                                      | 각 문단이 하나의 중심 내용을 유지하며 대체로 충분히 전개되고 있고, 문장 간 연결도 대체로 자연스럽고 논리적임.                                                                                                     | 문장 표현과 어휘 사용이 대체로 적절하며, 전반적으로 자연스럽게 표현됨.                                                                                                                                | 규범 사용이 대체로 적절하며, 오류가 거의 없음.       |
| 3        | 논제에 부합하는 주장이 제시되었으나, 다소 모호하거나 일부에서 일관성이 약함.                                                                                                              | 주장을 뒷받침하는 직접한 이유를 1가지 제시하였으며, 이를 비교적 구체적이고 타당하게 논증함.                                                                                                                                 | 다른 입장을 언급하였으나, 탐구나 대응이 제한적임.                                                                                                                                  | 기본적인 구조는 있으나 일부 부분이 모호하거나 전개가 다소 불안정함.                                                                                                                                         | 문단의 중심 내용도 대체로 유지되나 전개가 다소 부분적으로 자연스럽지 못함.                                                                                                                          | 문장 표현이나 어휘 선택이 다소 부정확하거나 단조로움.                                                                                                                                          | 일부 오류가 있으나, 이해에는 큰 지장이 없음.        |
| 2        | 주장이 불명확하거나 논제와의 관련성이 약하며, 일관성이 부족함.                                                                                                                      | 주장을 뒷받침하는 직접한 이유를 1가지 제시하였으나, 근거가 부족하거나 논증이 미흡함.                                                                                                                                     | 다른 입장을 형식적으로만 언급하거나, 대응이 부적절함.                                                                                                                                | 구조가 불명확하거나 전개가 일관되지 않음.                                                                                                                                                        | 문단의 중심 내용이 충분히 유지되나 전개가 미흡하며, 문장 간 연결이 부자연스러움.                                                                                                                      | 부적절한 어휘나 부자연스러운 문장이 자주 발견됨.                                                                                                                                             | 오류가 빈번하여 이해에 다소 어려움이 있음.          |
| 1        | 주장이 제시되지 않았거나 논제에서 벗어나 있음.                                                                                                                               | 적절한 이유나 근거가 제시되지 않음.                                                                                                                                                                 | 다른 입장을 전혀 고려하지 않음.                                                                                                                                            | 글의 전체 구조를 파악하기 어려움.                                                                                                                                                            | 문단 내 조직이 거의 이루어지지 않음.                                                                                                                                               | 문장과 어휘 사용이 매우 부적절하며 의미 전달이 어려움.                                                                                                                                         | 규범 오류가 매우 많아 이해가 어려움.             |

### 3. 평가 데이터 수집

자료 수집은 참여자별 개별·독립 채점 방식으로 이루어졌다. 모든 참여자에게 동일한 평가 자료와 평가 루브릭을 제공하여 평가 조건을 표준화하되, 실제 채점은 각 평가자가 평소 수행하는 방식에 따라 자율적으로 이루어지도록 함으로써 자연스러운 채점 특성이 반영될 수 있도록 하였다.

자료 배부 시에는 평가 루브릭과 평가 대상 글 20편을 하나의 PDF 파일로 제작하여 참여자에게 제공하였다. 평가 대상 글은 E01~E20으로 번호를 부여하여 모든 참여자에게 동일한 순서로 제시하였으며, 채점 결과는 별도로 제공한 엑셀 양식에 입력한 후 회신하도록 하였다. PDF 파일은 화면에서 열람할 수 있을 뿐 아니라 인쇄가 가능한 형태로 제공하였다. 이에 따라 참여자는 자료를 출력하여 메모하거나 여러 편의 글을 나란히 비교하는 등 자신의 평소 채점 방식에 맞는 환경에서 자유롭게 평가를 수행할 수 있도록 하였다.

평가 기간은 자료 배부 후 14일로 설정하였다. 다만 평가자의 내적 일관성을 유지하기 위하여 여러 날에 걸쳐 조금씩 채점하기보다는 일정 시간을 확보하여 가능한 한 한 차례의 연속된 평가로 완료해 줄 것을 요청하였다. 또한 채점이 진행되면서 평가 준거에 대한 이해와 적용 방식이 달라질 수 있음을 고려하여, 필요하다고 판단되는 경우에는 앞서 평가한 글을 다시 검토하고 점수를 수정할 수 있음을 함께 안내하였다.

회수된 자료는 모든 평가 대상 글과 평가 준거에 대해 점수가 입력되었는지를 확인하여 결측치 여부를 점검하였다. 결측이 발견된 경우에는 해당 참여자에게 확인을 요청하여 자료를 보완하였으며, 최종 자료에서는 결측치가 발생하지 않았다. 회수된 채점 결과를 검토한 결과, 특정 평가자가 전반적으로 지나치게 높거나 낮은 점수를 부여하는 등 분석에 영향을 미칠 수준의 극단적인 채점 양상은 확인되지 않았다.

#### 4. 평가 데이터 분석

이 연구에서는 두 개의 연구 문제와 다섯 개의 하위 연구 문제에 따라 평가 대상 글의 수행 수준, 평가 준거의 측정 특성, 평가자의 채점 특성 및 평가자 특성에 따른 채점 결과의 차이를 분석하였다.

먼저 연구 문제 1-1(평가 대상 글의 수행 수준)을 확인하기 위해 글별 및 평가 준거별 평균과 표준편차를 산출하였다. 각 글의 총점은 7개 평가 준거 점수의 평균으로 산출하였으며, 평균과 표준편차를 중심으로 평가 대상 글의 수행 수준과 평가 준거별 점수 분포를 기술하였다.

다음으로 연구 문제 1-2(평가 준거의 측정 특성)를 확인하기 위해 문항별 난이도와 변별도를 동시에 추정할 수 있는 다범주 문항반응모형인 일반화 부분점수모형(Generalized Partial Credit Model, GPCM)을 적용하였다. 분석은 R의 mirt 패키지를 이용하여 수행하였으며, 60명의 평가자가 20편의 글을 7개 평가 준거에 따라 채점한 자료를 활용하였다. 일반화 부분점수모형은 평가 준거의 상대적인 난이도와 변별도를 탐색하기 위해 적용하였으며, 분석 결과는 준거 간 상대적 측정 특성을 중심으로 해석하였다.<sup>3)</sup>

연구 문제 1-3(평가자의 채점 특성)을 확인하기 위해 다국면 라쉬 모형을 적용하였다. 모형은 평가 대상 글, 평가 준거, 평가자의 세 국면으로 구성하였으며, 평가 대상 글을 측정 대상, 7개 평가 준거를 문항으로 설정하였다. 분석은 R의 TAM 패키지를 이용하여 수행하였으며, 평가자별 엄격성 모수(logit)와 표준오차를 추정하였다. 또한 내적합(Infit) 평균제곱과 외적합(Outfit) 평균제곱을 산출하여 각 평가자의 채점 적합도를 검토하였다.

마지막으로 연구 문제 2-1(평가자 특성에 따른 평가자 효과)을 확인하기

---

3) 동일한 평가자가 20편의 글에 반복 응답하였으므로 지역독립성(local independence) 가정이 엄밀히 충족되지는 않으며, 이로 인해 준거별 난이도·변별도 추정치의 표준오차가 과소추정되었을 가능성이 있다. 이에 따라 준거 간 비교는 확정적 결론이 아닌 잠정적 경향으로 해석하였다.

위하여 성별과 담당 학교급에 따른 차이를 평가자의 전반적인 엄격성과 평가 준거별 채점 양상의 두 차원에서 분석하였다. 먼저 다국면 라쉬 모형에서 추정된 평가자별 엄격성(logit)을 대상으로 성별에 따른 차이는 독립표본 t 검정, 담당 학교급에 따른 차이는 일원분산분석(one-way ANOVA)을 실시하였으며, 학교급 간 차이가 통계적으로 유의한 경우에는 Tukey 사후검정을 실시하였다. 이와 함께 각 평가자가 20편의 글에 부여한 총점과 7개 평가 준거별 평균 점수를 대상으로 동일한 방법으로 원점수 분석을 실시하였다. 원점수 분석은 실제 평정 척도에서 나타나는 집단 간 차이의 크기와 방향을 확인하고 후속 면담에서 평가 준거별 판단 차이를 탐색하기 위한 근거를 마련하는 한편, 엄격성 추정치를 이용한 분석과 결과가 수렴하는지를 확인하기 위해 활용하였다.

또한 담당 학교급에 따른 차이가 평가자의 전반적인 엄격성보다 특정 평가 준거에 대한 차별적인 반응에서 나타나는지를 확인하기 위하여 다국면 라쉬 모형에서 담당 학교급과 평가 준거 간 편향 상호작용을 추가로 분석하였다. 담당 학교급은 개별 평가자에 종속된 배경 변인이므로 이를 평가자 국면과 독립된 주효과로 동시에 추정하지 않고, 평가자별 전반적인 엄격성을 통제된 상태에서 학교급에 따라 특정 평가 준거를 상대적으로 엄격하거나 관대하게 적용하는지를 검토하였다. 편향 상호작용의 통계적 유의성은 Wald 검정을 통해 확인하였으며, 전체 상호작용과 평가 준거별 학교급 차이를 각각 검정하였다. 이를 통해 원점수에서 확인된 평가 준거별 집단 차이가 평가 대상 글, 평가 준거 및 평가자의 전반적인 엄격성을 고려한 이후에도 나타나는지를 확인하였다.

## 5. 후속 면담

후속 면담은 연구 문제 2-2(평가 준거 해석의 차이)를 확인하기 위해 실시하였다. 양적 분석에서 학교급에 따라 일부 평가 준거의 평균 점수에 유의

한 차이가 나타남에 따라, 이러한 차이가 나타난 원인을 심층적으로 탐색하기 위해 평가자 초점집단면담(Focus Group Interview, FGI)을 실시하였다. 이 연구는 양적 분석 결과를 바탕으로 질적 자료를 수집하여 결과를 해석하는 설명적 순차 설계(explanatory sequential design)의 혼합연구 방법을 적용하였다. FGI는 양적 분석에서 확인된 학교급별 차이를 설명하기 위한 탐색적 절차로 실시하였으며, 학교급 전체의 특성을 일반화하기보다 참여자의 평가 준거 해석 양상을 심층적으로 이해하는 데 목적을 두었다.

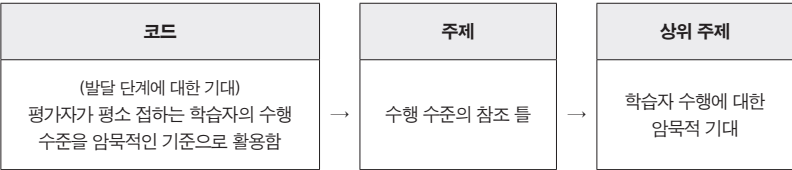
FGI 참여자는 학교급에 따른 평가 준거 해석의 차이를 비교하기 위한 목적 표집(purposive sampling) 방식으로 선정하였다. 인터뷰 결과가 특정 평가자의 채점 성향에 과도하게 영향을 받지 않도록 평균적인 채점 엄격성(엄격성 추정치가 0에 가까운 수준)을 보이면서 내적합과 외적합 평균제곱 값이 모두 적합 범위에 있는 평가자를 우선 선정하였다. 또한 학교급별 교육 경험을 비교하기 위하여 중학교 평가자 2명, 고등학교 평가자 1명, 대학교 평가자 1명을 선정하였다. 최종적으로 중학교 평가자 Mm09와 Mf09, 고등학교 평가자 Hf10, 대학교 평가자 Um04가 면담에 참여하였다.

면담은 약 90분 동안 반구조화된 방식으로 진행하였다. 먼저 양적 분석에서 학교급에 따른 차이가 확인된 평가 준거와 분석 결과를 참여자들에게 제시한 후, 동일한 글에 대한 평가 판단의 근거를 중심으로 면담을 진행하였다. 면담 자료는 양적 분석 결과를 바탕으로 선정한 평가 대상 글 4편(E08, E10, E12, E13)으로 구성하였다. 대상 글은 수행 수준이 상·중상·중·중하 수준을 고르게 포함하되, 특히 평가자 간 채점 결과의 편차가 비교적 크게 나타나 평가 준거 해석의 차이가 드러날 가능성이 높은 글을 우선 선정하였다. 이를 통해 서로 다른 수행 수준의 글에서 평가 준거 해석의 차이를 비교할 수 있도록 하였다.

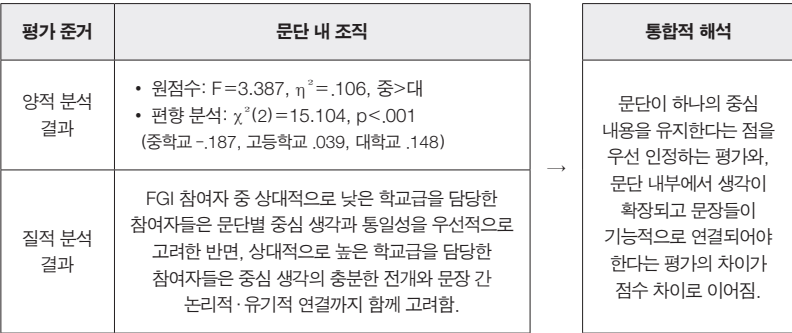
참여자들은 각 글을 함께 읽으며 ‘이유와 근거’, ‘반론 고려’, ‘글 전체 조직’, ‘문단 내 조직’, ‘문장과 어휘’ 준거에 대한 평가 근거를 설명하고, 동일한 글을 서로 다르게 평가한 이유를 중심으로 토론하였다. 또한 자신이 주로 지

도하는 학습자의 일반적인 글쓰기 특성, 학교급에서 우선적으로 형성해야 한다고 인식하는 쓰기 능력, 그리고 해당 교육 단계에서 중요하게 고려하는 평가 준거가 채점 판단에 어떠한 영향을 미쳤는지 등을 자유롭게 논의하였다.

모든 면담은 참여자의 동의를 얻어 녹음한 후 전사하였다. 전사 자료는 반복적으로 검토하면서 학교급별 교육 경험, 학습자에 대한 기대 수준, 평가 준거의 해석과 적용 방식에 관한 진술을 중심으로 개방 코딩을 실시하였다. 이후 유사한 의미 단위를 범주화하여 학교급에 따른 평가 준거 해석의 차이를 설명하는 주제를 도출하고(〈그림 1〉 참조), 이를 양적 분석 결과와 통합하여 해석하였다(〈그림 2〉 참조). 분석 과정에서는 코드와 원자료를 반복적으로 대조하고, 상위 주제가 참여자의 진술을 충분히 포괄하는지를 검토하였다.



〈그림 1〉 FGI 자료 분석 과정의 예



〈그림 2〉 학교급별 평가자 점수 차이에 대한 양적·질적 분석의 통합

## IV. 연구 결과

### 1. 평가 대상 글의 수행 수준

〈표 3〉은 평가 대상 글의 수행 수준과 평가 준거별 점수 분포를 제시한 것이다. 총점을 기준으로 평균 점수가 3.5점 이상인 글이 6편, 3.0점 이상 3.5점 미만인 글이 4편, 2.5점 이상 3.0점 미만인 글이 6편, 2.5점 미만인 글이 4편으로 나타났다. 전체적으로 다양한 수행 수준의 글이 비교적 고르게 분포하여 평가 자료가 특정 수행 수준에 편중되지 않았음을 확인할 수 있었다.

평가 준거별 평균을 살펴보면, ‘어문 규범 준수’의 평균 점수가 가장 높았으며 ‘반론 고려’의 평균 점수가 가장 낮게 나타났다. 또한 ‘이유와 근거’와 ‘문장과 어휘’ 역시 다른 평가 준거에 비해 상대적으로 낮은 평균을 보여, 이들 준거에서 평가 대상 글의 수행 수준이 상대적으로 낮게 나타났다.

〈표 3〉 글 점수(평균, 표준편차)

| 글 ID | 내용            |               |                | 조직            |               | 표현             |               | 총점            |
|------|---------------|---------------|----------------|---------------|---------------|----------------|---------------|---------------|
|      | 주장            | 이유와 근거        | 반론 고려          | 글 전체 조직       | 문단 내 조직       | 문장과 어휘         | 어문 규범 준수      |               |
| E01  | 3.57<br>(.79) | 3.33<br>(.93) | 2.83<br>(.99)  | 3.50<br>(.72) | 3.30<br>(.81) | 3.02<br>(1.07) | 3.82<br>(.60) | 3.34<br>(.50) |
| E02  | 2.73<br>(.69) | 2.58<br>(.85) | 2.75<br>(1.00) | 2.83<br>(.91) | 2.57<br>(.74) | 2.80<br>(1.05) | 3.68<br>(.68) | 2.85<br>(.44) |
| E03  | 3.60<br>(.74) | 2.93<br>(.99) | 2.93<br>(1.25) | 3.50<br>(.81) | 3.32<br>(.93) | 3.70<br>(.96)  | 3.63<br>(.80) | 3.37<br>(.53) |
| E04  | 2.17<br>(.67) | 1.92<br>(.72) | 2.02<br>(.83)  | 2.13<br>(.68) | 2.08<br>(.77) | 2.05<br>(.85)  | 3.52<br>(.68) | 2.27<br>(.46) |
| E05  | 4.02<br>(.57) | 4.27<br>(.63) | 3.57<br>(.87)  | 4.05<br>(.72) | 4.22<br>(.78) | 3.53<br>(.87)  | 3.88<br>(.56) | 3.93<br>(.43) |

|     |                |                |                |                |                |                |               |               |
|-----|----------------|----------------|----------------|----------------|----------------|----------------|---------------|---------------|
| E06 | 3.90<br>(.71)  | 4.00<br>(.90)  | 3.13<br>(1.05) | 4.02<br>(.85)  | 3.63<br>(.78)  | 3.40<br>(.96)  | 3.73<br>(.86) | 3.69<br>(.49) |
| E07 | 2.05<br>(.47)  | 1.85<br>(.88)  | 2.35<br>(.92)  | 1.90<br>(.73)  | 1.82<br>(.57)  | 2.32<br>(.79)  | 3.47<br>(.70) | 2.25<br>(.42) |
| E08 | 2.63<br>(.76)  | 2.45<br>(.79)  | 2.35<br>(.94)  | 2.67<br>(.91)  | 2.42<br>(.87)  | 2.43<br>(.87)  | 3.72<br>(.58) | 2.67<br>(.44) |
| E09 | 3.27<br>(.80)  | 3.63<br>(.82)  | 3.47<br>(.81)  | 3.55<br>(.79)  | 3.50<br>(.79)  | 2.92<br>(.77)  | 4.12<br>(.58) | 3.49<br>(.43) |
| E10 | 4.45<br>(.65)  | 4.48<br>(.70)  | 3.83<br>(.74)  | 4.37<br>(.80)  | 4.48<br>(.70)  | 3.98<br>(.89)  | 4.05<br>(.47) | 4.24<br>(.42) |
| E11 | 4.08<br>(.65)  | 4.13<br>(.72)  | 3.42<br>(.96)  | 4.13<br>(.54)  | 4.35<br>(.66)  | 3.70<br>(.81)  | 4.18<br>(.54) | 4.00<br>(.40) |
| E12 | 3.38<br>(.85)  | 3.27<br>(.86)  | 2.43<br>(.89)  | 3.35<br>(.92)  | 3.15<br>(.66)  | 2.95<br>(1.08) | 3.47<br>(.81) | 3.14<br>(.46) |
| E13 | 3.00<br>(.86)  | 2.40<br>(.85)  | 2.22<br>(.94)  | 2.88<br>(.87)  | 2.75<br>(.77)  | 3.08<br>(.83)  | 3.75<br>(.73) | 2.87<br>(.47) |
| E14 | 2.78<br>(.83)  | 2.40<br>(.87)  | 2.63<br>(.80)  | 2.63<br>(.97)  | 2.23<br>(.77)  | 2.45<br>(.81)  | 3.80<br>(.78) | 2.70<br>(.48) |
| E15 | 2.43<br>(.72)  | 2.02<br>(.85)  | 2.75<br>(.88)  | 2.20<br>(.73)  | 2.25<br>(.84)  | 2.65<br>(.80)  | 3.53<br>(.68) | 2.55<br>(.47) |
| E16 | 2.20<br>(.78)  | 1.98<br>(.95)  | 2.40<br>(1.06) | 1.95<br>(.53)  | 2.03<br>(.71)  | 2.52<br>(.89)  | 3.87<br>(.77) | 2.42<br>(.49) |
| E17 | 3.03<br>(.71)  | 2.78<br>(.87)  | 2.08<br>(.85)  | 3.10<br>(.77)  | 3.10<br>(.75)  | 2.77<br>(1.03) | 3.95<br>(.70) | 2.97<br>(.46) |
| E18 | 1.95<br>(.70)  | 1.75<br>(.77)  | 2.38<br>(.85)  | 1.68<br>(.81)  | 1.67<br>(.66)  | 2.27<br>(.82)  | 3.25<br>(.65) | 2.14<br>(.45) |
| E19 | 3.57<br>(.81)  | 3.83<br>(.78)  | 2.90<br>(.90)  | 3.78<br>(.80)  | 3.68<br>(.81)  | 3.25<br>(.77)  | 4.05<br>(.62) | 3.58<br>(.44) |
| E20 | 3.92<br>(.65)  | 3.70<br>(.91)  | 3.58<br>(.83)  | 3.97<br>(.69)  | 3.87<br>(.62)  | 3.72<br>(.90)  | 3.80<br>(.61) | 3.79<br>(.42) |
| 전체  | 3.14<br>(1.03) | 2.99<br>(1.20) | 2.80<br>(1.06) | 3.11<br>(1.13) | 3.02<br>(1.13) | 2.98<br>(1.04) | 3.76<br>(.71) | 3.11<br>(.77) |

## 2. 평가 준거의 측정 특성

〈표 4〉는 일반화 부분점수모형을 적용하여 추정한 평가 준거의 난이도와 변별도를 제시한 것이다. 난이도는 ‘반론 고려’가 ‘어려움’ 수준, ‘어문 규범 준수’가 ‘쉬움’ 수준으로 나타났으며, 나머지 5개 평가 준거는 모두 ‘보통’ 수준에 해당하였다. 또한 모든 평가 준거의 변별도는 ‘적절’ 수준으로 나타나, 7개 평가 준거 모두 수행 수준이 서로 다른 글을 안정적으로 변별하는 것으로 확인되었다.

〈표 4〉 평가 준거별 난이도, 변별도

|     | 내용    |        |       | 조직      |         | 표현     |          |
|-----|-------|--------|-------|---------|---------|--------|----------|
|     | 주장    | 이유와 근거 | 반론 고려 | 글 전체 조직 | 문단 내 조직 | 문장과 어휘 | 어문 규범 준수 |
| 난이도 | -.037 | .295   | .809  | .019    | .223    | .335   | -1.644   |
| 변별도 | 1.000 | 1.068  | .853  | 1.045   | 1.064   | .882   | .692     |

해석 기준(상태제, 2019: 313, 319)

난이도: -2.0 미만(매우 쉬움), -2.0~-.5(쉬움), -.5~.5(보통), .5~2.0(어려움), 2.0 이상(매우 어려움)

변별도: .35 미만(거의 없음), .35~.65(낮음), .65~1.35(적절), 1.35~1.70(높음), 1.70 이상(매우 높음)

## 3. 평가자의 채점 특성

〈표 5〉는 다국면 라쉬 모형을 적용하여 추정한 평가자의 엄격성과 채점 적합도를 제시한 것이다. 평가자의 엄격성은 -1.312~1.305 logit으로 약 2.62 logit 범위에 걸쳐 분포하여, 동일한 글과 평가 준거를 적용한 상황에서도 평가자 간 엄격성에 상당한 차이가 있는 것으로 나타났다. 엄격성 추정치가 클수록 상대적으로 엄격하게, 작을수록 상대적으로 관대하게 채점한 것으로 해석할 수 있다.

〈표 5〉 평가자별 채점 특성

| 평가자 ID | 엄격성<br>(logit) | 표준오차 | 내적합도<br>(Infit) | 외적합도<br>(Outfit) | 평가자 ID | 엄격성<br>(logit) | 표준오차 | 내적합도<br>(Infit) | 외적합도<br>(Outfit) |
|--------|----------------|------|-----------------|------------------|--------|----------------|------|-----------------|------------------|
| Mm01   | -.693          | .114 | .641            | .696             | Hf01   | -.003          | .110 | .913            | .917             |
| Mm02   | -.865          | .116 | .948            | .996             | Hf02   | .421           | .109 | 1.079           | 1.052            |
| Mm03   | -.551          | .113 | 1.018           | 1.132            | Hf03   | -.125          | .110 | .697            | .762             |
| Mm04   | -1.312         | .121 | 1.020           | 1.115            | Hf04   | -.015          | .110 | .532            | .544             |
| Mm05   | -.602          | .113 | 1.283           | 1.342            | Hf05   | 1.204          | .113 | 1.025           | 1.098            |
| Mm06   | 1.305          | .114 | 1.023           | 1.031            | Hf06   | -.772          | .115 | 1.307           | 1.274            |
| Mm07   | .505           | .109 | 1.264           | 1.275            | Hf07   | .204           | .109 | .740            | .722             |
| Mm08   | .252           | .109 | .747            | .759             | Hf08   | .216           | .109 | 1.065           | 1.072            |
| Mm09   | -.076          | .110 | 1.115           | 1.106            | Hf09   | -.199          | .110 | 1.079           | 1.072            |
| Mm10   | .481           | .109 | .880            | .849             | Hf10   | .058           | .109 | .927            | .905             |
| Mf01   | -.628          | .113 | .829            | .874             | Um01   | .083           | .109 | 1.182           | 1.165            |
| Mf02   | -.461          | .112 | .789            | .814             | Um02   | 1.242          | .113 | .875            | .888             |
| Mf03   | -.286          | .111 | .813            | .852             | Um03   | -.174          | .110 | .633            | .638             |
| Mf04   | .252           | .109 | 1.298           | 1.356            | Um04   | -.027          | .110 | 1.084           | 1.055            |
| Mf05   | .191           | .109 | 1.049           | 1.085            | Um05   | -1.197         | .120 | 1.290           | 1.208            |
| Mf06   | -.905          | .116 | 1.242           | 1.265            | Um06   | .348           | .109 | .843            | .825             |
| Mf07   | -1.140         | .119 | 1.261           | 1.197            | Um07   | 1.280          | .114 | .888            | .874             |
| Mf08   | 1.103          | .112 | 1.459           | 1.555            | Um08   | .143           | .109 | .959            | 1.001            |
| Mf09   | .022           | .109 | 1.074           | 1.085            | Um09   | .469           | .109 | .931            | .927             |
| Mf10   | .119           | .109 | .899            | .911             | Um10   | .167           | .109 | 1.057           | 1.090            |
| Hm01   | -.436          | .112 | .623            | .656             | Uf01   | -.113          | .110 | .914            | .895             |
| Hm02   | -.039          | .110 | 1.119           | 1.118            | Uf02   | -.772          | .115 | 1.276           | 1.326            |
| Hm03   | -1.070         | .118 | 1.064           | 1.071            | Uf03   | -.298          | .111 | .693            | .693             |
| Hm04   | -.311          | .111 | .612            | .627             | Uf04   | 1.153          | .112 | .969            | .981             |
| Hm05   | -.064          | .110 | 1.201           | 1.156            | Uf05   | -.323          | .111 | 1.431           | 1.565            |
| Hm06   | .674           | .110 | .867            | .884             | Uf06   | -.905          | .116 | 1.120           | 1.092            |

|      |        |      |       |       |      |      |      |       |       |
|------|--------|------|-------|-------|------|------|------|-------|-------|
| Hm07 | -1.312 | .121 | 1.191 | 1.178 | Uf07 | .918 | .111 | 1.061 | 1.052 |
| Hm08 | 1.254  | .113 | .888  | .946  | Uf08 | .276 | .109 | .992  | .954  |
| Hm09 | .022   | .109 | .817  | .824  | Uf09 | .360 | .109 | 1.028 | 1.076 |
| Hm10 | .397   | .109 | .855  | .878  | Uf10 | .553 | .109 | .987  | .988  |

M(중학교), H(고등학교), U(대학교), m(남성), f(여성)

**내적합도·외적합도 해석 기준**(Linacre, 2002: 878)

.5 미만(과적합), .5~1.5(적합), 1.5~2.0(주의 필요), 2.0 초과(부적합)

내적합 평균제곱은 모든 평가자가 .5~1.5의 적합 범위에 포함되었다. 외적합 평균제곱 역시 대부분 적합 범위에 있었으나 Mf08(1.555)와 Uf05(1.565)는 1.5를 초과하여 주의가 필요한 수준으로 나타났다. 다만 외적합 평균제곱이 2.0을 초과한 평가자는 없었으며, 두 평가자 모두 내적합 평균제곱은 적합 범위에 포함되어 있었다. 따라서 일부 글에서 모형의 기대와 다소 다른 평정이 나타났으나, 전체적인 채점 양상은 대체로 측정모형에 수용 가능한 수준으로 판단하였다.<sup>4)</sup>

평가자 간 엄격성 차이가 추정오차를 넘어 안정적으로 구별되는지를 확인하기 위하여 평가자 국면의 분리 지표와 동질성 검정을 추가로 살펴보았다. 평가자 엄격성 추정치의 표준편차는 .677, 추정 표준오차의 RMSE는 .112였으며, 이에 따른 분리 지수는 5.980, 분리 신뢰도는 .973으로 나타났다. 또한 약 8.31개의 구별 가능한 엄격성 층위가 확인되었으며, 평가자의 엄격성이 동일하다는 가설에 대한 동질성 검정 결과도 통계적으로 유의하였다( $\chi^2(59)=2059.81$ ,  $p<.001$ ).<sup>5)</sup>

4) 외적합 평균제곱 값만 높은 경우는 전반적인 채점 일관성은 유지되었으나, 모형의 기대에서 크게 벗어난 일부 글에서 예외적인 채점이 나타났음을 의미한다. 반면 채점 전반에 걸쳐 일관성이 부족한 경우에는 내적합 평균제곱 값도 함께 증가한다. 따라서 일반적으로는 외적합 평균제곱보다 내적합 평균제곱 지표를 채점 적합성 판단에서 더 중요하게 활용한다.

5) 이는 이 연구에서 확인된 평가자 간 엄격성의 차이가 단순한 측정오차를 넘어 안정적으로 구별되는 수준임을 보여준다. 다만 평가자 국면에서 높은 분리 신뢰도는 평가자 간 일

4. 평가자 특성에 따른 평가자 효과

1) 성별에 따른 평가자 효과

〈표 6〉은 평가자의 성별에 따른 엄격성과 원점수 채점 결과를 비교한 것이다. 먼저 다국면 라쉬 모형에서 추정한 평가자의 엄격성은 남성 평가자가 평균  $-.004$  logit( $SD=.749$ ), 여성 평가자가  $.004$  logit( $SD=.611$ )으로 나타났으며, 두 집단 간 차이는 통계적으로 유의하지 않았다( $t(58)=-.040$ ,  $p=.968$ ). 이는 성별에 따라 평가자가 전반적으로 더 엄격하거나 관대하게 채점하는 경향이 체계적으로 나타나지 않았음을 의미한다.

〈표 6〉 평가자의 성별에 따른 엄격성 및 채점 결과 비교

| 구분          |    |          | 성별          |            | t     | p    |
|-------------|----|----------|-------------|------------|-------|------|
|             |    |          | 남성 M(SD)    | 여성 M(SD)   |       |      |
| 엄격성(logit)  |    |          | -.004(.749) | .004(.611) | -.040 | .968 |
| 원<br>점<br>수 | 내용 | 주장       | 3.15(.48)   | 3.13(.42)  | .200  | .842 |
|             |    | 이유와 근거   | 3.03(.54)   | 2.95(.44)  | .619  | .538 |
|             |    | 반론 고려    | 2.79(.49)   | 2.82(.49)  | -.210 | .834 |
|             | 조직 | 글 전체 조직  | 3.10(.47)   | 3.12(.42)  | -.116 | .908 |
|             |    | 문단 내 조직  | 3.03(.52)   | 3.01(.44)  | .202  | .841 |
|             | 표현 | 문장과 어휘   | 2.96(.50)   | 2.99(.52)  | -.253 | .802 |
|             |    | 어문 규범 준수 | 3.74(.48)   | 3.79(.41)  | -.403 | .688 |
|             | 총점 |          | 3.11(.43)   | 3.11(.35)  | .009  | .993 |

치도가 높다는 의미가 아니라, 평가자들의 엄격성이 서로 구별되는 정도가 높다는 의미로 해석할 필요가 있다.

원점수에서도 총점과 7개 평가 준거 모두에서 성별에 따른 차이가 통계적으로 유의하지 않았다. 총점은 남성 평가자 3.11점(SD=.43), 여성 평가자 3.11점(SD=.35)으로 거의 동일하였으며, 평가 준거별 평균 차이 역시 .02~.08점에 그쳤다. 따라서 성별에 따른 분석에서는 측정모형에서 추정한 전반적인 엄격성과 실제 평정 척도에서 관찰된 총점 및 평가 준거별 점수가 일관되게 유의한 차이를 보이지 않았다.

### 2) 담당 학교급에 따른 평가자 효과

〈표 7〉은 평가자의 담당 학교급에 따른 엄격성과 원점수 채점 결과를 비교한 것이다. 먼저 다국면 라쉬 모형에서 추정한 엄격성의 평균은 중학교 평가자 -.164 logit(SD=.706), 고등학교 평가자 .005 logit(SD=.637), 대학교 평가자 .159 logit(SD=.682)으로 나타났다. 학교급이 높아질수록 평균 엄격성 추정치가 높아지는 방향이 나타났으나, 학교급에 따른 차이는 통계적으로 유의하지 않았다( $F(2, 57)=1.148, p=.325$ ). 따라서 담당 학교급에 따라 평가자의 전반적인 엄격성이 체계적으로 달라진다고 보기는 어려웠다.

〈표 7〉 평가자의 담당 학교급에 따른 엄격성 및 채점 결과 비교

| 구분          |    |        | 담당 학교급          |                |                | F<br>(p)        | $\eta^2$ | 사후<br>검정 |
|-------------|----|--------|-----------------|----------------|----------------|-----------------|----------|----------|
|             |    |        | 중학교<br>M(SD)    | 고등학교<br>M(SD)  | 대학교<br>M(SD)   |                 |          |          |
| 엄격성(logit)  |    |        | -.164<br>(.706) | .005<br>(.637) | .159<br>(.682) | 1.148<br>(.325) | .039     | -        |
| 원<br>점<br>수 | 내용 | 주장     | 3.12<br>(.46)   | 3.17<br>(.45)  | 3.12<br>(.46)  | .080<br>(.923)  | .003     | -        |
|             |    | 이유와 근거 | 3.21<br>(.50)   | 3.00<br>(.43)  | 2.75<br>(.44)  | 5.083<br>(.009) | .151     | 중>대      |
|             |    | 반론 고려  | 3.02<br>(.55)   | 2.83<br>(.34)  | 2.56<br>(.45)  | 5.332<br>(.008) | .158     | 중>대      |

|    |          |               |               |               |                 |      |       |
|----|----------|---------------|---------------|---------------|-----------------|------|-------|
| 조직 | 글 전체 조직  | 3.38<br>(.41) | 3.02<br>(.34) | 2.93<br>(.45) | 6.776<br>(.002) | .192 | 중)고·대 |
|    | 문단 내 조직  | 3.22<br>(.47) | 3.00<br>(.47) | 2.85<br>(.43) | 3.387<br>(.041) | .106 | 중)대   |
| 표현 | 문장과 어휘   | 2.75<br>(.44) | 2.96<br>(.48) | 3.22<br>(.52) | 4.701<br>(.013) | .142 | 대)중   |
|    | 어문 규범 준수 | 3.75<br>(.40) | 3.80<br>(.51) | 3.74<br>(.44) | .117<br>(.890)  | .004 | -     |
| 총점 |          | 3.21<br>(.40) | 3.11<br>(.36) | 3.02<br>(.39) | 1.148<br>(.325) | .039 | -     |

$\eta^2$  해석 기준(Cohen, 1988: 281, 285-287)

.01 미만(무시할 만한 효과), .01~.06(작은 효과), .06~.14(중간 효과), .14 이상(큰 효과)

원점수 총점에서도 학교급에 따른 차이는 통계적으로 유의하지 않았으며, 이는 다국면 라쉬 모형에서 추정된 평가자 엄격성의 분석 결과와 수렴하였다. 따라서 담당 학교급에 따른 차이가 평가자의 전반적인 채점 수준에서 나타난다고 보기는 어려웠다.

반면 평가 준거별 원점수에서는 학교급에 따른 차이가 일부 확인되었다. ‘이유와 근거’, ‘반론 고려’, ‘글 전체 조직’, ‘문단 내 조직’, ‘문장과 어휘’에서 학교급에 따른 평균 점수 차이가 통계적으로 유의하였으며, ‘주장’과 ‘어문 규범 준수’에서는 유의한 차이가 나타나지 않았다.<sup>6)</sup> Tukey 사후검정 결과, ‘이유와 근거’·‘반론 고려’·‘문단 내 조직’에서는 중학교 평가자가 대학교 평가자보다 높은 점수를 부여하였으며, ‘글 전체 조직’에서는 중학교 평가자가 고등학교·대학교 평가자보다 높은 점수를 부여하였다. 반면 ‘문장과 어휘’에서는 대학교 평가자가 중학교 평가자보다 높은 점수를 부여하였다. 효과크기는 ‘이유와 근거’( $\eta^2=.151$ ), ‘반론 고려’( $\eta^2=.158$ ), ‘글 전체 조

- 6) ‘어문 규범 준수’에서 학교급에 따른 유의한 차이가 나타나지 않은 것은 평가 대상이 대학 생의 글이었고, 글이 워드프로세서 환경에서 작성되었다는 점과 관련될 수 있다. 이러한 조건으로 어문 규범 오류가 전반적으로 적게 나타나 평가자 간 판단 차이가 제한되었을 가능성이 있다. 실제로 이 준거는 난이도가 ‘쉬움’ 수준이었으며, 변별도도 7개 평가 준거 가운데 가장 낮았다.

직'(η²=.192), '문장과 어휘'(η²=.142)에서 큰 수준으로 나타났으며, '문단 내 조직'(η²=.106)은 중간 수준의 효과를 보였다.

원점수에서 확인된 평가 준거별 차이가 평가자의 전반적인 엄격성과 평가 준거의 측정 특성을 고려한 이후에도 나타나는지를 확인하기 위하여 담당 학교급과 평가 준거 간 편향 상호작용을 분석하였다(〈표 8〉). 전체 편향 상호작용에 대한 Wald 검정 결과는 통계적으로 유의하였다( $\chi^2(12)=258.07$ ,  $p<.001$ ). 이는 담당 학교급에 따른 평가자 효과가 모든 평가 준거에서 동일한 방향의 엄격성 또는 관대성으로 나타나는 것이 아니라, 적용하는 평가 준거에 따라 그 방향과 정도가 달라짐을 의미한다.

〈표 8〉 담당 학교급과 평가 준거 간 편향 분석 결과

| 평가 준거 |          | 중학교<br>편향(SE) | 고등학교<br>편향(SE) | 대학교<br>편향(SE) | $\chi^2(2)$ | p     |
|-------|----------|---------------|----------------|---------------|-------------|-------|
| 내용    | 주장       | .197(.051)    | -.062(.051)    | -.134(.051)   | 15.728      | <.001 |
|       | 이유와 근거   | -.226(.051)   | -.032(.050)    | .258(.051)    | 30.755      | <.001 |
|       | 반론 고려    | -.223(.051)   | -.052(.050)    | .274(.051)    | 32.895      | <.001 |
| 조직    | 글 전체 조직  | -.314(.051)   | .160(.051)     | .154(.051)    | 37.555      | <.001 |
|       | 문단 내 조직  | -.187(.051)   | .039(.050)     | .148(.050)    | 15.104      | <.001 |
| 표현    | 문장과 어휘   | .562(.051)    | .028(.050)     | -.590(.051)   | 166.502     | <.001 |
|       | 어문 규범 준수 | .191(.054)    | -.080(.054)    | -.111(.054)   | 12.695      | .002  |

편향 분석에서 나타난 주요 양상은 원점수 분석과 대체로 수렴하였다. '이유와 근거'와 '반론 고려'에서는 중학교 평가자가 상대적으로 관대한 편향을, 대학교 평가자가 상대적으로 엄격한 편향을 보였다. '글 전체 조직'과 '문단 내 조직'에서도 중학교 평가자는 상대적으로 관대하게 평가한 반면, 고등학교와 대학교 평가자는 상대적으로 엄격한 방향을 보였다. 반대로 '문장과 어휘'에서는 중학교 평가자가 상대적으로 엄격한 편향을, 대학교 평가자가

상대적으로 관대한 편향을 나타냈다. 이는 원점수 분석에서 중학교 평가자가 ‘이유와 근거’, ‘반론 고려’, ‘글 전체 조직’, ‘문단 내 조직’에 상대적으로 높은 점수를 부여하고, 대학교 평가자가 ‘문장과 어휘’에 상대적으로 높은 점수를 부여한 결과와 동일한 방향이다.

한편 원점수 평균에서는 학교급에 따른 유의한 차이가 나타나지 않았던 ‘주장’과 ‘어문 규범 준수’에서도 편향 분석에서는 유의한 학교급×평가 준거 상호작용이 확인되었다. 이는 두 분석이 포착하는 평가자 효과의 측면이 다르기 때문으로 해석할 수 있다. 원점수 분석은 실제 평정 척도에서 각 집단이 부여한 절대적인 평균 점수의 차이를 보여 주는 반면, 편향 분석은 평가자의 전반적인 엄격성과 평가 준거의 측정 특성을 고려한 상태에서 특정 준거에 대해 상대적으로 엄격하거나 관대하게 반응한 정도를 보여 준다. 따라서 평균 점수 자체에는 유의한 집단 차이가 없더라도, 특정 학교급 평가자가 다른 준거에 비해 해당 준거를 상대적으로 엄격하거나 관대하게 적용하는 양상이 나타날 수 있다.

이상의 결과를 종합하면, 담당 학교급에 따른 평가자 효과는 전반적인 엄격성의 차이보다는 평가 준거에 따라 엄격성과 관대성의 방향이 달라지는 형태로 나타났다. 특히 원점수 분석에서도 학교급 간 차이가 나타난 ‘이유와 근거’, ‘반론 고려’, ‘글 전체 조직’, ‘문단 내 조직’, ‘문장과 어휘’에서는 편향 분석에서도 동일한 방향의 차이가 확인되었다. 이에 후속 면담에서는 이들 평가 준거를 중심으로 학교급에 따른 평가 준거의 해석과 적용 방식이 어떻게 달라지는지를 살펴보았다.

## 5. 후속 면담 결과

〈표 9〉는 학교급에 따른 채점 특성을 해석하기 위해 실시한 평가자 FGI의 분석 결과를 제시한 것이다. FGI 참여자들의 진술에서는 담당 학교급에 따라 평가 준거를 해석하고 적용하는 관점이 다르게 나타날 가능성을 확인할 수 있었다. 특히 학습자 수행에 대한 암묵적 기대, 글 전개에의 구성 원리에

대한 관점, 언어 사용에 대한 평가 관점에서 차이가 확인되었으며, 이는 양적 분석에서 확인된 학교급별 채점 특성을 해석하는 데 참고할 수 있는 질적 근거를 제공하였다.<sup>7)</sup>

〈표 9〉 FGI 결과

| 상위 주제             | 주제           | 코드                                                      | 학교급에 따른 채점 특성의 해석 가능성                                              |
|-------------------|--------------|---------------------------------------------------------|--------------------------------------------------------------------|
| 학습자 수행에 대한 암묵적 기대 | 논증 요소의 충족 여부 | (이유·근거 제시 여부) 주장에 대한 이유나 근거가 명시적으로 제시되었는지를 판단함.         | 학교급이 낮은 평가자는 이유·근거가 제시되었다는 사실 자체를 우선적으로 인정하는 경향                    |
|                   |              | (이유·근거의 충분성) 주장을 뒷받침하기 위해 이유와 근거가 충분히 제시되었는지를 판단함.      | 학교급이 낮은 평가자는 이유와 근거의 양적 충족 여부를 우선적으로 평가하는 경향                       |
|                   |              | (이유·근거의 구체성) 이유나 근거가 추상적인 진술에 머물지 않고 구체적으로 설명되었는지를 판단함. | 학교급이 높은 평가자는 이유·근거의 구체성을 함께 요구하는 경향                                |
|                   |              | (이유·근거의 관련성) 제시된 이유나 근거가 주장과 직접적으로 연결되었는지를 판단함.         | 학교급이 높은 평가자는 주장과 이유·근거의 논리적 관련성을 중시하는 경향                           |
|                   |              | (이유·근거의 질) 이유나 근거가 주장을 실질적으로 뒷받침하고 독자를 설득할 수 있는지를 판단함.  | 학교급이 높은 평가자는 이유·근거의 질을 중시하는 경향                                     |
|                   | 반론 고려        | (반론 제시 여부) 자신의 주장과 다른 견해가 글에 포함되었는지를 판단함.               | 학교급이 낮은 평가자는 반대 관점이 언급되었다는 사실을 긍정적으로 평가하는 경향                       |
|                   |              | (반론의 구체성) 예상 가능한 반대 입장이나 문제점이 구체적으로 제시되었는지를 판단함.        | 학교급이 높은 평가자는 반대 관점의 실질적 내용과 구체성을 요구하는 경향                           |
|                   |              | (반론 고려의 효과) 반대 관점의 검토가 자신의 주장을 정교화하거나 강화하는지를 판단함.       | 학교급이 높은 평가자는 반론이 전체 논증에 기여하는 정도를 중시하는 경향                           |
|                   | 수행 수준의 참조 틀  | (발달 단계에 대한 기대) 평가자가 평소 접하는 학습자의 수행 수준을 암묵적인 기준으로 활용함.   | 학교급이 낮은 평가자는 논증 요소의 제시 여부를 중심으로 판단하고, 학교급이 높은 평가자는 질적 완성도를 기대하는 경향 |

7) 〈표 9〉에 제시한 결과는 4명의 FGI 참여자에게서 확인된 평가 준거 해석 양상을 기술한 것으로, 각 학교급 평가자 전체의 특성을 일반화하기보다 양적 분석에서 확인된 학교급별 차이를 설명할 수 있는 하나의 기능성으로 해석하였다.

|                                       |                 |                                                                  |                                                                    |
|---------------------------------------|-----------------|------------------------------------------------------------------|--------------------------------------------------------------------|
| 글<br>전개의<br>구성<br>원리에<br>대한<br>관점     | 글 전체<br>조직      | (형식 문단과 내용 문단의 구분)<br>서론·본론·결론 등 글의 기본적인<br>형식이 구분되어 있는지를 판단함.   | 학교급이 낮은 평가자는 글의<br>기본적인 구성과 내용의 구분을<br>중시하는 경향                     |
|                                       |                 | (각 부분의 역할) 글의 각 부분이 문제<br>제기, 근거 제시, 정리 등 적절한<br>기능을 수행하는지를 판단함. | 학교급이 높은 평가자는 각 부분의<br>존재보다 논증 전개에서 수행하는<br>역할을 중시하는 경향             |
|                                       |                 | (글의 통일성) 글 전체가 하나의 중심<br>생각을 유지하는지를 판단함.                         | 학교급이 낮은 평가자는 중심<br>내용에서 벗어나지 않는 통일성을<br>중시하는 경향                    |
|                                       |                 | (논증의 단계적 심화) 논리가 단순히<br>나열되지 않고 일정한 방향으로<br>발전하는지를 판단함.          | 학교급이 높은 평가자는 글의 전개가<br>단계적으로 심화되는지를 함께<br>고려하는 경향                  |
|                                       | 문단 내<br>조직      | (문단의 통일성) 문단을 구성하는<br>문장들이 하나의 중심 생각에<br>집중하는지를 판단함.             | 학교급이 낮은 평가자는 문단에서<br>중심적으로 말하려는 내용이<br>확인되는지를 우선적으로 평가하는<br>경향     |
|                                       |                 | (문단의 완결성) 중심 생각이 뒷받침<br>문장을 통해 구체적이고 충분히<br>설명되는지를 판단함.          | 학교급이 높은 평가자는 중심 생각이<br>어느 정도 확장·심화되었는지를<br>중시하는 경향                 |
|                                       |                 | (문단의 일관성) 문단 내 문장들이<br>단절되지 않고 논리적·유기적으로<br>연결되는지를 판단함.          | 학교급이 높은 평가자는 문장<br>간 논리적 연결과 내용 전개의<br>자연스러움을 중시하는 경향              |
|                                       | 구성의<br>기여       | (논증에 대한 기여) 글 전체와 각<br>문단의 구성이 논증의 형성과 발전에<br>기여하는지를 판단함.        | 학교급이 높은 평가자는 조직을<br>형식적 배열이 아니라 논증 수행을<br>위한 구성 원리로 해석하는 경향        |
| 언어<br>사용의<br>정확성과<br>효과성에<br>대한<br>관점 | 규범적<br>정확성      | (문장의 정확성) 문장이 문법적으로<br>완결되고 자연스럽게 구성되었는지를<br>판단함.                | 학교급이 낮은 평가자는 비문,<br>호응 오류, 어색한 문장 구조 등에<br>민감하게 반응하는 경향            |
|                                       |                 | (어휘 사용의 적절성) 사용된 어휘가<br>문맥과 의미에 맞는지를 판단함.                        | 학교급이 낮은 평가자는<br>부정확하거나 부적절한 어휘 사용을<br>중요한 감점 요인으로 보는 경향            |
|                                       |                 | (표현의 규범성) 일상적·구어적<br>표현이나 부자연스러운 표현이 글의<br>격식에 적절한지를 판단함.        | 학교급이 낮은 평가자는 표준적이고<br>규범적인 표현 여부를 중시하는 경향                          |
|                                       | 의미<br>전달의<br>효과 | (의미의 명료성) 문장의 의미가<br>모호하지 않고 쉽게 이해되는지를<br>판단함.                   | 학교급이 높은 평가자는 표현의<br>화려함이나 깊이가 부족하더라도<br>의미가 명확하면 긍정적으로<br>평가하는 경향  |
|                                       |                 | (독자에 대한 전달 효과) 표현이 필자의<br>의도와 의미를 독자에게 정확하고<br>효과적으로 전달하는지를 판단함. | 학교급이 높은 평가자는 문장과<br>어휘의 기능을 독자와의 의사소통<br>측면에서 해석하는 경향              |
|                                       |                 | (맥락의 적절성) 문장과 어휘가 글의<br>목적, 내용 및 문맥에 적절하게<br>사용되었는지를 판단함.        | 학교급이 높은 평가자는 규범적<br>정확성뿐<br>아니라 글의 맥락에서 표현이<br>수행하는 역할을 함께 고려하는 경향 |

### 1) 학습자 수행에 대한 암묵적 기대

양적 분석에서는 ‘이유와 근거’와 ‘반론 고려’에서 학교급이 낮은 평가자의 평균 점수가 상대적으로 높게 나타났다. FGI에서는 평가자들이 모두 성인의 글이라는 사실을 전제로 평가하였음에도, 글이 어느 정도의 수행을 보여야 하는가를 판단하는 기준에는 자신이 평소 접하는 학습자의 수행 수준이 비명시적으로 반영되고 있음을 확인할 수 있었다.

FGI 참여자들의 진술에서는 상대적으로 낮은 학교급을 담당한 참여자가 이유·근거를 제시하거나 다른 관점을 고려하려는 시도 자체를 의미 있는 수행으로 해석하는 양상이 확인되었다.

- 평소 우리가 접하는 학생들의 경우에는 이유를 붙여 보려고 한 것만으로도 의미가 있습니다. 아직 근거가 충분하지 않더라도 방향은 잡고 있다고 봅니다. (Mf09)
- 학생들은 자기 생각을 쓰는 데 집중하다 보니 반대 입장을 굳이 생각하지 않는 경우가 많습니다. 이 글은 반론을 자세히 다루지는 않았지만, 다른 의견이 있을 수 있다는 것을 알고 글에 넣으려고 했어요. 중학교 단계에서는 그런 관점의 확장이 중요한 성취라고 생각했습니다. (Mm09)

반면 학교급이 높은 평가자는 이유·근거와 반론을 개별 요소의 제시 여부보다 논증 전체에서 수행하는 기능을 중심으로 해석하고 평가하는 양상이 나타났다.

- 글에서 이유가 제시되어 있다는 사실만으로 충분하지는 않습니다. 그 이유가 주장을 실제로 뒷받침하는지, 근거가 구체적이고 충분한지, 서로 비슷한 내용을 반복하고 있지는 않은지를 함께 보아야 합니다. (Hf10)
- 반대하는 사람이 있을 수 있다고 한 문장 쓰는 것과 반론을 실제로 검토하는 것은 다릅니다. 어떤 반론이 가능한지 구체적으로 제시하고, 그 반론의 타당

성을 검토한 뒤 자신의 주장을 다시 강화해야 합니다. (Um04)

이러한 결과는 동일한 글을 평가하더라도 학교급에 따른 평가자의 교육 경험이 평가 준거 해석과 관련되어 있을 가능성을 보여준다. FGI 참여자들의 진술에서는 학교급이 낮은 평가자는 논증을 형성하려는 시도를 중요한 수행으로 해석하는 반면, 학교급이 높은 평가자는 논증의 완결성과 각 요소가 논증에 기여하는 정도를 중심으로 수행 수준을 판단하는 양상이 나타났다.

## 2) 글 전개와 구성 원리에 대한 관점

양적 분석에서는 ‘글 전체 조직’과 ‘문단 내 조직’에서 학교급이 낮은 평가자의 평균 점수가 상대적으로 높게 나타났다. FGI에서는 학교급에 따라 글이 어떻게 구성되고 전개되어야 하는가를 바라보는 관점에 차이가 나타났다.

먼저 ‘글 전체 조직’에서 학교급이 낮은 평가자는 글의 기본적인 구조가 형성되어 있는지에 주목하는 양상이 나타났다. 특히 서론·본론·결론이 형식적으로 완전하지 않더라도 내용이 몇 개의 단위로 구분되어 전개되는 경우에는 글의 전체적인 틀이 형성된 것으로 판단하였다.

- 실제 학생들의 글을 보면 문단 구분 없이 계속 쓰거나, 앞에서 했던 이야기가 뒤에 다시 나오는 경우가 많습니다. 반면 이 글은 글의 단위마다 형식적으로 완벽하지는 않더라도, 서론·본론·결론의 큰 흐름이 나타납니다. 글의 기본적인 틀은 갖추었다고 보았습니다. (Mf09)

반면 학교급이 높은 평가자는 서론·본론·결론의 형식적 구분을 기본적으로 고려하면서도, 더 나아가 각 부분이 논증의 전개 과정에서 적절한 역할을 수행하고 논증이 단계적으로 발전하는지를 판단하는 양상이 나타났다.

- 서론, 본론, 결론처럼 보이는 부분이 있다고 해서 글 전체의 조직이 잘되었다

고 보기는 어렵습니다. 이 글은 큰 순서는 있었지만 논증이 단계적으로 발전한다고 보기는 어려웠습니다. (Hf10)

‘문단 내 조직’에서도 평가 초점의 차이가 나타났다. 학교급이 낮은 평가자는 하나의 문단 안에서 중심 내용이 유지되는 ‘통일성’을 기본적인 조직 능력으로 판단하였다.

- 한 문단 안에서 전혀 다른 이야기가 섞이지 않고 비슷한 내용이 이어지면, 기본적으로 어느 정도의 조직을 보이고 있다고 생각합니다. 문장 사이의 연결이 아주 치밀하지는 않아도 중심 내용이 크게 흔들리지 않는 점을 긍정적으로 평가했습니다. (Mm09)

반면 학교급이 높은 평가자는 문단의 통일성뿐 아니라 중심 내용을 충분히 전개하는 ‘완결성’, 문장 간 논리적 연결을 의미하는 ‘일관성’을 비중 있게 고려하는 양상이 나타났다.

- 문단마다 하나의 중심 생각이 드러나야 한다는 통일성 외에 완결성이나 일관성 같은 요인도 중요하다고 생각합니다. 그러니까, 중심 생각을 충분히 설명하는 뒷받침 문장들이 있어야 하고, 이러한 문장들 사이의 연결이 논리적이고 유기적이어야 합니다. (Um04)

이처럼 FGI 참여자들의 진술에서는 학교급이 낮은 평가자가 조직 준거를 기본적인 구조 형성과 통일성을 중심으로 해석한 반면, 학교급이 높은 평가자는 글과 문단의 구조뿐 아니라 조직의 완성도까지 포함하는 보다 복합적인 기준으로 해석하는 양상이 나타났다. 이는 양적 분석에서 확인된 학교급별 채점 특성을 설명하는 하나의 가능성을 시사한다.

### 3) 언어 사용의 정확성과 효과성에 대한 관점

양적 분석에서는 ‘문장과 어휘’에서 학교급이 높은 평가자의 평균 점수가 상대적으로 높게 나타났다. FGI에서는 학교급에 따라 문장과 어휘를 평가할 때 중요하게 고려하는 언어적 특성이 서로 다르게 나타났다.

학교급이 낮은 평가자는 문장의 정확성이나 어휘 사용의 적절성에 대한 민감도가 높게 나타났으며, 규범적 정확성을 중심으로 평가하고 있었다.

- 학생들의 정확한 문장 사용에 평소 많은 관심을 가지고 지도하기 때문에 이러한 부분이 더 많이 보였던 것 같습니다. (Mm09)

반면 학교급이 높은 평가자는 표현이 독자에게 전달하는 효과와 맥락적 적절성까지 함께 고려하는 양상이 나타났다. 글의 표현이 다소 미흡하더라도 문장이 의미를 명확하게 전달한다면 해당 준거에서는 긍정적으로 평가할 수 있다는 인식이 확인되었다. 또한 일부 평가자는 이 준거에서 특별히 눈에 띄는 결함이 없었다는 점을 근거로 비교적 높은 평가를 부여하기도 하였다.

- 문장이 반드시 어렵고 복잡해야 좋은 것은 아닙니다. 오히려 문장을 지나치게 길게 쓰거나 추상적인 용어를 많이 사용해서 의미가 불분명해지는 경우도 많습니다. 이 글은 표현의 세련됨은 다소 부족했지만, 문장 자체는 비교적 직접적이고 뜻도 어렵지 않게 전달되었습니다. (Hf10)
- 특별히 눈에 밝히는 흠결이 보이지 않았습시다. (Um04)

이처럼 FGI 참여자들의 진술에서는 학교급이 낮은 평가자가 문장과 어휘를 규범적 정확성을 중심으로 해석한 반면, 학교급이 높은 평가자는 표현이 독자에게 의미를 효과적으로 전달하는 기능까지 함께 고려하는 양상이 나타났다. 이는 양적 분석에서 확인된 학교급별 채점 특성을 설명하는 하나의 가능성을 시사한다.

## V. 논의 및 제언

### 1. 논의

이 연구는 쓰기 평가에서 나타나는 평가자 효과를 평가자의 엄격성과 평가 준거 해석이라는 두 차원에서 분석하였다. 이를 위해 평가 대상 글, 평가 준거 및 평가자의 측정 특성을 확인하고, 성별과 담당 학교급에 따른 채점 결과를 비교한 뒤, 학교급 간 차이가 나타난 평가 준거를 중심으로 후속 면담을 실시하였다. 연구 결과를 바탕으로 논의하면 다음과 같다.

첫째, 평가자 효과는 평가자의 전반적인 엄격성뿐 아니라 평가 준거를 해석하고 적용하는 과정에서도 나타났다. 다국면 라쉬 모형 분석 결과 평가자 간 엄격성에는 일정한 차이가 존재하였으나, 대부분의 평가자는 수용 가능한 수준의 채점 일관성을 유지하였다. 반면 총점에서는 담당 학교급에 따른 차이가 나타나지 않았음에도 일부 평가 준거에서는 유의한 차이가 확인되었다. 이는 평가자 효과를 단일한 채점 성향으로 이해하기보다, 평가자가 각 평가 준거의 의미를 어떻게 해석하고 적용하는지를 함께 고려할 필요가 있음을 시사한다. 특히 총점 수준에서는 상쇄되는 차이가 평가 준거 수준에서는 드러날 수 있다는 점은 평가자 효과를 분석할 때 총점과 준거별 점수를 함께 검토할 필요가 있음을 보여준다. 이러한 결과는 평가자가 동일한 수행에 대해서도 평가 준거에 따라 서로 다른 판단을 내릴 수 있음을 보고한 Eckes(2005)의 논의와도 맥락을 같이한다.

둘째, 후속 면담 참여자들의 진술은 평가 준거별 점수 차이가 준거의 의미와 적용 범위를 해석하는 방식과 관련되어 있을 가능성을 보여주었다. 상대적으로 낮은 학교급을 담당한 참여자들은 논증의 기본 구조를 학습자의 수행 수준과 관련지어 긍정적으로 평가하는 한편 언어 사용에서는 규범적 정확성에 주목하였고, 상대적으로 높은 학교급을 담당한 참여자들은 논증의

완결성과 표현의 의사소통적 효과를 함께 고려하였다. 이러한 면담 결과는 양적 분석에서 확인된 학교급별 차이를 설명할 수 있는 하나의 가능성으로서, 동일한 루브릭을 사용하더라도 평가자가 어떤 텍스트 특성을 특정 평가 준거의 수행으로 인정하고 어느 수준까지 수행을 요구하는지에 따라 평가 판단이 달라질 수 있음을 시사한다. 이는 평가자의 채점을 단순한 점수 부여가 아니라 평가 준거와 텍스트의 특성을 연결하여 의미를 구성하는 인지적 판단 과정으로 이해한 Heidari et al. (2022)의 논의와도 맥락을 같이한다.

셋째, 담당 학교급은 평가자가 평가 준거를 해석하는 과정과 관련된 교육적 맥락으로 이해할 필요가 있다. 이 연구에서는 성별에 따른 차이는 확인되지 않았으나, 담당 학교급에 따라 일부 평가 준거에서 통계적으로 유의한 차이가 나타났다. 이러한 결과는 평가자가 오랫동안 경험해 온 교수·학습 환경과 학습자 수행에 대한 기대가 평가 준거의 적용 방식과 관련될 가능성을 보여준다.<sup>8)</sup> 특히 학교급이 다른 평가자들은 학생의 발달 수준과 교육 목표에 대한 서로 다른 기대를 형성해 왔을 수 있으며, 이러한 차이가 논증의 완성도나 언어 사용처럼 질적 판단이 요구되는 평가 준거에서 서로 다른 참조 기준으로 나타났을 가능성이 있다. 따라서 평가자 효과를 이해하기 위해서는 평가자의 엄격성뿐 아니라 평가자가 형성해 온 교육 경험과 그에 따른 준거 해석 과정을 함께 고려할 필요가 있다.

## 2. 제언

이 연구의 결과를 바탕으로 쓰기 평가의 신뢰성과 타당성을 높이기 위

---

8) 이 연구에서 담당 학교급은 평가자의 교육 경험과 학습자 수행에 대한 기대가 형성되는 맥락을 나타내는 변인으로 보았다. 다만 채점 판단에는 평가 경험, 쓰기 지도 경험, 전문적 배경 등 다른 요인도 함께 작용할 수 있으므로, 확인된 차이를 담당 학교급 자체의 독립적인 효과로 단정하기보다 학교급과 연계된 교육적·전문적 경험과 관련된 결과로 해석할 필요가 있다.

한 실천적 방향을 루브릭 설계, 채점 운영, 평가자 훈련의 측면에서 제안하면 다음과 같다.

첫째, 루브릭 설계에서는 평가 준거의 의미와 적용 범위를 구체화할 필요가 있다. 이 연구에서는 동일한 루브릭을 사용하더라도 평가자가 어떤 텍스트 특성을 준거의 증거로 인정하고 어느 수준의 수행을 요구하는지에 따라 점수 차이가 나타날 가능성이 확인되었다. 특히 ‘이유와 근거’, ‘반론 고려’, ‘글 전체 조직’, ‘문단 내 조직’과 같은 준거는 수행 요소의 존재 여부뿐 아니라 내용의 구체성, 논증에 대한 기여, 조직의 완성도 등을 함께 판단해야 하므로 평가자의 해석이 개입될 여지가 크다. 따라서 루브릭에는 준거의 명칭과 일반적인 수행 기술문뿐 아니라, 평가자가 관찰해야 할 텍스트 특성과 점수 수준을 구분하는 판단 기준을 구체적으로 제시할 필요가 있다. 아울러 수준별 예시 글과 기준 점수, 점수 판단의 근거를 함께 제공하여 평가자가 준거의 적용 범위와 점수 수준의 경계를 확인할 수 있도록 해야 한다.

둘째, 채점 운영에서는 총점뿐 아니라 평가 준거별 점수와 판단의 일관성을 함께 점검할 필요가 있다. 이 연구에서는 총점에서 확인되지 않은 평가자 집단 간 차이가 일부 평가 준거에서 나타났다. 이는 총점만을 점검할 경우 준거별 점수 차이가 서로 상쇄되어 평가자 효과가 드러나지 않을 수 있음을 보여준다. 따라서 채점 과정에서는 준거별 점수 분포와 평가자별 엄격성·관대성의 양상을 함께 확인하고, 특정 평가자가 일부 준거에서 지속적으로 다른 점수 경향을 보이는지 점검할 필요가 있다. 또한 이러한 점검 결과는 평가자 간 차이가 두드러지는 평가 준거를 파악하고, 이후 평가자 훈련의 내용을 구성하는 기초 자료로 활용될 수 있다.

셋째, 평가자 훈련은 기준 점수와의 일치를 높이는 절차를 넘어 평가 준거에 대한 공통의 해석 범위를 형성하는 과정으로 운영할 필요가 있다. 이를 위해 평가자가 어떤 텍스트 특성을 준거의 증거로 판단하였는지, 수행 요소의 존재와 질을 어떻게 구분하였는지, 점수 수준의 경계를 어디에 설정하였는지를 설명하고 비교하도록 해야 한다. 특히 학교급이나 교육 경험이 다른

평가자들이 동일한 글을 채점한 뒤 판단 근거를 공유하면, 각자가 암묵적으로 전제한 학습자 수준과 수행 기대를 확인할 수 있다. 채점 전에는 평가 목적과 대상, 평가 척도의 구성 및 의미를 충분히 검토하고, 채점 중에는 실제 텍스트와 수행 기술문을 반복적으로 대조하면서 판단 근거를 점검하도록 지원할 필요가 있다.

다만 평가자 훈련의 목적을 모든 평가자의 판단을 획일적으로 일치시키는 데 두는 것은 적절하지 않을 수 있다. 평가자가 축적한 교육 경험과 전문적 배경은 서로 다른 참조 기준을 형성하는 한편, 학습자의 수행을 다각도로 해석하는 데 필요한 전문성을 제공하기 때문이다. 따라서 평가자 훈련에서는 이러한 경험을 배제하기보다, 경험에서 비롯된 암묵적인 기대와 판단 기준을 드러내고 평가 목적과 루브릭이 허용하는 범위 안에서 조정하는 데 초점을 둘 필요가 있다. 이를 통해 평가자의 전문적 판단을 존중하면서도 평가 결과의 일관성과 설명 가능성을 높일 수 있을 것이다.

## VI. 연구의 기여와 향후 과제

이 연구는 쓰기 평가에서 나타나는 평가자 효과를 평가자의 엄격성과 평가 준거 해석이라는 두 차원에서 통합적으로 분석하였다는 점에서 의의가 있다. 특히 총점 수준에서는 드러나지 않았던 평가자 간 차이를 평가 준거별 점수에서 확인하고, 이러한 차이가 평가자의 교육 경험과 평가 준거 해석 방식에 관련되어 있을 가능성을 양적·질적 자료를 통합하여 설명하였다. 이를 통해 평가자 효과를 단순한 채점 성향의 차이가 아니라 평가 준거를 이해하고 적용하는 판단 과정의 차원까지 포함하는 현상으로 이해할 필요가 있음을 제시하였다.

또한 이 연구는 평가자 효과를 설명하는 배경 요인의 범위를 확장하였

다는 점에서도 의미가 있다. 기존 연구에서는 평가자의 전문 분야, 숙련도 및 평가 경험과 같은 전문적 배경을 중심으로 평가자 효과를 설명해 왔다 (Leckie & Baird, 2011; Li et al., 2025; Royal-Dawson & Baird, 2009; Shohamy, Gordon, & Kraemer, 1992; Soler, 2000). 이에 비해 이 연구는 담당 학교급이라는 교육적 맥락을 평가자의 전문적 배경 요인으로 제시하고, 학교급에 따라 형성된 교수·학습 경험과 학습자 수행에 대한 기대가 일부 평가 준거의 해석과 적용에 관련되어 있을 가능성을 제시하였다. 이는 평가자 효과 연구를 평가 결과의 측정과 보정 중심에서 평가자의 판단 과정과 교육적 맥락을 함께 설명하는 방향으로 확장하였다는 점에서 의의가 있다.

다만 이 연구는 대학생의 논증적 글쓰기와 중학교·고등학교·대학교 평가자를 대상으로 하였으므로, 연구 결과를 다른 쓰기 장르나 평가 상황으로 일반화하는 데에는 한계가 있다. 또한 후속 면담은 학교급 간 차이가 나타난 일부 평가 준거를 중심으로 양적 분석 결과를 해석하기 위한 탐색적 절차로 실시되었기 때문에 모든 평가 준거의 해석 과정을 포괄하지 못하였다. 특히 후속 면담 참여자가 4명으로 제한되어 있었으므로, 확인된 평가 준거 해석 양상을 각 학교급 평가자의 일반적인 특성으로 일반화하는 데에는 신중할 필요가 있다.

향후 연구에서는 다양한 학교급과 쓰기 장르를 대상으로 평가 경험, 쓰기 신념 및 평가 전문성 등 다양한 변인이 평가 준거의 해석과 적용에 어떠한 관련성을 가지는지를 체계적으로 검토할 필요가 있다. 또한 사고구술, 자극 회상 및 안구추적 등의 방법을 활용하여 평가자가 텍스트의 어떤 특성에 주목하고 이를 평가 준거와 점수 판단으로 연결하는 과정을 보다 직접적으로 분석함으로써, 평가자 효과의 인지적 메커니즘을 보다 심층적으로 규명할 필요가 있다.

\* 본 논문은 2026.07.31. 투고되었으며, 2026.08.14. 심사가 시작되어 2026.09.09. 심사가 종료되었음.

## 참고문헌

- 민병곤·남가영·김선희·장성민·이성준·권은선(2023), 『2023년 국민의 글쓰기 능력 진단 체계 개발』, 서울: 국립국어원.
- 박영민·최숙기(2009), 「현직 국어교사와 예비 국어교사의 쓰기 평가 비교 연구」, 『교육과정평가연구』 12(1), 123-143.
- 박영민·최숙기(2010ㄱ), 「예비 국어교사의 성별 및 학생 성별 인식에 따른 평가 차이 분석」, 『교육과정평가연구』 13(2), 239-258.
- 박영민·최숙기(2010ㄴ), 「Rasch 모형을 활용한 국어교사의 쓰기 평가 특성 분석: 중학생 설명문 쓰기 평가를 중심으로」, 『국어교육학연구』 37, 367-391.
- 박혜영·이상하·이명진·이미영·주현우·백은진(2021), 『평가를 활용한 초·중등학생 글쓰기 능력 신장 방안 연구(Ⅱ)』, 진천: 한국교육과정평가원.
- 백은진·가은아(2021), 「텍스트 분량 요인에 대한 현직 국어교사와 예비 국어교사의 쓰기 평가 특성 차이 분석」, 『청람어문교육』 82, 69-96.
- 성태제(2019), 『현대교육평가(5판)』, 서울: 학지사.
- 송지언(2016), 「평가자에 따른 작문 평가 결과의 비교 연구: 채점(scoring) 및 총평(description)의 양상을 중심으로」, 『작문연구』 29, 29-63.
- 장성민(2026), 「인공지능 작성 여부에 대한 출처 정보가 쓰기 평가에 미치는 영향: 글 수준 및 분석적 채점 준거별 변화 분석」, 『국어교육학연구』 61(2), 225-256.
- 장성민·김동섭·남가영·권은선·이성준·김선희·오예림·이수진·정은선·송규민·민소연·홍태경·민병곤(2024), 「2023 국민의 글쓰기 능력 진단 체계 개발 및 시행 결과 분석: 성인 대상의 논증적 글쓰기 능력 진단을 중심으로」, 『국어교육』 185, 1-32.
- 정미경(2011), 「국어교사의 성별에 따른 쓰기 평가 특성 분석」, 『교원교육』 27(4), 73-93.
- Cohen, J. (1988), *Statistical power analysis for the behavioral sciences* (2nd ed.), Hillsdale, NJ: Erlbaum.
- Eckes, T. (2005), "Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis", *Language Assessment Quarterly* 2(3), 197-221.
- Eckes, T. & Jin, K. Y. (2021), "Measuring rater centrality effects in writing assessment: A Bayesian facets modeling approach", *Psychological Test and Assessment Modeling* 63(1), 65-94.
- Heidari, N., Ghanbari, N., & Abbasi, A. (2022), "Raters' perceptions of rating scales criteria and its effect on the process and outcome of their rating", *Language Testing in Asia* 12(1), 1-19.
- Johnson, J. S. & Lim, G. S. (2009), "The influence of rater language background on writing performance assessment", *Language Testing* 26(4), 485-505.

- Leckie, G. & Baird, J. A. (2011), "Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience", *Journal of Educational Measurement* 48(4), 399-418.
- Li, Y., Liao, X., & Jia, J. (2025), "Effects of raters' professional backgrounds on assessing interpreting quality: An exploratory mixed-methods investigation into rater behavior", *System* 133, 103772.
- Linacre, J. M. (2002), "What do infit and outfit, mean-square and standardized mean", *Rasch Measurement Transactions* 16(2), 878.
- Palermo, C. (2022), "Rater characteristics, response content, and scoring contexts: Decomposing the determinates of scoring accuracy", *Frontiers in Psychology* 13, 937097.
- Royal-Dawson, L. & Baird, J. A. (2009), "Is teaching experience necessary for reliable scoring of extended English questions?", *Educational Measurement: Issues and practice* 28(2), 2-8.
- Shohamy, E., Gordon, C. M., & Kraemer, R. (1992), "The effect of raters' background and training on the reliability of direct writing tests", *The Modern Language Journal* 76(1), 27-33.
- Soler, H. H. (2000), "The effect of gender and working place of raters on university entrance examination scores", *Revista Española de Lingüística Aplica* 14, 161-180.
- Yan, X. (2014), "An examination of rater performance on a local oral English proficiency test: A mixed-methods approach", *Language Testing* 31(4), 501-527.

## 쓰기 평가에서 평가자 효과는 어떻게 나타나는가?

— 평가자의 엄격성과 평가 준거 해석을 중심으로

장성민

이 연구는 쓰기 평가에서 나타나는 평가자 효과를 평가자의 엄격성과 평가 준거 해석의 측면에서 분석하고, 성별과 담당 학교급에 따른 차이를 탐색하는 데 목적이 있다. 이를 위해 평가 대상 글, 평가 준거 및 평가자의 측정 특성을 확인하고, 성별과 담당 학교급에 따른 채점 결과를 비교한 후 그 차이에 대한 후속 면담을 실시하였다. 연구 결과, 성별에 따른 평가자 효과는 나타나지 않았으나 담당 학교급에 따라서는 일부 평가 준거에서 유의한 차이가 나타났다. 후속 면담 참여자들의 진술에서는 양적 분석에서 확인된 학교급별 차이를 설명할 수 있는 평가 준거 해석과 적용 방식의 차이가 나타났다. 이러한 결과는 쓰기 평가의 평가자 효과가 평가자의 엄격성뿐 아니라 평가 준거 해석에서도 나타날 수 있음을 보여주며, 평가자 훈련에서는 평가 준거의 해석과 적용에 대한 일관성을 함께 점검할 필요가 있음을 시사한다.

**핵심어** 쓰기 평가, 평가자 효과, 평가자 엄격성, 평가 준거 해석, 평가자 배경 요인

## How Do Rater Effects Manifest in Writing Assessment?

— Focusing on Rater Severity and Criterion Interpretation

Chang Sungmin

In this study, we investigated how rater effects manifest in writing assessment by examining rater severity and criterion interpretation, with particular attention paid to differences associated with the gender of the raters and the school level they teach. To this end, the measurement characteristics of writing samples, scoring criteria, and raters were examined, and the rating results were compared across the raters' genders and school levels. Follow-up interviews were conducted to determine the factors underlying the observed differences. The results revealed no significant rater effects associated with gender, whereas significant differences across school levels were observed for several scoring criteria. Follow-up interviews further revealed that raters at different school levels differed in how they interpreted and applied the scoring criteria. These findings suggest that rater effects in writing assessment may arise from differences in not only the severity of the raters but also their interpretation of scoring criteria. This study highlights the need for rater training to promote the consistent interpretation and application of scoring criteria.

**KEYWORDS** Writing assessment, rater effects, rater severity, criterion interpretation, rater background characteristics