http://dx.doi.org/10.20880/kler.2021.56.151.

국어교사의 쓰기평가 일관성 유형에 따른 채점 - 재채점 신뢰도 분석

윤금준 청명고등학교 교사(제1저자)

박영민 한국교원대학교 국어교육학과 교수(교신저자)

I. 서론

쓰기 능력은 개인의 사유와 경험을 표현하고 이를 공동체와 공유하게 함으로써, 인류의 지적 성과가 면면히 이어지도록 이끈 원동력이다. 이런 이유에서 세계 여러 나라에서 쓰기 능력의 중요성을 인식하고 쓰기 교육을 장려하고 있으며, 학생들의 쓰기 능력을 정확하게 평가하기 위한 다양한 방법을 모색하고 있다. 그러나 현재 우리나라의 경우, 고부담 검사에서는 학생들의 쓰기 수행에 대한 직접 평가를 시행하지 않고 있으며, 다른 대단위 평가에서도 간접 평가가 주로 쓰이고 있는 실정이다.

고부담 검사나 대단위 평가에 직접 평가를 적용하지 못하는 이유로 비용이 많이 든다는 점을 흔히 꼽지만, 평가의 신뢰도 확보의 문제도 그 이유 중 하나라 할 수 있다. 이러한 평가 장면에 직접 평가를 적용하려면 평가자 신뢰도가 전제되어야 하는데, 쓰기평가 중에 개입하는 평가자의 주관으로 인해 평가자 신뢰도 확보가 쉽지 않다. 평가자 신뢰도는 직접 평가의 전제에 해당하지만 이를 확보하는 일이 쉽지 않다 보니 쓰기평가 전문성에 대한 논의는 평가자 신뢰도 확보 방안에 집중되는 경향이 있다.

쓰기평가의 검사-재검사 연구에 해당하는 채점-재채점 신뢰도 분석은 같은 학생 글을 반복적으로 채점했을 때 그 결과가 얼마나 일관성이 있는지를 판단하는 것으로, 평가의 신뢰도를 정확하게 가늠할 수 있는 중요한 연구 방법이다. 대단위로 이루어지는 고부담 검사에서 평가자 신뢰도 검증을 위해 채점-재채점을 일부 시행하기도 하지만, 쓰기평가에서 채점-재채점 기반의 신뢰도 연구는 거의 찾아보기 어렵다. 쓰기평가 연구라 하더라도 평가자에게 동일한 학생 글을 반복적으로 채점하도록 요구하는 것은 현실적으로 어려운 일이기 때문이다.

이에 비해 의학 분야에서는 검사-재검사 기반의 신뢰도 분석 연구를 흔히 발견할 수 있다. 이재창·김지은·김민영·양지선·한명훈·권혁찬 외(2017)의 주의력 네트워크 검사에 대한 검사-재검사 신뢰도 분석, 이혜원·이경원(2014)의 어음 청각 검사에 대한 검사-재검사 신뢰도 분석 연구를 예로 꼽을 수 있다. 이러한 선행 연구에서는 특정 검사를 반복적으로 시행하여 검사 도구의 신뢰도를 평가하거나, 이를 바탕으로 한 피험자 진단의 적절성을 주로 다룬다.

평가자 신뢰도의 향상 방안을 모색하려면 먼저 평가자 신뢰도에 대한 분석이 필요하다. 이 연구의 목적은 쓰기평가의 일관성 유형에 따른 국어교사의 채점[1]-재채점 신뢰도를 분석하는 데 있다. 이에, 이 연구에서는 국어교사 36명을 평가자로 선정하여 학생 글 20편을 반복적으로 채점하게 하고 이를 바탕으로 채점-재채점 신뢰도를 분석하였다. 먼저 다국면 Rasch 분석으로 평가자의 일관성 유형을 살펴보고, 이를 바탕으로 채점-재채점의 기술통계, 급내 상관계수 및 Pearson 상관계수를 계산하여 일관성의 유형별 채점-재채점의 신뢰도를 분석하였다.

---

1) 이 연구에서 '채점'은 학생 글에 대해 점수를 부여하는 것을 말하며, '재채점'은 이전에 채점했던 동일한 학생 글에 대해 다시 점수를 부여하는 것을 의미한다. 이와 용어가 혼동되지 않도록 하기 위해 쓰기 능력을 측정하고 해석하는 일련의 과정은 '쓰기평가'로 표기하였다.

이 연구는 일부의 국어교사를 대상으로 연구 결과를 도출하였다는 점에서 그 결과를 일반화하는 데 한계가 있다. 하지만 그동안 국어교사를 대상으로 한 채점-재채점 신뢰도 분석 연구가 없었다는 점을 고려할 때, 이 연구의 결과는 평가자 신뢰도에 대한 면밀한 탐색을 가능하게 한다는 점에서 의의가 있다. 더 나아가 채점-재채점에 대한 이 연구의 결과는 앞으로 평가자 신뢰도 확보 방안을 위한 후속 연구에도 기여할 수 있을 것이다.

II. 이론적 배경

1. 쓰기평가 신뢰도

쓰기평가는 학생의 쓰기 능력에 관한 정보를 수집하고 교육적인 진단과 피드백을 제공하는 일련의 과정이라고 할 수 있다. 이러한 쓰기평가의 궁극적인 목적은 학생 개개인의 쓰기 능력에 대한 정보를 제공함으로써 학생의 쓰기 능력을 신장하고 의사소통 능력을 향상하는 데 있다고 할 수 있다.

쓰기평가에서 타당도는 검사 도구가 측정하고자 하는 바를 얼마나 충실하게 측정하는지를 의미하는 것으로, 신뢰도와 함께 쓰기평가에서 중요한 요소라 할 수 있다. 신뢰도는 측정하려는 바를 얼마나 일관성 있게 측정하는지를 의미하며, 동일한 검사를 반복하여 시행할 경우 그 결과의 유사성이 검사의 신뢰도를 보여 준다고 할 수 있다. 즉 신뢰도는 어떤 검사의 결과 값이 얼마나 일관성이 있는지를 통계적으로 나타내는 수치라 할 수 있다(강승호·김양분, 2004; 박도순·권순달·김명화·김영애·김정민, 2012; Gwet, 2012).

그런데 쓰기평가에서 직접 평가를 실시할 경우 학생의 쓰기 수행 과정이나 결과를 직접 관찰하고 평가하는 과정에서 평가자의 주관적 판단이 개입된다. 이러한 평가자의 주관성으로 인해 쓰기평가에서 평가의 신뢰도에

문제가 발생하게 되는 것이다. 이와 관련하여 평가자 판단과 해석에는 여러 요인들이 영향을 미친다는 연구 결과가 있으며(Ruth & Murphy, 1988), 평가자의 역할을 강조하기 위해 평가 결과를 산출해 내는 '평가도구의 역할'로 평가자의 역할을 설명하는 연구도 있었다(권대훈, 2008).

또한 평가자의 주관성을 해결하기 위한 방법으로 대단위 평가에서 신뢰도를 높이기 위해 여섯 단계의 실행 절차를 제안하였으며(White, 1984), 쓰기평가 협의 과정에서의 평가자 인식에 주목하여 평가자의 주관성을 해결하기 위한 대안을 제시하기도 하였다(서수현, 2012).

평가의 신뢰도는 평가자 간 신뢰도와 평가자 내 신뢰도로 구분할 수 있다(성태제, 2002). 평가자 간 신뢰도는 여러 평가자가 산출한 평가 결과 간의 일치도를 의미하며, 평가자 내 신뢰도는 한 평가자가 동일한 글을 반복적으로 평가했을 때 평가 결과가 얼마나 일치하는지를 말한다. 평가자 내 신뢰도가 확보되지 못하는 경우 평가에 대한 일관성이 확보되기 어려우므로 평가자 간 신뢰도 역시 유지되기 어렵다.

평가 또는 검사 도구의 신뢰도를 통계적 검증 방법으로는 내적 일관성을 분석하는 방법과 검사-재검사 신뢰도를 분석하는 방법 등이 있다(허소희, 2019). 내적 일관성은 일반적으로 Cronbach α를 산출하여 검증할 수 있으며, 검사-재검사 신뢰도는 급내 상관계수 또는 Pearson 상관계수 등을 산출하여 검증할 수 있다(이재창 외, 2017). 평가 결과나 검사 도구에 대한 신뢰도를 가장 정확하게 검증할 수 있는 방법은 검사-재검사 신뢰도를 분석하는 것이라고 할 수 있다.

그러나 그동안 쓰기평가 연구에서의 신뢰도 추정은 평가의 검사-재검사에 해당하는 채점-재채점을 실시하기 어려운 현실적 한계가 있다. 이로 인해 평가자 내 신뢰도는 다국면 Rasch 모형의 내적합 지수를 통해 추정하거나, 평가자 간 신뢰도는 일반적으로 내적 합치도 계수인 Cronbach α로 추정하는 경우가 다수를 이루고 있다(성낙수·김슬옹·김홍범·안주호·양정석·이정택 외, 2015: 252-255)

2. 평가자의 평가 특성

쓰기평가 과정에서 평가자의 특성은 다양하게 나타나지만, 평가의 신뢰도와 관련하여 중요하게 고려해야 하는 평가자 특성으로는 평가자의 엄격성과 일관성 등이 있다. Eckes(2008)는 평가자 간에는 엄격성과 일관성, 평가 기준의 준수와 해석 등에 차이가 있음을 밝히기도 하였다.

평가자의 엄격성은 평가자의 고유한 특성으로, 관대한 경향이나 엄격한 경향으로 나타난다. 예를 들어 평가자에 따라 어떤 평가자는 관대한 경향을 보이면서 높은 점수를 부여하지만 어떤 평가자는 엄격한 경향을 보이면서 낮은 점수를 부여할 수도 있다. 이러한 관대함과 엄격성은 서로 대립되는 개념을 나타내는 것으로, 일반적으로는 엄격성이라는 용어를 사용하여 엄격성의 정도를 표현한다. Barrett(2001)에서는 평가자의 엄격성이 평가 수행 중에 지속적으로 나타나는 경향성임을 보고하기도 하였다. 다음으로 평가자의 일관성은 평가자가 다수의 학생 글에 대해 동일한 잣대를 일관되게 적용하여 평가하는 것을 말하며, 평가자 내 신뢰도라는 명칭으로 알려져 있다(박영민, 2016: 422). 이러한 일관성이 유지되지 못 한 경우 평가의 결과를 신뢰하기 어렵다.

일관성을 기준으로 볼 때, 평가자의 유형은 적합, 과적합, 부적합으로 범주화할 수 있다. 적합은 평가자가 동일한 잣대를 일관되게 적용하여 평가하는 것을 의미한다. 과적합은 평가자가 특정한 점수만을 주로 부여하여 일관성이 과도한 것을 의미한다. 과적합이 지나칠 경우 학생 글의 수준을 특정한 점수 위주로 평가하게 되어 수준의 차이를 충분히 보여 주지 못할 수 있다. 부적합은 평가자의 일관성이 유지되지 못하거나 평가자가 임의로 점수를 부여하는 것을 말한다. 이러한 경우 실제 학생 글의 수준과 다른 점수를 부여할 수 있어 학생 글에 대한 변별이 어려울 뿐만 아니라 학생 글에 대한 정확한 평가가 어렵다. Myford & Wolfe(2003)의 연구와 같이, 일반적으로는 과적합인 경우보다는 부적합인 경우가 평가의 신뢰도에 더 문제가 있는 것으로 보고되고 있다.

III. 연구 방법

1. 연구 대상

국어교사의 쓰기평가 일관성 유형에 따른 채점-재채점의 신뢰도를 분석하기 위하여 고등학교에 재직하고 있는 국어교사 36명을 대상으로 20편의 학생 글에 대한 쓰기평가 채점-재채점을 실시하였다. 연구 대상은 성별, 교육 경력, 학교 소재지 등을 고려하여 전국적인 수준에서 표집하였으며, 성별 및 경력별로 정리한 연구 대상의 분포는 〈표 1〉과 같다.

〈표 1〉 연구 대상의 성별 및 경력별 분포

경력(년) / 성별	5년 미만	5~10년	10~15년	15~20년	20년 이상	계
남	1	1	6	3	3	14
여	1	5	9	4	3	22
계	2	6	15	7	6	36

〈표 1〉에 따르면, 교육 경력이 10이상 15년 미만인 경우가 15명으로 41.67%로 나타나 가장 높은 비율을 나타냈다. 또한 남자 교사는 14명으로 38.89%, 여자 교사는 22명으로 61.11%이었다. 남자 교사에 비해 여자 교사의 비율이 높지만, 현재 교육 현장에서의 성별 분포와 크게 다르지 않은 것으로 보인다. 학교 소재지에 따른 지역별 비율은, 서울·경기 지역 28명(77.78%), 충청 지역 2명(5.55%), 전라 지역 3명(8.33%), 경상 지역 1명(2.78%), 강원 지역 1명(2.78%), 제주 지역 1명(2.78%)이었다. 연구 대상이 서울·경기 지역에 다소 편중되어 있기는 하나 전국 단위 표집으로 구성되었음을 확인할 수 있다.

2. 검사 도구

이 연구에서는 국어교사를 대상으로 쓰기평가의 채점-재채점을 실시하기 위해 고등학교 2학년 학생이 쓴 논설문으로 쓰기평가 검사지를 구성하였다. 쓰기평가 검사지는 경기도 소재 A 고등학교의 2학년 3개 학급을 대상으로 글을 수집하였으며, 글이 지나치게 짧거나 글의 내용을 판별하기 어려운 경우 등을 제외하여 최종적으로 20편을 선정하였다. 쓰기 과제는 제시한 자료를 읽고 '동물원 폐지'에 대한 자신의 의견을 기술하도록 구성하였다. 쓰기평가 기준의 경우, 가은아(2011), 권태현(2014), 장은주(2015)의 쓰기평가 기준을 참고하여 일부 내용을 수정하여 사용하였다. 평가 척도는 6점 척도로, 내용·조직·표현으로 범주화하였으며 쓰기평가 기준은 〈표 2〉와 같다.

〈표 2〉 쓰기평가 기준표

평가 요소	평가 기준	척도
내용	1. 주장이 명료하고 타당한가?	1 ~ 6
	2. 주장을 뒷받침하는 근거가 타당하고 풍부한가?	1 ~ 6
	3. 내용이 통일성을 갖추고 있는가?	1 ~ 6
조직	1. 글의 전체 구조가 완결성을 갖추고 있는가?	1 ~ 6
	2. 문단을 적절히 나누었으며 문단 연결이 자연스럽고 긴밀한가?	1 ~ 6
표현	1. 독창적이며 설득력 있게 표현되었는가?	1 ~ 6
	2. 어문규범(단어 표기, 맞춤법, 문장 주술 호응 등)에 맞게 표현되었는가?	1 ~ 6

3. 연구 절차

이 연구는 국어교사의 쓰기평가 일관성 유형에 따른 채점-재채점의 신뢰도를 분석하기 위해 국어교사 36명을 대상으로 학생 글 20편을 채점-

재채점하게 하였다. 먼저 2019년 11월에 채점을 실시하였으며 3개월 후인 2020년 2월에 재채점을 실시하였다.

일반적으로 설문지 형태로 진행되는 검사-재검사에서는 검사와 재검사 사이의 기간이 너무 길면 정확한 검사-재검사 신뢰도 측정이 어려우며, 반대로 이 기간이 너무 짧으면 학습 효과의 영향을 받을 수 있다. 이 연구에서는 이를 고려하여 3개월의 시차를 두고 채점-재채점을 실시하였다.

4. 분석 도구

이 연구에서 진행된 평가결과는 SPSS ver 26.0을 이용하여 분석하였으며, 채점-재채점 결과의 신뢰도를 분석하기 위해 두 채점 결과 간 절대값의 차이를 검증하는 데 효과적인 급내 상관계수를 산출하였다. 또한 채점-재채점 결과의 상관관계를 파악하여 신뢰도를 분석하기 위해 Pearson 상관분석을 실시하였다.

이 연구에서는 FACETS ver 3.83을 활용하여 다국면 Rasch 모형에 따라 평가 대상인 글, 평가자인 국어교사, 논설문의 평가 요인의 3국면을 분석하고, 이에 따라 평가자의 일관성 유형을 분류하였다. 다국면 Rasch 모형[2]에서 각 국면의 분석 결과를 logit 척도로 변환하면, 고등학생의 쓰기 능력인 논설문 점수, 평가 요인의 난도에 따른 국어교사의 엄격성에 대한 값과 내적합 표준화값이 산출된다. 평가의 일관성은 내적합 표준화값 또는 내적합 지수를 기준으로 판단할 수 있으며 이에 따라 평가자를 적합, 과적합, 부적합 유형으로 분류하였다.

---

2) Linacre(1989)에 따르면, 다국면 Rasch 모형에서는 평가자 r이 고등학생 h가 쓴 논설문의 평가 요인 j에 대해 평가 점수가 k-1이 아닌 k를 점수로 부여할 확률과 그 확률을 log로 변환한 값인 logit 값을 얻을 수 있다.

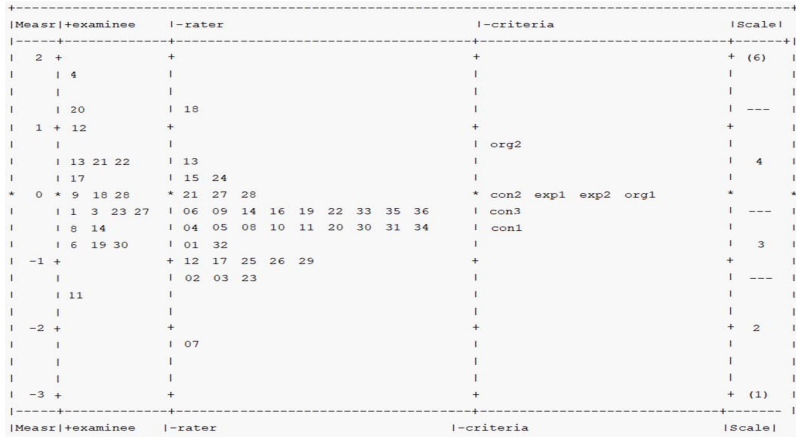
IV. 연구 결과 및 분석

1. 평가자의 일관성 유형 분석

평가자의 일관성 유형에 따라 채점-재채점 신뢰도를 분석하기 위하여 먼저 FACETS 프로그램을 활용하여 다국면 Rasch 모형 분석을 실시하였다. 이러한 결과를 통해 평가자의 엄격성과 일관성, 편향 등을 파악할 수 있는데, 이 연구에서는 다국면 Rasch 분석 결과에 따라 평가자의 일관성 유형을 적합, 과적합, 부적합으로 범주화하였다.

먼저 채점-재채점 결과가 다국면 Rasch 모형으로 분석하기에 적합한지를 확인해 보기 위해, FACETS 프로그램 실행 결과에 도출되는 모형적합도 그래프를 확인하였다. 그 결과 채점과 재채점 모두 모형적합도에서 관찰된 평가 점수인 '×'가 대부분 두 실선 사이에 있어 신뢰구간 내에 위치하고 있으므로, 다국면 Rasch 모형에 적합함을 알 수 있었다. 다국면 Rasch 분석 결과, 이 연구에서 설정한 세 가지 국면에 대한 정보는 〈그림 1〉과 같은 표로 제시된다. 〈그림 1〉은 채점 결과의 학생 글 × 평가자 × 평가 요인 분포도이다.

Vertical = (1N,2A,2*,3A,S) Yardstick (columns lines low high extreme)= 0,4, -3,2,End

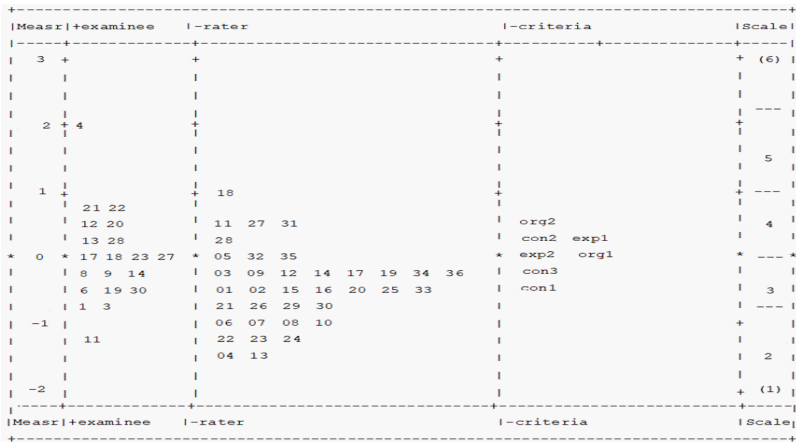


〈그림 1〉 채점 결과의 학생 글 × 평가자 × 평가 요인 분포도

〈그림 1〉의 'text'에는 학생 글 20편이 수준별로 표기되어 있다. 글 4는 최상위 수준의 글임을 의미하며, 글 11은 최하위 수준의 글임을 의미한다. 'rater'에는 엄격성의 정도에 따라 평가자 번호가 표기되므로 36명의 평가자 중, 평가자 18이 가장 엄격하게 평가하였으며 평가자 7이 가장 관대하게 평가하였음을 알 수 있다. logit 값이 0 정도인 평가자는 3명으로, 평가자 21, 27, 28이었다.

'criteria'에는 난도에 따라 평가 요인이 표기된다. 평가 요인 중 조직 2(문단을 적절히 나누었으며 문단 연결이 자연스럽고 긴밀한가?)가 가장 엄격하게 평가된 평가 요인이고, 내용 1(주장이 명료하고 타당한가?)이 가장 관대하게 평가된 평가 요인이다. 다음으로 〈그림 2〉는 재채점 결과의 학생 글 × 평가자 × 평가 요인 분포도이다.

Vertical = (1N,2A,2*,3A,S) Yardstick (columns lines low high extreme)= 0,4, -2,3,End



〈그림 2〉 재채점 결과의 학생 글 × 평가자 × 평가 요인 분포도

〈그림 2〉에 따르면, 글 4는 최상위 수준이며, 글 11은 최하위 수준임을 알 수 있다. 글 4와 11은 각각 채점, 재채점 결과 모두에서 최상위 수준의 글과 최하위 수준의 글로 나타났다. 또한 36명의 평가자 중, 평가자 18이 가장 엄격하게 평가하였고, 평가자 4, 13이 가장 관대하게 평가하였다. logit 값이 0 정도인 평가자는 3명으로, 평가자 5, 32, 35이었다. 평가자 18은 채점과 재채점 모두에서 가장 엄격하게 평가하였는데, 엄격성 정도를 구체적으로 살펴보면 채점 결과에서는 1.36 logit(SE=.10)을 보였으나, 재채점 결과에서는 1.11 logit(SE=.09)로 나타났다. 또한 채점에서 가장 관대하게 평가한 평가자 7의 경우, 채점 결과에서는 −2.21 logit(SE=.12), 재채점 결과에서는 −1.06 logit(SE=.09)으로 나타나, 재채점에서는 엄격성의 정도가 변화하였다.

평가 요인별 난도를 살펴보면, 평가 요인 중 조직 2(문단을 적절히 나누었으며 문단 연결이 자연스럽고 긴밀한가?)가 가장 엄격하게 평가된 평가 요인이며, 내용 1(주장이 명료하고 타당한가?)이 가장 관대하게 평가된 평가 요인으로 나타났다. 이는 채점 결과와도 동일하며, 조직 1(글의 전체 구조가 완

결성을 갖추고 있는가?)과 표현 2(어문규범(단어 표기, 맞춤법, 문장 주술 호응 등)에 맞게 표현되었는가?)가 logit 값 0 정도에 위치한 것 역시 채점과 재채점 결과에서 동일하게 나타났다. 이로 보아 국어교사들은 논설문 평가에 있어 조직 2를 가장 엄격하게, 내용 1을 가장 관대하게 평가하는 경향이 있음을 알 수 있다. 다음으로 채점 - 재채점 결과에 대해 다국면 Rasch 분석을 실시하여 평가자의 엄격성과 일관성을 살펴본 결과는 〈표 3〉과 같다.

〈표 3〉 채점-재채점의 엄격성과 일관성

평가자	채점				재채점			
	엄격성	측정값의 표준오차	내적합 지수	내적합 표준화값	엄격성	측정값의 표준오차	내적합 지수	내적합 표준화값
1	-0.77	0.10	0.50	-5.2	-0.59	0.08	0.68	-3.1
2	-1.19	0.10	1.37	2.8	-0.47	0.08	1.04	0.3
3	-1.16	0.10	1.41	3.1	-0.15	0.08	0.71	-2.9
4	-0.59	0.10	0.81	-1.7	-1.50	0.10	0.81	-1.5
5	-0.44	0.10	0.54	-4.7	0.05	0.08	0.66	-3.5
6	-0.34	0.10	0.47	-5.6	-1.12	0.09	0.80	-1.8
7	-2.21	0.12	0.76	-2.0	-1.06	0.09	0.97	-0.1
8	-0.42	0.10	0.55	-4.6	-1.04	0.09	1.74	5.2
9	-0.34	0.10	0.84	-1.4	-0.35	0.08	0.71	-2.8
10	-0.49	0.10	0.61	-3.8	-0.97	0.09	0.75	-2.3
11	-0.53	0.10	0.70	-2.8	0.43	0.08	1.01	0.1
12	-1.05	0.10	0.95	-0.3	-0.27	0.08	0.93	-0.6
13	0.46	0.09	0.75	-2.4	-1.38	0.09	0.92	-0.6
14	-0.29	0.10	0.48	-5.4	-0.29	0.08	0.58	-4.4
15	0.24	0.09	0.90	-0.8	-0.50	0.08	0.76	-2.3
16	-0.26	0.10	0.87	-1.1	-0.40	0.08	0.77	-2.2

17	-0.95	0.10	0.96	-0.3	-0.28	0.08	0.71	-2.8
18	1.36	0.10	0.96	-0.3	1.11	0.09	1.19	1.4
19	-0.19	0.09	0.82	-1.6	-0.23	0.08	0.70	-3.0
20	-0.59	0.10	1.03	0.2	-0.55	0.08	1.52	4.0
21	0.09	0.09	0.74	-2.4	-0.79	0.08	0.90	-0.8
22	-0.34	0.10	1.57	4.1	-1.31	0.09	1.49	3.5
23	-1.26	0.10	2.52	9.0	-1.26	0.09	0.97	-0.2
24	0.15	0.09	0.72	-2.6	-1.18	0.09	0.97	-0.2
25	-0.94	0.10	0.68	-2.9	-0.55	0.08	0.98	-0.1
26	-1.00	0.10	1.05	0.4	-0.74	0.08	1.29	2.4
27	-0.30	0.09	1.51	3.8	0.47	0.08	1.22	1.8
28	0.07	0.09	0.93	-0.6	0.24	0.08	0.81	-1.7
29	-1.00	0.10	2.47	8.9	-0.67	0.08	1.23	1.9
30	-0.47	0.10	0.48	-5.5	-0.86	0.08	0.76	-2.3
31	-0.39	0.10	0.77	-2.0	0.38	0.08	1.01	0.1
32	-0.77	0.10	1.34	2.6	-0.12	0.08	0.96	-0.2
33	-0.20	0.09	0.67	-3.2	-0.57	0.08	2.13	7.7
34	-0.40	0.10	0.59	-4.0	-0.33	0.08	1.12	1.1
35	-0.26	0.10	1.46	3.4	-0.12	0.08	1.27	2.3
36	-0.28	0.10	2.33	8.3	-0.31	0.08	0.94	-0.5
	표준오차 평균 0.46 측정값 표준편차 0.45, 분리도 8.39 분리지수 6.04, 분리신뢰도 0.97 실제 일치도 30.7%, 기대 일치도 29.0%				표준오차 평균 0.65 측정값 표준편차 0.59, 분리도 9.38 분리지수 6.79, 분리신뢰도 0.98 실제 일치도 25.7%, 기대 일치도 24.7%			

평가자 국면에 포함되는 모든 평가자들의 엄격성이 동일하다는 영가설을 검증한 결과, 채점 결과에서는 χ^2=1220.9, p=0.000으로 나타나 영가설이 기각되어 평가자들의 엄격성은 동일하지 않다고 판단할 수 있다. 또한 분리

신뢰도가 높을수록 평가자의 엄격성에 차이가 있다는 것을 의미하는데, 채점 결과의 분리신뢰도는 0.97로 나타나, 평가자의 엄격성은 변별된다고 볼 수 있다. 측정값의 표준오차는 0.09~0.12로, 0에 가깝게 나타났으므로 측정값은 정확한 것으로 볼 수 있다. 이에 채점에서 국어교사의 엄격성을 살펴보면 -2.21 logit(SE=.12)부터 1.36 logit(SE=.10)까지의 범위에 분포하고 있음을 알 수 있다.

다음으로 재채점 결과를 살펴보면, χ^2=1501.7, p=0.000으로 나타나 영가설이 기각되어 평가자들의 엄격성은 동일하지 않으며, 분리신뢰도 역시 0.98로 나타나 평가자의 엄격성이 변별된다고 볼 수 있다. 측정값의 표준오차는 0.08~0.10으로 나타나 재채점의 측정값 역시 정확한 것으로 볼 수 있다. 이에 재채점에서 국어교사의 엄격성을 살펴보면, -1.50 logit(SE=.10)부터 1.11 logit(SE=.09)까지의 범위에 분포하고 있음을 알 수 있다.

Linacre(1998)에 따르면, 내적합 표준화값을 기준으로 할 경우 -2.0 ~ +2.0 범위 안에 내적합 표준화값이 있을 때 일관성이 적합한 것으로 판정한다. 또한 McNamara(1996), 장소영·신동일(2009)에서 밝힌 바와 같이, 일반적으로는 다국면 Rasch 분석 결과 내적합 지수가 1.30 이상이면 부적합으로, 0.75 이하이면 과적합으로 볼 수 있다. Linacre & Wright(1990), McNamara(1996)에 따르면, 부적합은 평가자의 엄격성에 전혀 일관성이 없어 모형으로 예측이 불가능한 것을 말하며, 과적합은 평가자의 엄격성이 학생의 능력에 따라 변화하지 않고 특정한 점수에 편중되는 것을 말한다.

이 연구에서는 가장 일반적인 범위인 0.75~1.30을 기준으로 평가자의 일관성을 판단하였다. 따라서 1.30 이상이면 부적합, 0.75 이하이면 과적합으로 분류하였다. 일관성에 따라 평가자의 유형을 분석한 결과는 〈표 4〉와 같다.

〈표 4〉 채점-재채점의 일관성에 따른 평가자 유형

집단	채점의 일관성에 따른 평가자 유형 (채점-재채점)	평가자	인원 (%)
1	적합 – 적합	4, 7, 12, 15, 16, 18, 26, 28, 31	9(25.00)
2	과적합 – 적합, 적합-과적합	6, 9, 11, 13, 17, 19, 21, 24, 25, ,30, 34	11(30.56)
3	부적합 – 적합, 적합-부적합	2, 20, 23, 27, 29, 32, 35, 36	8(22.22)
4	과적합-과적합	1, 5, 10, 14	4(11.11)
5	과적합-부적합, 부적합-과적합	3, 8, 33	3(8.33)
6	부적합-부적합	22	1(2.78)

〈표 4〉에 따르면, 집단 1은 채점, 재채점 결과 모두 일관성이 적합한 경우로, 9명(25.00%)이 해당된다. 집단 2는 채점, 재채점 중 한 번은 적합이었으나 다른 한 번의 채점에서는 과적합으로 나타난 경우로, 11명(30.56%)이 해당된다. 집단 3은 채점, 재채점 중 한 번은 적합이었으나 다른 한 번의 채점에서는 부적합으로 나타난 경우로, 8명(22.22%)이 해당된다.

그리고 집단 4는 채점, 재채점 결과 모두 일관성이 과적합한 경우로, 4명(11.11%)이었다. 집단 5는 채점, 재채점 중 한 번은 과적합이었으나 다른 한 번의 채점에서는 부적합으로 나타난 경우로, 3명(8.33%)이었다. 마지막으로 집단 6은 채점, 재채점 결과 모두 일관성이 부적합한 경우로, 1명(2.78%)이었다. 즉 채점 – 재채점에서 일관성이 모두 적합한 경우는 25.00%, 한 번만 적합한 경우는 52.78%였다. 또한 채점 – 재채점 모두 과적합이나 부적합이 나타난 경우는 22.22%였다.

다음으로 국어교사 36명을 대상으로 채점 – 재채점 결과의 신뢰도를 분석하였다. 이를 위해 채점 – 재채점의 급내 상관계수를 평가자별로 산출하였으며, 채점 – 재채점의 상관관계를 파악하여 신뢰도를 분석하기 위해 Pearson 상관분석을 실시하였다. 그 결과는 〈표 5〉와 같다.

〈표 5〉 평가자별 채점-재채점 결과의 상관계수

평가자	집단	ICC	p	r	p
1	4	.701	.006	.584	.007
2	3	.655	.013	.487	.030
3	5	.459	.095	.300	.198
4	1	.318	.206	.191	.421
5	4	.894	.000	.809	.000
6	2	.564	.039	.407	.075
7	1	.830	.000	.775	.000
8	5	.595	.028	.494	.027
9	2	.675	.009	.512	.021
10	4	.798	.001	.712	.000
11	2	.488	.077	.466	.038
12	1	.662	.011	.513	.021
13	2	.548	.046	.397	.083
14	4	.386	.148	.282	.228
15	1	.582	0.32	.422	.064
16	1	.832	.000	.721	.000
17	2	.876	.000	.808	.000
18	1	.770	.001	.632	.003
19	2	.375	.157	.235	.319
20	3	.850	.000	.744	.000
21	2	.646	.014	.483	.031
22	6	.641	.015	.474	.035
23	3	.656	.012	.549	.012
24	2	.755	.002	.699	.001
25	2	.798	.001	.682	.001
26	1	.887	.000	.840	.000

27	3	.800	.000	.755	.000
28	1	.752	.002	.619	.004
29	3	.350	.178	.214	.366
30	2	.532	.053	.370	.109
31	1	.502	0.69	.373	.105
32	3	.698	.006	.545	.013
33	5	.631	.018	.485	.030
34	2	.648	.014	.522	.018
35	3	.605	.025	.435	.055
36	3	.822	.000	.763	.000

〈표 5〉에 따르면, 국어교사 36명 중 유의확률이 $p < 0.05$로 통계적으로 유의한 상관관계를 보이지 못한 경우는 11명으로 나타났다. 이는 전체 인원의 30.56%에 해당하는 것으로, 쓰기평가에서 채점-재채점의 신뢰도가 확보되지 못한 경우가 있음을 시사한다. 또한 통계적으로 유의한 상관관계를 보인 25명의 상관계수는 0.466에서 0.840까지 폭넓게 분포되어 있는 것으로 나타났다.

앞서 밝힌 바와 같이, 집단 1은 채점과 재채점의 일관성이 모두 적합한 것으로 나타났으며, 집단 4는 모두 과적합, 집단 6은 모두 부적합한 것으로 나타났다. 따라서 집단 1은 일관성 유형 중 적합에 해당하는 특징을, 집단 4는 과적합의 특징을, 집단 6은 부적합의 특징을 잘 보여 준다고 할 수 있다. 반면에 집단 2는 과적합과 적합의 양상이 혼재되어 있으며, 집단 3은 부적합과 적합의 양상이, 집단 5는 과적합과 부적합의 양상이 혼재되어 있다. 이에 일관성 유형에 따른 채점-재채점의 신뢰도를 분석하기 위하여 각각의 일관성 유형에 대한 특징을 잘 보여 주는 집단 1, 4, 6에 해당하는 평가자를 대상으로 하여 채점-재채점의 신뢰도를 구체적으로 분석하였다.

2. 일관성 유형에 따른 채점 - 재채점 신뢰도 분석

집단 1, 4, 6에 해당하는 국어교사 14명을 대상으로, 일관성 유형에 따른 채점 - 재채점 결과의 신뢰도를 분석하였다. 이를 위해 각 집단에 해당하는 평가자별로 채점 - 재채점의 기술 통계와 함께 급내 상관계수, Pearson 상관계수를 분석하였다. 그 결과는 〈표 6〉과 같다.

〈표 6〉 일관성 유형에 따른 채점-재채점 기술 통계 및 상관계수

집단	평가자		평균	표준 편차	ICC	p	r	p
1	4	채점	28.35	3.617	.318	.206	.191	.421
		재채점	34.55	3.220				
	7	채점	35.65	4.591	.830	.000	.775	.000
		재채점	31.95	7.037				
	12	채점	30.70	6.148	.662	.011	.513	.021
		재채점	26.25	8.065				
	15	채점	23.70	4.769	.582	0.32	.422	.064
		재채점	27.95	6.030				
	16	채점	26.55	5.835	.832	.000	.721	.000
		재채점	27.20	6.849				
	18	채점	17.30	8.392	.770	.001	.632	.003
		재채점	16.50	7.302				
	26	채점	30.45	7.193	.887	.000	.840	.000
		재채점	29.70	9.969				
	28	채점	24.70	7.012	.752	.002	.619	.004
		재채점	22.35	5.566				
	31	채점	27.25	4.166	.502	.069	.373	.105
		재채점	21.30	6.681				

4	1	채점	29.30	3.840	.701	.006	.584	.007
		재채점	28.65	5.733				
	5	채점	27.55	5.781	.894	.000	.809	.000
		재채점	23.75	5.999				
	10	채점	27.80	4.607	.798	.001	.712	.000
		재채점	31.30	6.729				
	14	채점	26.70	3.556	.386	.148	.282	.228
		재채점	26.35	6.409				
6	22	채점	27.00	5.341	.641	.015	.474	.035
		재채점	33.50	5.925				
전체		채점	27.36	6.688				
		재채점	27.24	8.137				

〈표 6〉에 따르면, 채점 결과의 평균 점수는 27.36, 표준 편차는 6.688이며, 재채점 결과의 평균 점수는 27.24, 표준 편차는 8.137로 나타나 채점 - 재채점 결과의 평균 점수가 상당히 유사함을 알 수 있다. 급내 상관계수는 두 가지 이상의 검사나 두 명 이상의 평가자가 실시한 평가 결과에 대한 일치도를 분석하는 방법으로, 일반적으로 급내 상관계수가 0.75 이상이면 높은 일치도를 보인 것으로 볼 수 있다. Pearson 상관계수는 두 변수가 서로 얼마나 관련이 있는지를 수치로, 이에 대한 해석은 연구자의 관점에 따라 설정하는 범위에 다소 차이가 있다. Mukaka(2012)에 따르면, 상관계수가 $r < 0.3$일 경우에는 상관관계가 나타나지 않는 것으로 보았으며, $0.3 \leq r < 0.5$일 경우 낮은 상관관계, $0.5 \leq r < 0.7$일 경우 중간 정도의 상관관계, $0.7 \leq r < 0.9$일 경우 높은 상관관계, $0.9 \leq r$일 경우 매우 높은 상관관계로 보았다. 이 연구에서는 Mukaka(2012)에서 밝힌 기준에 따라 상관계수를 분석하였다.

그 결과 통계적으로 유의한 상관관계가 나타나지 않은 경우가 4명(28.57%)이었으며, $0.3 \leq r < 0.5$로 나타나 낮은 상관관계를 보인 경우

가 1명(7.14%), 0.5 ≤ r 〈 0.7로 나타나 중간 정도의 상관관계를 보인 경우가 4명(28.57%), 0.7 ≤ r 〈 0.9로 나타나 높은 상관관계를 보인 경우가 5명(35.71%)이었다. 이는 평가자별로 쓰기평가의 채점 – 재채점 신뢰도가 상이함을 보여 준다. 통계적으로 유의하지 않은 상관관계를 나타낸 평가자 4명은 분석에서 제외하였으며, 상관계수에 따라 일관성 유형 집단에 따른 채점 – 재채점의 신뢰도를 분석한 결과는 〈표 7〉과 같다.

〈표 7〉 일관성 유형에 따른 집단별 채점-재채점의 신뢰도

상관의 정도 / 집단(명)	높은 (0.7≤ r 〈0.9)	중간 (0.5≤ r 〈0.7)	낮은 (0.3≤ r 〈0.5)	계
1	3	3	0	6
4	2	1	0	3
6	0	0	1	1
계				10

〈표 7〉에서 높은 상관관계를 보인 경우를 살펴보면, 집단 1은 3명으로 평가자 7, 16, 26이 해당되며, 집단 4는 2명으로, 평가자 5, 10이었다. 집단 6의 경우에는 높은 상관관계를 보인 평가자는 없었으며, 평가자 22가 낮은 상관관계를 나타냈다. 반면에 집단 1과 4의 경우에는 낮은 상관관계를 보인 평가자가 없는 것으로 나타났다. 이를 볼 때 평가자의 일관성이 부적합한 경우 채점 – 재채점의 신뢰도가 낮은 것을 알 수 있다.

다음으로 일관성 유형에 따라 특징적인 결과를 보인 평가자의 채점 – 재채점 신뢰도를 좀 더 구체적으로 살펴보고자 한다. 먼저 집단 1의 경우, 평가자 26은 14명의 평가자 중 상관계수가 가장 높게 나타나 채점 – 재채점의 신뢰도가 가장 높은 평가자였다. 상관계수는 0.840, p=0.000으로 나타나 가장 높은 상관관계를 보였으며, 급내 상관계수 역시 0.887로 나타나 채점 – 재채점 결과 간의 높은 일치도를 보였다. 평가자 26은 앞선 다국면 Rasch 분석에

서도 채점, 재채점 모두에서 일관성이 적합으로 나타나, 평가자 내 신뢰도가 높은 것을 알 수 있다.

그리고 평가자 26의 기술 통계를 살펴본 결과, 채점 결과 평균 점수는 30.45, 표준 편차는 7.193으로 나타났으며, 재채점 결과의 평균 점수는 29.70, 표준 편차는 9.969로 나타났다. 이를 볼 때 평가자 26의 경우 채점과 재채점 결과 간의 평균 차이는 0.75점으로 아주 근소한 것으로 나타나, 채점-재채점 결과의 평균이 유사한 평가자임을 확인할 수 있다.

또한 집단 1에서 상관계수가 0.7보다 높게 나타난 평가자들을 살펴보면, 평가자 7과 16으로, 평가자 7은 상관계수= 0.775, p=0.000로 나타났으며, 평가자 16은 상관계수= 0.721, p=0.000으로 나타나 높은 상관관계를 보였다. 다만 채점, 재채점 결과 간의 평균 차이는 두 평가자가 상이하게 나타났다. 평가자 16의 경우 채점과 재채점 결과 간의 평균 차이는 0.65점으로 아주 근소하게 나타났으나, 평가자 7은 재채점의 평균이 채점 결과에 비해 3.70점 낮은 것으로 나타나, 재채점에서 더 엄격하게 채점한 것으로 나타났다. 이는 다국면 Rasch 분석 결과에서도 동일하게 나타났는데, 〈표 3〉에서 평가자 7의 엄격성을 살펴보면, 채점에서는 -2.21 losit(SE=.12)으로 나타났으며 재채점에서는 -1.06 losit(SE=.09)으로 나타나 재채점에서 좀 더 엄격하게 채점하였음을 알 수 있다. 이러한 결과는 평가자 7과 16의 경우, 채점과 재채점 사이의 상관계수는 높게 나타나 채점의 신뢰도는 높다고 할 수 있으나, 각각에 적용한 엄격성이 달라 채점과 재채점의 결과가 상이함을 보여 준다.

평가자 18은 채점-재채점에서 모두 국어교사 14명의 평균 점수에 비해 현저히 낮은 점수를 부여하였다는 점에서 특징적이다. 국어교사 14명의 채점 결과의 평균 점수는 27.36점, 재채점 결과의 평균 점수는 27.24점인 반면, 평가자 18의 채점 결과의 평균 점수는 17.30이었으며 재채점 결과의 평균 점수는 16.50으로 나타났다. 이러한 평가 결과는 평가자의 엄격성과 관련된 것으로, 〈표 3〉에 따르면 평가자 18의 엄격성 정도는 채점에서 1.36 losit(SE=.10)으로 나타나 36명의 국어교사 중 가장 극단적인 엄격성을 보

였다. 또한 재채점에서도 1.11 losit(SE=.09)으로 나타나 36명의 국어 교사 중 가장 극단적인 엄격성을 보였다. 이러한 결과는 평가자의 엄격성이 평가자가 지니는 고유한 특성일 수 있음을 보여 준다. Lumley & McNamara(1995)에서 밝힌 바와 같이 평가자 교육 후에도 엄격성의 차이는 뚜렷하게 나타난다는 연구 결과도 이를 뒷받침한다.

그런데 평가자 18의 경우 상관계수는 0.632, p=0.003으로 나타나 중간 정도의 상관관계를 보였으며, 급내 상관계수 역시 0.770으로 나타나 채점-재채점 결과 간의 높은 일치도를 보였다. 즉 평가자 18은 다른 평가자들에 비해 학생 글에 대한 점수를 현저히 낮게 부여하였으나, 채점-재채점의 신뢰도는 양호한 것으로 나타났다. 또한 집단 1에 해당하므로 앞선 다국면 Rasch 분석 결과에서도 채점, 재채점 모두에서 일관성이 적합한 것으로 나타나, 평가자 내 신뢰도 역시 높은 것을 알 수 있다.

이를 볼 때, 평가자 18은 채점-재채점 신뢰도 및 평가자 내 신뢰도는 적합한 평가자이나, 극단적인 엄격성을 보여 20편의 학생 글에 대해 다른 평가자들에 비하여 현저히 낮은 점수를 부여한 것으로 보인다. 이러한 경우 평가자 간의 신뢰도가 확보되기 어려우므로 평가자의 평가 결과 일치 여부에 대한 정도를 제공하는 등 평가자 간의 신뢰도를 높이기 위한 방안이 요구된다.

또한 평가자 4, 15, 31을 살펴보면, 다국면 Rasch 분석 결과에서는 채점, 재채점이 모두 적합한 것으로 나타났으나 채점-재채점의 신뢰도를 보여주는 상관계수는 통계적으로 유의한 것으로 나타나지 않았다. 채점-재채점 간의 엄격성을 〈표 3〉을 통해 살펴보면, 해당 평가자들은 모두 채점-재채점 간의 엄격성에 차이가 있음을 확인할 수 있다. 예를 들어 평가자 4는 채점에서 −0.59 losit(SE=.10)이었으나 재채점에서는 −1.50 losit(SE=.10)으로 나타나 재채점에서 더 관대하게 평가하였음을 알 수 있다. 이러한 엄격성의 차이로 인해 채점-재채점 간의 평균 점수 차이도 크게 나타났는데, 평가자 4, 15, 31의 평균 점수 차이는 각각 6.20점, 4.25점, 5.95점으로 나타났다. 이러

한 결과는 다국면 Rasch 분석 결과만으로 평가자 내 신뢰도를 추정하기에는 어려움이 있을 수 있음을 시사한다. 즉 평가자 내 신뢰도를 추정하기 위한 분석 도구를 사용함에 있어 다국면 Rasch 모형 이외에도 좀 더 다양한 분석 도구가 활용되어야 함을 보여 준다.

다음으로 채점, 재채점에서 모두 과적합으로 나타난 집단 4에 해당하는 평가자를 살펴보면, 평가자 5와 10은 상관계수가 0.70이상으로 나타나 채점 – 재채점의 신뢰도가 높게 나타났다. 평가자 5의 경우 상관계수= 0.809, p=0.000으로 나타나 평가자 26 다음으로 가장 높은 상관관계를 보였다. 급내 상관계수 역시 0.894로 나타나 채점 – 재채점 결과 간의 높은 일치도를 보였다. 평가자 10 역시 상관계수=0.712, p=0.000으로 나타나 높은 상관관계를 보여, 평가자 5와 10은 과적합 평가자임에도 불구하고 채점 – 재채점의 상관계수는 높은 것으로 나타났다.

그러나 평가자 5와 10은 모두 채점과 재채점 결과 간의 평균 차이가 각각 3.80점과 3.50점으로, 채점과 재채점 간의 평균 차이가 크게 나타났다. 평가자 5는 재채점에서 채점을 엄격하게 하여 평균 점수가 3.80점 하락하였으며, 이러한 결과는 다국면 Rasch 분석 결과에서 나타난 엄격성의 차이로도 확인할 수 있다. 〈표 3〉에 따르면, 평가자 5의 엄격성 정도는 채점에서 –0.44 losit(SE=.10)으로 나타났으나, 재채점에서는 0.05 losit(SE=.08)으로 나타났다. 반면에 평가자 10의 경우 재채점에서 채점을 관대하게 하여 평균 점수가 3.50점 상승하였으며, 다국면 Rasch 분석 결과에서도 0.48 losit의 차이를 보여 재채점에서 관대하게 평가하였음을 알 수 있다.

이러한 결과를 볼 때, 평가자 5와 10은 채점과 재채점 간의 상관계수는 높게 나타났으나 다국면 Rasch 분석 결과와 채점 – 재채점의 평균 차이를 분석해 볼 때, 평가자 내 신뢰도가 높은 평가자로 분류하기에는 다소 어려움이 있어 보인다. 따라서 집단 4의 경우에는 집단 1과 같이 채점 – 재채점의 상관관계가 높으면서도 채점 – 재채점의 평균 차이도 근소하여 반복된 채점에서 동일한 양상의 결과를 보인 평가자는 없음을 알 수 있다. 이는 평가자 내 신

뢰도를 파악함에 있어, 다국면 Rasch 분석 이외에도 채점-재채점의 신뢰도와 같은 다양한 방법을 종합적으로 검토해 볼 필요가 있음을 시사한다.

마지막으로 채점, 재채점에서 모두 부적합으로 나타난 집단 6에 해당하는 평가자를 살펴보면, 평가자 22는 상관계수=0.474, p=0.035로 나타나, 유의확률이 $p < 0.05$로 통계적으로 유의한 상관관계를 나타낸 경우 중 가장 낮은 상관관계를 보였다. 즉 평가자 22는 앞선 다국면 Rasch 분석 결과와 채점과 재채점 결과 모두에서 일관성이 부적합하여 평가자 내 신뢰도가 낮은 것으로 나타났다. 그리고 평가자 22의 기술 통계를 살펴본 결과, 채점 결과 평균 점수는 27.00, 표준 편차는 5.341로 나타났으며, 재채점 결과의 평균 점수는 33.50, 표준 편차는 5.925로 나타났다. 즉 평가자 22는 채점과 재채점 결과의 평균 점수 차이가 6.5점으로 크게 나타나 통계적으로 유의한 평가자 중 가장 큰 차이를 보여, 평가자 22는 채점-재채점의 결과가 서로 상이함을 알 수 있다.

이를 종합해 볼 때 평가자의 일관성이 적합한 집단 1에서는 채점-재채점 신뢰도가 높게 나타나 동일한 글에 대한 반복적인 채점에서도 유사한 결과가 나타나는 경향이 있음을 보여 준다. 과적합 양상을 보인 집단 4에서도 채점-재채점의 신뢰도가 높게 나타나는 경우가 있었으나 이 경우에는 채점과 재채점에서의 엄격성이 상이하여 평균 차이가 크게 나타났다. 이는 평가자의 신뢰도를 분석하기 위해서는 채점-재채점의 신뢰도와 더불어 평가자의 엄격성 역시 고려되어야 함을 보여 준다. 또한 부적합 양상을 보인 집단 6의 경우 채점-재채점 결과 낮은 상관관계를 보여 신뢰도가 가장 낮게 나타나, 동일한 글에 대한 반복적인 채점에서 상이한 결과가 도출됨을 알 수 있다. 이는 평가자의 일관성 유형이 부적합으로 나타나는 경우에는 평가의 신뢰도를 확보하기 위한 다양한 방안이 모색되어야 함을 시사한다.

이외에도 평가자 14명의 상관계수가 0.474에서 0.840까지 폭넓게 분포되어 있어 채점-재채점의 신뢰도가 다양하게 나타난 점, 통계적으로 유의한 상관관계를 보이지 못할 정도로 채점과 재채점 결과 간의 관련성이 떨어진

평가자가 있음을 고려할 때, 쓰기평가에서 채점-재채점 신뢰도의 차이는 학교 현장의 다양한 쓰기평가 상황에서 학생 글에 대한 평가 결과의 차이로 이어질 가능성이 높다. 이러한 결과는 간접 평가에서 타당성과 신뢰성이 높은 평가 도구를 개발하여 학생의 능력을 정확하게 평가하고자 노력하는 것과 마찬가지로, 직접 평가에서는 학생의 능력이 정확하게 평가될 수 있도록 평가자 연수나 교육에 많은 노력이 요구된다는 것을 시사한다. 더불어 다국면 Rasch 분석 결과와 채점-재채점 결과의 신뢰도를 살펴본 결과, 평가자 내 신뢰도를 추정하기 위한 분석 도구의 정합성에 대한 좀 더 심층적인 연구가 요구됨을 알 수 있다.

그동안 평가자의 일관성은 평가자 내 신뢰도로 명명되어 왔다. 그러나 다국면 Rasch 모형에서의 내적합 표준화값이나 내적합 지수를 기준으로 판단되는 평가자의 일관성은 수치로 표현될 뿐이므로, 평가 결과에 어떠한 차이를 나타내는지 또는 어느 정도의 평가 신뢰도를 보이는지를 직접적으로 가늠하기는 어려웠다. 앞서 제시한 연구 결과는 적합, 과적합, 부적합에 해당하는 평가자의 일관성에 따라 채점-재채점의 상관관계를 관련지어 살펴봄으로써, 동일한 채점을 반복하여 시행한 경우에 평가자의 일관성 유형에 따라 어떠한 평가 결과가 나타나는지를 가시적으로 보여 주었다는 점에서 의의가 있다.

V. 결론

2015 개정 교육과정의 도입과 더불어 쓰기평가는 국어 교과의 경계를 넘어 다양한 교과의 평가 장면에서 활용되고 있다. 이에 쓰기평가의 직접 평가 도입에 대한 관심이 증대됨에 따라, 쓰기평가와 관련된 다양한 연구와 논의가 진행되고 있다. 이 연구는 국어교사의 쓰기평가 일관성 유형에 따라 채

점-재채점 신뢰도를 분석하는 데 목적이 있으며, 이를 통해 쓰기 평가자로서의 국어교사의 역할을 재고해 보고자 하였다.

이 연구에서는 국어교사 36명을 대상으로 학생 글 20편에 대한 채점-재채점을 실시하였으며, 먼저 다국면 Rasch 분석을 통해 평가자의 유형을 적합, 과적합, 부적합으로 범주화하였다. 그 결과 다국면 Rasch 분석을 통해 평가자의 일관성이 적합한 것으로 나와 채점-재채점 모두 평가자 내 신뢰도를 유지한 경우는 25.00%로 나타났으며, 두 번의 채점에서 한 번만 적합한 것으로 나타난 경우는 52.78%였다. 또한 채점-재채점 모두 과적합이나 부적합으로 나와 평가자 내 신뢰도가 유지되지 못한 경우는 22.22%로 나타났다. 다음으로 평가자별 채점-재채점의 신뢰도를 살펴보면, 국어교사 36명 중 11명의 평가자가 통계적으로 유의한 상관관계를 보이지 못하였으며, 25명의 상관계수는 0.466에서 0.840까지 폭넓게 분포되어 있는 것으로 나타났다.

일관성 유형에 따라 집단을 분류한 결과, 집단 1은 채점-재채점의 일관성이 모두 적합한 것으로 나타났으며, 집단 4는 모두 과적합, 집단 6은 모두 부적합한 것으로 나타났다. 이에 집단 1, 4, 6은 각각 적합, 과적합, 부적합의 특징을 잘 보여 주는 평가자로 볼 수 있으므로, 일관성 유형에 따른 채점-재채점의 신뢰도를 분석하기 위하여 집단 1, 4, 6에 해당하는 평가자 14명을 대상으로 채점-재채점의 신뢰도를 분석하였다.

그 결과, 평가자의 일관성이 적합한 집단 1에서는 채점-재채점 신뢰도가 높게 나타나 동일한 글에 대한 반복적인 채점에서도 유사한 결과가 나타나는 경향을 보였다. 과적합 양상을 보인 집단 4에서도 채점-재채점의 신뢰도가 높게 나타나는 경우가 있었으나 이 경우에는 채점과 재채점에서의 엄격성이 상이하여 평균 차이가 크게 나타났다. 이는 평가자의 신뢰도를 면밀하게 분석하기 위해서는 채점-재채점의 신뢰도와 더불어 평가자의 엄격성 역시 고려되어야 함을 보여 준다. 그리고 부적합 양상을 보인 집단 6의 경우, 채점과 재채점의 결과가 낮은 상관관계를 보임으로써 신뢰도가 가장 낮게 나타나, 동일한 글에 대한 반복적인 채점에서 상이한 결과가 도출됨을 보여

준다. 이는 평가자의 일관성이 부적합한 경우 채점과 재채점의 신뢰도가 낮아, 학생 글에 대한 평가 결과의 차이로 이어질 수 있음을 시사한다. 이와 더불어 평가자별 상관계수가 0.474에서 0.840까지 채점-재채점의 신뢰도가 다양하게 나타난 점, 통계적으로 유의한 상관관계를 보이지 못할 정도로 채점과 재채점 결과 간의 관련성이 떨어진 평가자가 있음을 고려할 때, 평가의 신뢰도를 확보하기 위한 다양한 방안이 모색되어야 함을 알 수 있다.

쓰기평가의 궁극적 목적은 학생들의 쓰기 능력 신장에 있다. 이를 위해서는 학생의 쓰기 능력과 학생 글의 특성을 정확히 분석하여 적절한 피드백을 제공할 필요가 있다. 평가자의 신뢰도는 학생의 쓰기 능력을 정확하게 판단하도록 이끌며, 이를 바탕으로 한 피드백이 가능하도록 할 수 있다. 이 연구는 일관성의 유형에 따라 채점-재채점 신뢰도를 분석함으로써 신뢰도와 관련된 평가자의 특성을 좀 더 면밀히 탐색해 보고자 하였다. 그 결과 채점-재채점에서의 평가자 신뢰도가 서로 상이함은 물론 평가자 내 신뢰도를 정확하게 추정하기 위해서는 다국면 Rasch 분석 이외에도 채점-재채점 신뢰도 분석과 같이 다양한 분석 도구가 함께 활용될 필요가 있음을 확인하였다. 쓰기평가에서 평가자를 대상으로 채점-재채점을 실시하여 그 결과에 대한 신뢰도를 분석한 연구가 부재함을 고려할 때, 이 연구 결과가 평가자의 특성과 신뢰도를 탐색하는 데에 있어 기초 자료가 되기를 기대한다. 더 나아가 이 연구 결과를 바탕으로 평가자의 신뢰도를 높일 수 있는 다양한 방안을 모색하는 후속 연구가 이어지길 바란다.

* 본 논문은 2021.1.31. 투고되었으며, 2021.2.18. 심사가 시작되어 2021.3.16. 심사가 종료되었음.

참고문헌

가은아(2011),「쓰기 발달의 양상과 특성 연구」, 한국교원대학교 박사학위논문

강승호 · 김양분(2004),『신뢰도』, 서울: 교육과학사.

권대훈(2008),『교육평가』, 서울: 학지사.

권태현(2014),「쓰기 성취기준에 따른 학생 예시문 선정에 관한 연구」, 한국교원대학교 박사학위논문.

박도순 · 권순달 · 김명화 · 김영애 · 김정민(2012),『교육평: 이해와 적용』, 서울: 교육과학사.

박영민 · 이재기 · 이수진 · 박종임 · 박찬홍(2016),『작문교육론』, 서울: 역락.

서수현(2012),「쓰기평가 협의 과정에 나타난 쓰기 평가자의 인식 연구」,『국어교육학연구』44, 335-367.

성낙수 · 김슬옹 · 김홍범 · 안주호 · 양정석 · 이정택 · 이창덕 · 한길 · 박영민 · 박종임 · 박형우 · 유혜령 · 윤천탁 · 이재형 · 최숙기(2015),『국어학과 국어교육학』, 서울: 채륜

성태제(2002),『타당도와 신뢰도』, 서울: 학지사.

이재창 · 김지은 · 김민영 · 양지선 · 한명훈 · 권혁찬 · 김기웅 · 임상현 · 정은의 · 김지웅 · 임우영 · 이상민 · 김승준(2017),「조현병 환자가 시행한 주의력 네트워크 검사 점수의 검사-재검사 신뢰도」,『정신신체의학』25(2), 210-217.

이혜원 · 이경원(2014),「학령전기용 어음청각검사 어표의 검사-재검사 신뢰도」,『Audiology and Speech Research』10(1), 25-34.

장소영 · 신동일(2009),『언어교육평가 연구를 위한 FACETS 프로그램』, 서울: 글로벌콘텐츠.

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

허소희(2019),「한국어 이명주요기능설문지의 검사-재검사 신뢰도 검증 및 간소화 버전 제안」, 한림대학교 석사학위논문.

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

Eckes, T. (2008), "Rater types in writing performance assessments: A classification approach to rater variability", *Language Testing 25*(2), 155-185.

Gwet, K. L. (2012), *Handbook of Inter-Rater Reliability*, (3rd Ed.), Gaithersburg, MD: Advanced Analytics, LLC.

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

Linacre, J. M. (1998), "Rating, judges, and fairness", *Rasch Measurement Transactions*, *12*(2), 630-631.

Linacre, J. M. & Wright, B. D. (1990), *FACETS*, Chicago, IL: MESA Press.

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

McNamara, T. F. (1996), *Measuring Second Language Performance*, London: Longman.
Mukaka, M. M. (2012), "A guide to appropriate use of correlation coefficient in medical research", *Malawi Medical Journal 24*(3), 69-71.
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.
Ruth, L., & Murphy, S. (1988), *Designing writing tasks for the assessment of writing*, new york, NY: Ablex Publishing Corporation.
White, E. M. (1984), "Holisticism", *College Composition and Communication 35*(4), 400-409.

초록

국어교사의 쓰기평가 일관성 유형에 따른 채점-재채점 신뢰도 분석

윤금준 · 박영민

이 연구의 목적은 국어교사의 쓰기평가 일관성 유형에 따라 채점-재채점 신뢰도를 분석하는 데 있다. 이를 위해 국어교사 36명을 대상으로 학생 글 20편에 대해 채점-재채점을 실시하였으며, 다국면 Rasch 분석 결과에 따라 일관성의 유형을 기준으로 평가자를 집단별로 범주화하였다. 그 결과 평가자의 일관성이 적합한 경우는 25.00%였으며, 한 번만 적합한 것으로 나타난 경우는 52.78%, 채점-재채점 모두 과적합이나 부적합인 경우는 22.22%로 나타났다. 일관성 유형에 따른 채점-재채점의 신뢰도를 분석한 결과 일관성이 적합한 경우에는 채점-재채점 신뢰도가 높게 나타나는 경향을 보였다. 또한 과적합일 때 채점-재채점의 신뢰도가 높게 나타나는 경우가 있었으나 이 경우에는 채점과 재채점 간의 엄격성이 상이하여 평균 차이가 크게 나타났다. 마지막으로 부적합인 경우에는 채점-재채점의 신뢰도가 가장 낮게 나타났다. 이는 평가자에 따라 평가 결과가 상이할 수 있음을 보여 주는 것으로, 쓰기평가의 신뢰도를 확보하기 위한 다양한 방안이 모색되어야 함을 시사한다.

핵심어 국어교사, 쓰기평가, 평가자 일관성, 채점-재채점 신뢰도, 다국면 Rasch 모형

ABSTRACT

Reliability Analysis of Scoring-Rescoring Depending on the Consistency Type of Korean Language Teachers' the Writing Assessment

Yoon Keumjoon · Park Youngmin

The purpose of this study is to analyze the reliability of raters based on the results of the writing assessment scoring-rescoring of Korean language teachers. For this purpose, 36 Korean language teachers scored and re-scored 20 students' writings, and they were categorized into groups based on the type of consistency according to the results of the multi-faceted Rasch analysis. As a result, 25.00% of raters showed consistency in both scoring and rescoring and 52.78% of raters showed consistency only once. In addition, 22.22% of the raters were found to have failed to maintain consistency. Looking at the reliability of scoring and rescoring, the scoring-rescoring reliability tends to be high when consistency is suitable. In case of over-fitting, the reliability of scoring and rescoring was high, but the mean difference was significant because of the difference of the severity in scoring and rescoring. In case of mis-fitting, the reliability of scoring-rescoring was the lowest. This suggests that various measures should be sought to ensure the reliability of the writing assessment.

KEYWORDS Korean language teachers, writing assessment, rater consistency, scoring-rescoring reliability, many-facet Rasch measurement