『이중언어학』 제61호(2015.12.31). 二重言語學會

쓰기 평가에서 채점 특성에 대한 피드백이 채점에 미치는 영향 연구* **

강민석 · 심재경***

Abstract

Kang Min-suk & Sim Je-kyoung. 2015. 12. 31. **The Effects of Individualized Feedback on Rating Behaviors in Korean Writing Assessment.** *Bilingual Research 61*, 1-29. The purpose of this study is to investigate the effect of individualized feedback to raters of Korean writing assessment on their severity, consistency, and bias. 20 scripts were rated independently by 22 Korean language teachers. The many-facet Rasch model was used to generate individualized feedback reports on each rater's relative severity, overall consistency, and significant bias patterns with respect to particular categories of the rating scale. Reports were given and explicated to each rater at a feedback session. Raters were then asked to rate a further set of 20 scripts. A comparison of ratings before and after feedback revealed that individual feedback had positive effects on their ratings (again using the many-facet Rasch model). The range of raters' severity was reduced after the feedback session, indicating a higher level of agreement between raters compared to the ratings before the feedback. The individualized feedback also had an influence on the rater consistency; raters who were classified as over-fitted or mis-fitted improved their consistency after receiving the feedback. In addition, rater bias in relation to particular categories was all removed. **(Korea University)**

---

* 본 논문은 국제한국어교육학회 제25차 국제학술대회에서(2015년 8월) 발표한 내용을 수정, 보완한 것임.

** 이 논문은 BK21 플러스 고려대학교 한국어문학 미래인재육성사업단의 지원으로 작성되었음.

*** 강민석: 제1저자, 심재경: 교신저자

【Key words】 rater training(채점자 훈련), rater(채점자), many-facet Rasch model(다국면 라쉬 모형), writing assessment(쓰기 평가)

1. 서론

본 연구는 한국어 쓰기 평가에서 채점자 개인의 채점 특성에 대한 피드백이 채점자의 엄격성, 일관성, 편향성에 긍정적인 영향을 미칠 수 있는지를 검증하는 것을 목적으로 한다.

지금의 한국어 교육 현장에서 쓰기 평가는 대부분 직접평가의 형태로 이루어지고 있다. 주로 객관식 문항을 통해 쓰기 지식을 간접적으로 측정하던 전통적인 지필평가에서는 수험자의 응답이 정·오답으로 분류되어 곧바로 점수화되었으나, 현재와 같은 직접평가 방식의 쓰기 평가에서는 수험자의 쓰기 결과물을 바탕으로 채점자가 수험자의 쓰기 능력을 추정해 나가는 채점 과정이 포함되기 때문에, 평가에 있어서 채점자의 역할이 매우 중요해졌다.

그런데 복잡한 인간의 수행을 채점하는 과정에는 채점자의 주관적인 판단이 필연적으로 개입되며, 따라서 채점자들 간의 불일치가 발생하기 쉽다(McNamara, 1996:117). 비록 이러한 불일치가 필연적이라 하더라도 그 정도에 따라 평가의 신뢰성 및 공정성에 큰 문제를 일으킬 수도 있다. 이에 여러 연구들이 채점자 요인을 평가의 신뢰도와 공정성에 영향을 줄 수 있는 가장 결정적인 요인으로 주목하고 있으며(Shohamy et al, 1992; Linacre, 1994; Eckes, 2011 등), Ebel & Frisbie(1991:196)는 채점자의 능숙도가 작문형 시험의 유효성을 좌우하는 채점 과정에 있어서 가장 결정적인 요인이라고도 하였다.

따라서 공정하고 신뢰할 수 있는 쓰기 평가의 시행을 위해서는 채점자의 올바른 채점이 필수적이다. 그리고 제2 언어 쓰기 평가와 관련하여 채점자들의 주관성을 조정하거나 채점 전문성을 향상시키기 위한 방안

으로 제안되어 온 가장 대표적인 방안은 채점자 훈련이다. 한국어교육의 경우에도, 한국어능력시험 쓰기 영역의 채점에서 신뢰도를 확보하기 위한 채점자 훈련이 실시되고 있다. 그러나 이는 선발된 소수의 채점자만을 대상으로 시행되며, 대부분의 한국어 교사들은 공식적인 채점자 훈련을 경험하지 못한 채 교육 현장에서 쓰기 채점을 담당하고 있다.

또한 채점자 훈련은 주로 집체교육의 형태로 이루어지게 되는데, 이러한 형태의 교육은 시간이나 비용 등의 면에서 다수의 교사들을 대상으로 지속적으로 시행되기에는 한계가 있다. 대부분의 한국어 교육 기관에서 채점자 훈련의 필요성을 공감하면서도 실제 훈련이 이루어지지 못하고 있는 상황도 이러한 이유에서 기인하였을 가능성을 배제할 수 없다. 한편으로는 채점자 훈련이 이루어진다고 할지라도 단순히 채점 기술이나 전략에 대한 교육, 채점 결과의 비교를 통해서는 개별 채점자의 채점 특성이나 채점자 간 점수 불일치에 대한 원인 등을 알 수 없으며, 이렇게 이루어지는 훈련에서는 채점자들의 채점 자체의 개선보다는 총점의 일치를 위한 노력이 우선되는 경향이 없지 않다.

이에 본 연구는 채점자들의 채점을 개선시킬 수 있는 한 방안으로 채점 특성에 대한 개별적인 피드백을 제안하고자 한다. 이 피드백의 주된 목적은 채점자들이 자신의 채점 특성을 인지하고, 문제가 되는 부분이 있다면 스스로 조정해 나가도록 유도하는 것이다. 이와 관련하여, 한국어 쓰기 채점자들의 채점 과정을 사고 구술법을 통해 분석한 김동수(2014)에서는 채점자가 자신의 채점 특성을 인식하고 있는 경우 채점 과정에서 스스로 엄격성을 조정하는 모습을 관찰한 바 있으며, 인접학문인 제2 언어로서의 영어 교육 분야의 Elder et al(2005)에서는 채점자들에게 개별적인 피드백을 주었을 때 편향성의 출현 정도가 감소하는 것이 발견되기도 하였는데 본 연구에서는 이러한 점에 착안하여, 한국어 쓰기 평가에서 채점자들이 자신의 채점 특성을 인지하는 것이 채점의 엄격성, 일관성, 편향성에 미치는 효과를 검증해 보고자 한다.

이를 위하여 본 연구에서는 한국어 교사 22명이 20편의 글에 대하여 채점한 결과에서 각 채점자들의 채점 특성을 분석하여 개별적으로 피드백을 실시하는 한편, 피드백이 이루어진 후에 채점자들에게 다시 20편의 글을 채점하도록 하여 채점에 어떠한 변화가 있는지를 비교하였다. 채점 결과의 분석에는 다국면 라쉬 모형(many-facet Rasch model)을 활용하였다. 본 연구의 결과는 실증적인 근거를 통해 한국어 쓰기 채점자들의 채점을 개선할 수 있는 하나의 구체적인 방안을 제시할 수 있을 뿐만 아니라, 나아가 쓰기 평가에서의 채점과 채점자에 대한 이해를 증진시키는 데에도 보탬이 될 수 있을 것으로 예상된다.

2. 쓰기 평가에서의 채점자

2.1. 선행 연구

한국어교육 분야에서 쓰기 평가의 신뢰도 확보에 대한 논의는 지속적으로 이루어져 왔다. 그리고 이러한 논의들은 쓰기 평가의 특성상 '채점 방식' 또는 '채점자'를 중심으로 진행되었고, 채점 방식에 있어서는 대표적으로 김정숙(2010), 이인혜(2012)를 통해 분석적 채점 방식(analytic scoring)이 종합적 채점 방식(holistic scoring)에 비해 신뢰도 확보에 더 효과적임이 밝혀진 바 있다. 그런데 앞에서도 언급하였듯이 평가의 신뢰성 및 공정성에 영향을 줄 수 있는 가장 결정적인 요인은 채점자이다. 따라서 최근에는 채점자의 채점 과정(최은지, 2013; 김동수, 2014), 채점자 훈련(최윤곤·허유리, 2013; 허유리, 2014), 또는 채점자의 다양한 배경 변인에 따른 채점 특성(이정화, 2010; 이인혜, 2014; 강민석, 2015) 등 쓰기 평가에서 채점자에 대한 이해를 높이거나 채점 결과에 영향을 줄 수 있는 채점자 요인에 주목한 연구들이 주로 이루어지고 있다.

최은지(2013)에서는 사고 구술법을 통해 채점자들이 채점을 하는 과정

에서 ‘세부 평가 기준에 의존한 형식적 평가’, ‘엄격성의 차이’, ‘인상 평가에 의한 피드백의 부재’, ‘평가 항목에 대한 서로 다른 인식’, ‘내적 평가 기준의 흔들림’ 등으로 인해 채점자 간의 점수 차이가 일어남을 분석하였고, 김동수(2014), 강민석(2015)에서는 채점자들이 숙련도 차이에 따라 서로 다른 채점 특성을 나타낸다는 것을 각각 사고 구술법과 다국면 라쉬 모형을 통해 확인하기도 하였다. 또한 이인혜(2014)에서는 한국어 교사들의 평가 효능감과 채점 특성 사이에 일정한 관련성이 있음을 밝혀내기도 하였는데, 효능감이 낮은 교사들은 일관성에서 문제를 보이거나 편향된 채점을 하는 경향을 보였다.

위의 연구들은 한국어 쓰기 평가에서 채점자의 채점 과정과 채점 특성을 본격적으로 논의하였다는 것에 의의가 있으나, 이를 바탕으로 채점자들의 채점을 개선할 수 있는 방안을 모색하는 것까지는 이어지지 못하였다. 구체적인 방안을 제안하는 것에 초점을 둔 연구로는 최윤곤·허유리(2013), 허유리(2014)가 유일하다고 할 수 있는데, 이 연구에서는 채점자들의 쓰기 채점 표본과 설문조사를 바탕으로 채점자 훈련 모형을 제안하고 있다. 그러나 제안된 훈련 모형이 다소 구체적이지 못하며, 연구에서도 스스로 밝히고 있듯이 채점자 훈련을 실제로 진행하지 않아 그 효과를 검증하지 못하였다는 한계를 지닌다.

이처럼 한국어 쓰기 평가에서 채점자들의 채점을 개선시킬 수 있는 구체적인 방안을 제안하고 그 효과를 검증한 연구는 아직까지 이루어지지 않았다. 따라서 본고와 같이 특정한 방안의 효과를 실증적으로 확인하는 연구들이 축적되어 나가야 할 시점이라고 할 수 있겠다.

2.2. 채점자의 채점 특성

쓰기 채점에서 채점자들이 보이는 채점 특성에는 대표적으로 엄격성,

일관성, 편향성이 있다. 그리고 이들에서 나타나는 문제는 채점의 신뢰성과 공정성에 직접적인 영향을 주게 된다.

먼저 엄격성(severity)은 특정한 채점자가 지니는 채점 기준이 다른 채점자들과 비교하였을 때 엄격하거나 관대한 정도를 나타낸다. 엄격성은 채점자들이 가지고 있는 고유의 특성이기 때문에 마치 개개인의 성향이 모두 다르듯이 엄격성의 정도도 매우 다양하게 나타날 수 있으며, 실제로 쓰기 평가에서 채점자들의 엄격성에 큰 차이가 나타나는 것이 많은 연구들을 통해 증명되어 왔다(Lumley & McNamara, 1995; Weigle, 1998; Eckes, 2008; 이인혜, 2014; 강민석, 2015).

그런데 이러한 엄격성의 차이는 동일한 쓰기 결과물이 서로 다른 점수로 측정되는 가장 큰 원인이 된다. 이와 관련하여 Tyndall & Kenyon(1996)에서는 엄격성이 가장 높은 채점자와 가장 낮은 채점자의 엄격성 격차로 인해 무려 50%에 가까운 점수 차이가 발생할 수 있음을 지적하기도 하였다. 비록 엄격성의 차이가 채점자들이 가진 고유의 채점 기준으로부터 비롯된다고 하더라도, 수용되기 어려운 정도의 차이는 신뢰도의 문제로 이어질 수 있기 때문에 객관적이고 신뢰할 수 있는 평가의 시행을 위해서는 채점자들의 엄격성 차이를 좁혀 나가는 것이 바람직하다.

다음으로 일관성(consistency)은 채점자가 채점의 전 과정에서 엄격성을 일관되게 유지하는 정도를 의미한다. 즉, 일관성은 채점자가 얼마나 일관된 기준을 가지고 채점을 했는지를 나타내는 특성이다. 일관성은 채점자의 역량, 전문성과 가장 밀접한 관련이 있으며, 채점자의 일관성에 문제가 있다는 것은 채점을 하는 글마다 적용하는 엄격성이 흔들려 엄격한 채점자인지 관대한 채점자인지를 판단하는 것 자체가 불가능하다는 것을 의미한다.

McNamara(1996:125)는 일관성에 문제를 보이는 채점자는 반드시 채점자 훈련을 받아야 하며, 만약 훈련 이후에도 일관성이 나아지지 않는

경우에는 해당 채점자를 채점에서 제외시키는 방법 외에는 해결책이 없다고 하였을 정도로 채점자의 일관성 문제를 심각하게 바라보고 있다. 그런데 McNamara(1996)가 말하고 있는 일관성의 문제는 낮은 일관성만을 지칭하는 것이 아니다. 채점자의 일관성이 지나치게 높아도 문제가 되는데, 이는 채점자가 지나치게 예측 가능한 채점을 한다는 것을 의미한다. 즉, 이러한 채점자는 대부분의 글에 엇비슷한 점수를 주거나 혹은 인상 평가와 같이 잘 쓴 글에는 적당히 높은 점수를 주고 나머지 글에는 비슷하게 낮은 점수를 부여하는 등의 채점을 하고 있을 가능성이 높다. 이러한 채점자들은 채점의 변별력이 떨어질 뿐만 아니라 분석적 채점에 어울리지 않게 일부 채점 요소에 이끌려 채점하는 등의 문제를 드러낸다(Eckes, 2011; 박영민, 2012).[1)]

마지막으로 편향성(bias)이란 채점자가 특정 과제나 수험자 집단 등에 대해서만 지속적으로 엄격하거나 관대한 채점을 하는 경향을 의미한다(Lumley & McNamara, 1995:56). 그리고 분석적 채점에서 채점자들의 편향성은 특정 평가 범주에 대해서 나타나는 경우가 잦다. 예를 들어, 어떤 채점자는 언어 사용에는 상대적으로 매우 엄격하면서도 글의 전개 구조에는 대부분 최고점을 줄 정도로 관대할 수도 있다. 이러한 채점은 글을 조리 있게 조직하는 것에는 서툴지만 언어 사용에 능숙한 수험자에게 매우 유리하게 작용할 것이다.[2)]

---

1) 다국면 라쉬 모형에서는 일관성의 부족을 '부적합(misfit)'으로, 일관성의 과잉을 '과적합(overfit)'이라는 용어를 사용하여 나타낸다. 4장의 연구 결과의 기술에서는 이러한 용어를 사용하도록 하겠다.

2) 일본인 학습자들을 대상으로 한 영어 쓰기 평가의 채점 결과를 분석한 Schaefer(2008)에서는 총 40명의 채점자 중 24명의 채점자가 하나 이상의 평가 범주에 대한 편향성을 가지고 있었다. 마찬가지로 한국어 쓰기 평가의 채점 결과를 분석한 강민석(2015)에서도 연구에 참여한 20명의 채점자 중 12명에게서 총 27회의 평가 범주에 대한 편향성이 관찰되기도 하는 등, 분석적 채점에서 특정 평가 범주에 대한 채점자들의 편향성은 빈번하게 관찰되는 채점 특성이다.

이러한 편향성에 대하여 McNamara(1990:69)는 수험자들이 어느 정도는 자신의 능력과 관계없이 채점자가 중요하다고 생각하는 특정 범주에 의해 평가되고 있다고 언급하였는데, 이는 평가 범주에 대한 채점자의 편향성이 평가 결과에 미치는 영향을 잘 보여주는 진술이다. 이처럼 채점자의 편향성은 특정한 수험자들의 쓰기 능력, 또는 그 일부분의 추정 결과를 왜곡할 수 있으며, 평가의 공정성과 관련한 문제를 일으킬 수도 있기 때문에 제거되어야 할 필요가 있는 특성이다.

사실 이러한 채점 특성은 채점자 간 신뢰도(inter-rater reliability), 채점자 내 신뢰도(intra-rater reliability)와 맞닿아 있는 개념이다. 채점자들의 엄격성 차이는 채점자 간 신뢰도로, 일관성의 문제는 채점자 내 신뢰도의 문제로 이어진다. 그러나 평가 점수 간의 상관계수를 통해 신뢰도를 분석하는 고전적 통계방법은 평가 점수가 일치하는 정도를 보여주기는 하지만 채점 결과에서 차이가 나타났을 때 그러한 차이가 나타나는 원인을 살피기는 어려웠다(이인혜, 2014:235). 반면 다국면 라쉬 모형을 통해 분석할 수 있는 개별 채점자의 엄격성, 일관성, 편향성 등은 채점의 신뢰도를 나타내 줄 뿐만 아니라, 어떠한 측면에서 채점의 문제가 발생하였는지를 구체적으로 진단할 수 있다는 장점을 지닌다. 본고에서도 채점자들의 이러한 엄격성, 일관성, 편향성에 초점을 두고 채점에 대한 피드백 및 채점 결과의 분석을 진행하였다.

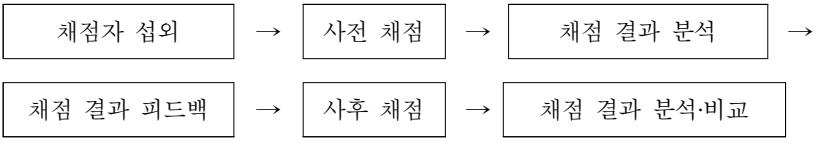
3. 연구 방법

3.1. 연구 문제 및 절차

연구의 목적에 근거하여 본 연구에서 초점을 두고 살펴보고자 하는 연구 문제는 다음과 같다.

첫째, 채점 특성에 대한 피드백이 채점자들 간 엄격성 차이의 정도를 감소시킬 수 있는가?
둘째, 채점 특성에 대한 피드백이 채점자 개인의 일관성을 향상시킬 수 있는가?
셋째, 채점 특성에 대한 피드백이 채점자 개인의 평가 범주에 대한 편향성을 제거할 수 있는가?

위와 같은 연구 문제의 해결을 위해, 본 연구에서는 22명의 채점자들을 대상으로 사전 채점을 실시한 후에 채점자 개인의 채점 특성에 대한 피드백을 제공하였다. 그리고 이러한 피드백이 채점자들의 채점에 긍정적인 영향을 줄 수 있는지를 알아보기 위해 사후 채점을 실시하였고, 그 결과를 사전 채점과 비교하였다. 본 연구의 절차를 요약하면 아래 <그림 1>과 같다.



<그림 1> 연구 절차

3.2. 연구 대상 및 도구

3.2.1. 채점자와 채점 기준

본 연구에 참여한 채점자는 8개의 한국어 교육 기관에 재직 중인 교사 22명이며, 이들의 한국어 교육 경력은 1년 이하부터 10년 이상까지 다양하게 분포하고 있다.[3)]

<표 1> 채점자 관련 정보

채점자	경력	채점자	경력	채점자	경력
R01	3개월	R08	1년 1개월	R15	2년 7개월
R02	3개월	R09	1년 1개월	R16	2년 9개월
R03	4개월	R10	1년 10개월	R17	3년
R04	7개월	R11	2년	R18	3년 4개월
R05	7개월	R12	2년	R19	4년
R06	7개월	R13	2년	R20	4년 7개월
R07	1년	R14	2년 5개월	R21	10년
				R22	13년 6개월

채점에 사용된 글은 국내 대학 부설 한국어 교육기관의 4급 학습자들이 쓴 글 40편이며, 40편의 글은 무작위로 양분되어 채점 특성 피드백이 제공되기 전과 후에 각각 20편씩 채점자들에게 제공되었다. 채점은 분석적 채점으로 이루어졌으며 이를 위한 채점 기준으로는 한국어능력시험(TOPIKⅡ) 작문형 문항의 채점 기준을 사용하였다. 구체적인 쓰기 과제와 평가 범주별 배점 및 채점 근거는 아래의 <표 2>와 같다.[4]

3) 채점자들은 'rater'의 머리글자를 빌려 'R01, R02…'와 같이 기호화하였다. 이후의 논의에서도 '첫 번째 채점자의 경우에는…'과 같은 진술을 대신하여 'R01의 경우에는…'과 같이 기술하도록 하겠다.

4) 한국어능력시험(TOPIKⅡ) 작문형 문항의 채점 기준은 53번 문항의 채점기준과 54번 문항의 채점 기준이 있는데, 두 채점 기준은 채점 범주와 근거, 범주별 점수 비율 등은 동일하고 총점에서만 차이가 난다. 전자의 경우 각 범주별 배점이 7점, 7점, 16점(8점×2)이며, 후자는 12점, 12점, 26점(13점×2)이다. 이들 중 본고에서는 53번 문항의 채점 기준을 사용하였는데, 그 이유는 따로 교육 단계를 거치지 않은 본 연구의 채점자들이 12점이나 13점과 같이 범위가 큰 척도를 가지고 수험자들의 글을 적절히 구분하여 채점하는 것이 불가능하기 때문이다. Hamp-Lyons & Henning(1991:364)에서도 6점 척도를 쓰기 평가의 일반적인 채점 척도라고 소개하면서, 더 큰 범위의 척도는 채점자들에게 상당한 인지적 부담을 줄 뿐이며 채점자들이 제대로 단계를 구분하지 못한다고 하였다.

<표 2> 쓰기 과제 및 채점 기준

※ 다음을 주제로 하여 자신의 생각을 400~600자로 쓰십시오.

최근 한국 사회가 고령화 사회로 진입하면서 가족과 떨어져 혼자 사는 노인이 점점 증가하는 것으로 밝혀졌습니다. 이에 대해 아래의 내용을 중심으로 자신의 생각을 쓰십시오.

· 노인 가구의 증가 원인은 무엇입니까?
· 노인 가구 증가로 인해 발생되는 문제와 이를 해결할 수 있는 방법은 무엇입니까?

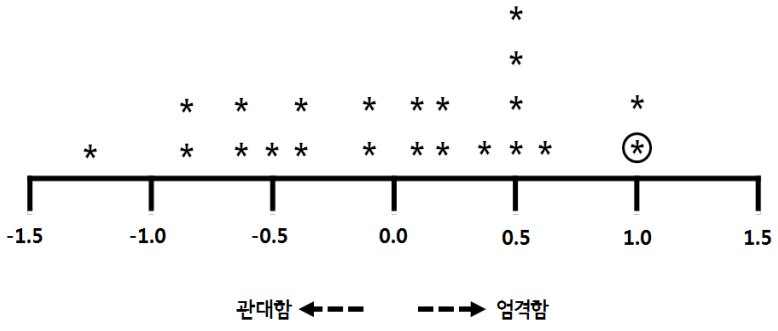
평가 범주	채점 근거	점수 구분		
		상	중	하
내용 및 과제 수행 (7점)	1) 주어진 과제를 충실히 수행하였는가? 2) 주제와 관련된 내용으로 구성하였는가? 3) 주어진 내용을 풍부하고 다양하게 표현하였는가?	7~6	5~3	2~0
글의 전개 구조 (7점)	1) 글의 구성이 명확하고 논리적인가? 2) 글의 내용에 따라 단락 구성이 잘 이루어졌는가? 3) 논리 전개에 도움이 되는 담화 표지를 적절하게 사용하여 조직적으로 연결하였는가?	7~6	5~3	2~0
언어 사용 (8x2=16점)	1) 문법과 어휘를 다양하고 풍부하게 사용하며 적절한 문법과 어휘를 선택하여 사용하였는가? 2) 문법, 어휘, 맞춤법 등의 사용이 정확한가? 3) 글의 목적과 기능에 따라 격식에 맞게 글을 썼는가?	8~7 (x2)	6~4 (x2)	3~0 (x2)

3.2.2. 채점 특성 피드백

20편의 글에 대한 22명의 채점자들의 사전 채점 결과를 바탕으로 채점 특성에 대한 피드백이 이루어졌다. 사전 채점 결과는 다국면 라쉬 모형에 기초한 프로그램인 FACETS 3.71.4를 통해 분석되었으며,[5] 분석된 채점 특성은 채점 결과 안내지를 통해 채점자들에게 개별적으로 제공되

었다.

채점 결과 안내지에 포함된 내용은 다음과 같다.[6] 먼저, 채점자가 본인의 엄격성 정도를 쉽게 알 수 있도록 22명의 채점자들의 엄격성 등급을 그래프로 시각화하여 제공하였다. 아래의 <그림 2>와 같이 0을 기준으로 왼쪽으로 멀어질수록 상대적으로 관대한 편임을, 오른쪽으로 멀어질수록 엄격한 편임을 표시하였다.



<그림 2> 엄격성 분석 결과 그래프 예시

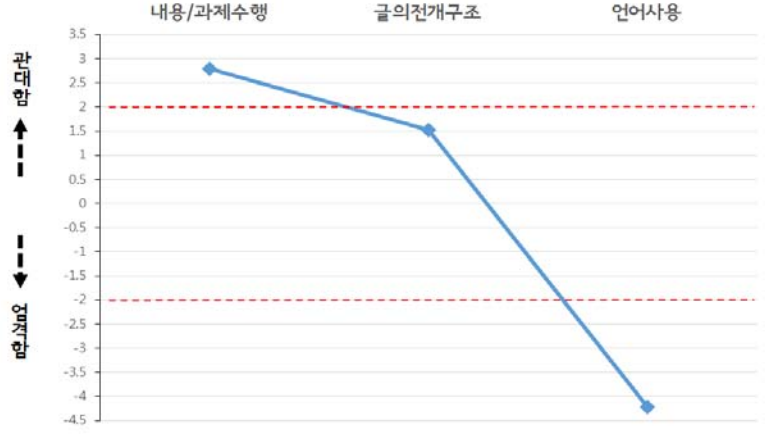
다음으로는 채점자의 일관성 정도를 점수로 제시하였다. FACETS는 채점자 개인이 일관된 기준을 가지고 채점하였는지를 적합도 지수를 통해 보고해 주는데, 적합도 지수가 1.00에 가까울수록 높은 일관성을 가진 것으로 볼 수 있으며 0.6~1.4의 범위를 벗어나는 경우에는 일관성에

5) FACETS는 평가의 결과에 영향을 미치는 여러 국면들의 데이터를 종합적으로 고려하여 문항특성곡선(item characteristic curve)을 확률적으로 추정해낸다. 본 연구에서 사용된 국면은 '학습자(글), 채점자, 평가 범주'의 3개 국면이다.
FACETS는 특히 채점자와 관련해서 채점자의 엄격성, 일관성에 관한 정보나 특정 채점자 혹은 채점자 집단이 평가 과제나 평가 범주 등에 대해 보일 수 있는 편향성의 정보를 제공해 주기 때문에, 계량화가 쉽지 않은 수행 기반의 언어 평가 연구에서 많이 사용되고 있다(장소영·신동일, 2009).

6) 실제로 사용된 채점 결과 안내지는 <부록 1>에 첨부하였다.

문제가 있는 것으로 판단된다.[7] 채점자들에게는 본인의 일관성 지수와 그에 대한 해석이 제공되었으며, 일관성에 문제를 보이는 채점자의 경우에는 분석적 채점에서 일관성 향상을 위해 사용할 수 있는 채점 전략을 함께 제공하여 사후 채점에서 참고할 수 있도록 하였다.[8]

마지막으로 채점자들의 범주별 채점 특성을 그래프로 제시하였고, 채점 결과에 유의미하게 영향을 줄 수 있는 편향성에 대해서는 채점자가 해당 범주에 대한 채점 기준을 조정해야 할 필요가 있음을 서술하였다. 아래 <그림 3>의 그래프는 채점자의 범주별 채점 편향성을 나타내고 있는데, 내용 및 과제 수행 범주에 대해서는 관대한 편향성이, 언어 사용 범주에 대해서는 엄격한 편향성이 존재함을 알 수 있다.



<그림 3> 편향성 분석 결과 그래프 예시

7) 적합도 지수가 1.4를 초과하는 경우는 일관성이 부족한 부적합 일관성(misfit), 0.6 미만인 경우는 일관성이 지나치게 높은 과적합 일관성(overfit)이다.

8) 채점 전략은 강민석(2015)의 연구 과정에서 채점자들에게 쓰기 채점에서 일관성 유지를 위해 사용하는 전략에 대하여 조사한 것들 중 일관성이 상대적으로 높은 채점자들이 주로 사용하고 있는 채점 전략을 추려 제시하였다.

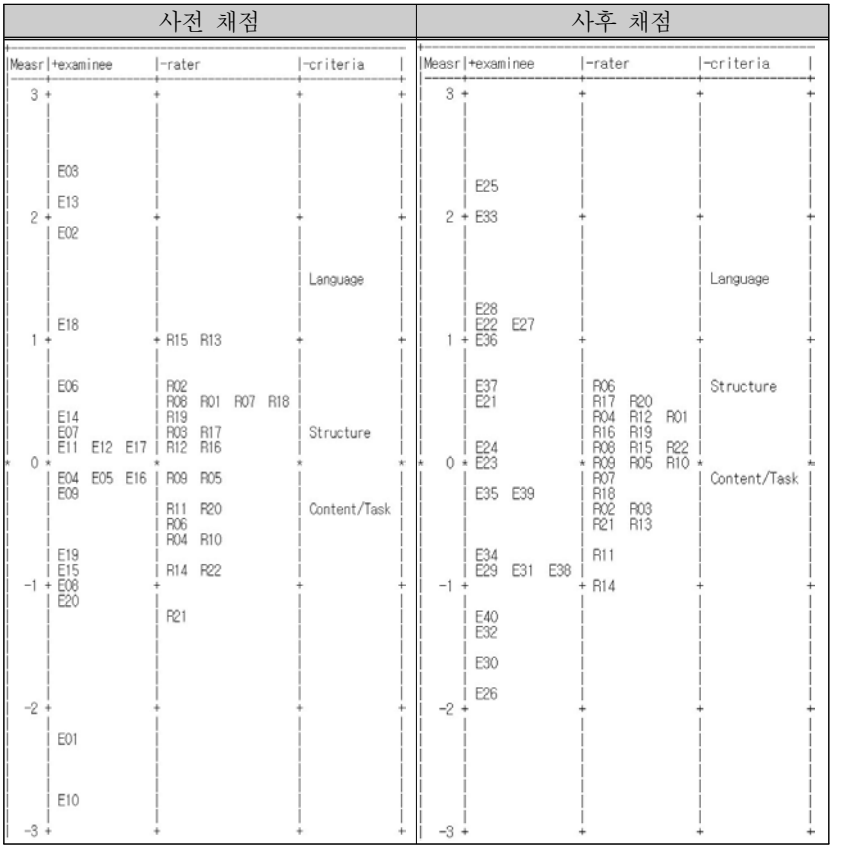
본 연구에서 채점 특성 피드백의 역할이 막중한 만큼, 전술한 내용들은 채점 결과 안내지에 상세하게 기술되어 채점자들에게 개별적으로 안내되었다. 또한 안내지의 앞부분에는 엄격성, 일관성, 편향성 등 채점 특성의 개념과 해석에 대한 설명을 덧붙여 채점자들의 이해를 돕고자 하였다. 채점자들은 사전 채점 결과에 대한 안내를 받은 후에 사후 채점을 하였는데, 사후 채점을 시작하기 전에 다시 한 번 개인적으로 본인의 사전 채점 결과를 정독하여 줄 것을 채점자들에게 요청하였다.

4. 연구 결과

4.1. 엄격성 분석 결과

먼저 첫 번째 연구문제에 기초하여 채점 특성에 대한 피드백을 받기 전과 받은 후의 채점자들의 엄격성이 변화가 있는지를 살펴보도록 하겠다. FACETS는 채점자들의 엄격성을 로짓(logit) 값으로 보고해준다.[9] 채점자들의 엄격성은 0로짓을 기준으로 0보다 클 경우 엄격하고 0보다 작을 경우 관대하다고 해석된다.

9) 로짓(logit)은 로그 승산(log odds unit)을 가리키는 말이며, 로짓은 등간척도이기 때문에 문항의 난이도나 능력 수준을 비교하여 나타내기에 적합하다(McNamara, 1996:165).



<그림 4> 사전-사후 채점의 전체 국면의 분포도

위의 <그림 4>는 사전 채점과 사후 채점의 전체 국면의 분포도를 비교한 것이다. 두 표의 가장 왼쪽 열의 'Measr'는 로짓 값을 단위로 하는 일종의 세로로 된 자(vertical ruler)로, 나머지 국면들을 하나의 기준에 놓고 비교하여 볼 수 있도록 해준다. 왼쪽에서 두 번째의 'examinee'열은 각각 20편의 글에 대한 정보를 나타내주고 있으며 위쪽에 위치할수록 점수가 높은 글이다. 세 번째 열은 채점자의 엄격성 정도, 네 번째 열은 각

평가 범주에 적용된 채점자 전체의 엄격성을 표시하며, 두 열 모두 로짓 값이 클수록, 즉 위쪽에 위치할수록 엄격함을 나타낸다.

본 연구에서 살펴보고자 하는 채점자의 엄격성은 세 번째 열에 제시되어 있다. 사전 채점에서보다 본인의 채점 특성에 대한 피드백을 받은 후의 채점자들의 엄격성이 0을 중심으로 더 밀집되어 있음을 시각적으로 확인할 수 있다. 이는 구체적인 수치로도 확인할 수 있는데, 사전 채점에서는 채점자들의 엄격성이 -1.26~1.02로짓의 범위 내에서 분포하였으나, 피드백을 받은 후의 채점에서는 그 범위가 -0.99~0.64로짓으로 좁혀진 것으로 나타났다. 이러한 변화는 채점자들의 엄격성 차이가 채점자 훈련을 거친 이후에도 쉽게 조정되지 않는 경향임을 고려할 때(Lumley & McNamara, 1995; Weigle, 1998), 매우 긍정적인 결과로 볼 수 있겠다.

아래의 <표 3>을 통해 개별 채점자들의 사전-사후의 엄격성의 변화에 대하여 조금 더 자세히 살펴보도록 하겠다.

<표 3> 사전-사후 채점의 엄격성

채점자	사전 엄격성	사후 엄격성	채점자	사전 엄격성	사후 엄격성
R01	0.52	0.36	R12	0.14	0.43
R02	0.66	-0.34	R13	1.02	-0.54
R03	0.23	-0.32	R14	-0.83	-0.99
R04	-0.60	0.36	R15	0.96	0.18
R05	-0.16	0.01	R16	0.12	0.30
R06	-0.55	0.64	R17	0.27	0.48
R07	0.45	-0.10	R18	0.54	-0.19
R08	0.56	0.18	R19	0.41	0.23
R09	-0.18	-0.01	R20	-0.43	0.52
R10	-0.64	-0.04	R21	-1.26	-0.51
R11	-0.37	-0.75	R22	-0.87	0.10

위의 표에서 음영 처리된 채점자들은 피드백을 받은 후 엄격성이 사전 채점에 비해 0로짓에 가깝게 조정된 채점자들이다. 22명의 채점자들 중 14명의 채점자가 본인의 엄격성 정도를 인지한 후에 엄격성을 조정한 모습을 볼 수 있다. 물론 나머지 8명의 채점자는 상대적으로 더 엄격해지거나 관대해진 경향을 보이고 있는데, 이는 그들의 문제라기보다는 14명의 채점자가 자신의 엄격성을 조정한 것에서 비롯된 결과로 해석할 수도 있을 것이다. 또한, 사전 채점에서 가장 관대했던 채점자 3명(R14, R21, R22)과 가장 엄격했던 3명(R02, R13, R15) 중에서 R14을 제외하고는 모든 채점자가 사후 채점에서 자신의 엄격성을 조정하였다는 것도 의미가 있는 결과라고 할 수 있다.

4.2. 일관성 분석 결과

여기에서는 두 번째 연구문제의 해결을 위해 채점자의 내적 일관성이 채점 특성에 대한 피드백을 받기 전과 후에 어떻게 나타났는지를 살펴보도록 하겠다. 다국면 라쉬 모형에 근거한 FACETS는 채점자들의 일관성을 적합도 지수를 통해 보고해 준다. 3장에서 언급하였듯이, 적합도 지수가 본 연구에서 설정한 0.6~1.4의 범위에 있으면 채점자의 일관성에 문제가 없는 것으로 볼 수 있으며, 0.6 미만인 경우 과적합, 1.4를 초과하는 경우에는 부적합으로 해석할 수 있다.[10)]

10) 본 연구에서 설정한 적합도 지수의 범위는 Wright & Linacre(1994)에서 제시된 범위를 따르고 있다.

<표 4> 사전-사후 채점의 일관성

채점자	사전 적합도 지수	사후 적합도 지수	채점자	사전 적합도 지수	사후 적합도 지수
R01	1.34	1.16	R12	0.92	1.02
R02	0.64	1.27	R13	1.66*	0.68
R03	0.92	1.00	R14	1.22	0.72
R04	1.15	0.98	R15	1.47*	1.32
R05	0.98	0.61	R16	0.65	0.71
R06	1.12	1.00	R17	0.92	0.63
R07	0.63	0.90	R18	0.75	0.93
R08	0.90	1.07	R19	0.65	0.64
R09	1.11	0.83	R20	2.17*	2.87*
R10	0.57*	0.67	R21	0.70	1.00
R11	0.73	0.80	R22	0.99	1.31

* = 과적합 또는 부적합

사전 채점에서 일관성에 문제를 보인 채점자는 22명 중 4명이었다. 일관성이 부적합 양상을 보이는 채점자가 3명(R13, R15, R20), 과적합 양상을 보이는 채점자가 1명(R10)이었다. 사후 채점에서는 사전 채점에서 일관성에 문제를 보였던 R10, R13, R15의 일관성이 정상 범위 안으로 교정된 것을 볼 수 있다. 또한, 사전 채점에서 일관성에 문제를 보이지 않던 채점자들 전원이 사후 채점에서도 적정한 일관성 수준을 유지하였다는 것도 매우 긍정적인 현상인 것으로 보인다. 영어교육 분야에서 채점자 훈련을 실시하고 그 효과를 연구한 Weigle(1998)이나 최연희(2002) 등의 연구에서 사전 채점에서는 문제가 없다가도 채점자 훈련의 영향을 받아 훈련 이후에 오히려 일관성에 문제를 보이는 채점자들이 더러 출현하였음과 비교해 볼 때, 본인의 현재 일관성 수준을 제공한 피드백이 일관성에 문제가 있는 채점자들에게는 개선의 노력을 기울이도록 영향을

주었으며 일관성에 문제가 없는 채점자들에게는 보다 높은 효능감을 가지고 현재의 채점을 이어나갈 수 있도록 해주었을 것으로 짐작해 볼 수 있겠다.[11)]

한편, R20의 경우 사후 채점에서 오히려 부적합 양상이 더 두드러진 결과가 나타났다. R20은 사전 채점에서도 가장 심각한 부적합 양상을 보인 채점자였는데, 일관성의 문제가 매우 심각한 수준일 경우에는 피드백을 통해 본인의 일관성 수준을 인지하였다고 하더라도 이에 대한 조정이 쉽지 않았던 것을 알 수 있다. 이러한 채점자의 경우에는 지속적인 훈련을 통한 교정이 필요해 보인다. 그러나 이 경우를 제외한다면, 본 연구에서 실시한 채점 특성 피드백은 채점자들의 일관성 향상 및 유지에 긍정적인 영향을 줄 수 있는 것으로 판단된다.

4.3. 편향성 분석 결과

마지막으로 세 번째 연구 문제와 관련하여, 채점자들의 사전-사후 채점에서 나타나는 평가 범주에 대한 편향성을 분석해 보기로 한다. FACETS는 <그림 3>과 같이 채점자의 범주별 채점 경향을 그래프로 나타내 주는데, 위의 일관성의 적합도 지수와 마찬가지로 통계적으로 유의미한 편향성을 가려낼 수 있는 범주가 존재한다. 본 연구의 경우에는 양측검정 하에서 α=0.05이고, df=19일 때의 기각값인 t=2.093이 기준이 되어, 편향성의 t값이 －2.093보다 작거나 2.093보다 큰 경우에 유의미하다고 판정된다.

이와 같은 기준에 근거하여 본 연구의 사전-사후 채점에서 평가 범주

11) 이와 관련하여 이인혜(2014)에서는 쓰기 평가 효능감과 채점자들의 일관성 이 양의 상관관계를 가지고 있음을 밝힌 바 있다. 채점 특성 피드백을 통해 본인이 일관성 있는 채점을 하고 있다는 사실을 확인하는 것은 쓰기 채점자들의 채점 효능감에 긍정적인 영향을 줄 수 있을 것으로 생각된다.

에 대하여 관찰된 채점자들의 편향성은 아래와 같다.

<표 5> 사전-사후 채점의 편향성

채점자	사전 채점			
	범주	편향성 크기	t값	p값
R03	언어 사용	-0.68	-2.75	0.0128
R08	언어 사용	0.56	2.19	0.0412
R14	언어 사용	-0.64	-2.53	0.0206
R15	내용 및 과제 수행	0.68	2.79	0.0117
R15	언어 사용	-0.92	-4.21	0.0005
R17	내용 및 과제 수행	0.77	3.07	0.0063
R20	글의 전개 구조	0.83	3.31	0.0037
R20	내용 및 과제 수행	-0.59	-2.44	0.0244
R22	내용 및 과제 수행	-0.61	-2.49	0.0249
채점자	사후 채점			
	범주	편향성 크기	t값	p값
R14	내용 및 과제 수행	-0.56	-2.23	0.0380
R22	언어 사용	-0.57	-2.12	0.0474

위의 표에서 편향성 크기나 t값이 양수인 경우는 채점자가 전체 채점 결과에 영향을 줄 만큼 특정 범주에만 관대한 채점을 한 경우를 말하며, 편향성 크기나 t값이 음수인 경우는 그 반대의 경우를 뜻한다. 예를 들어 표의 가장 위쪽에 위치한 R08의 경우에는 언어 사용 범주에서 0.56만큼의 편향성을 가지고 있는 것을 볼 수 있는데, 따라서 R08은 언어 사용 범주에만 상대적으로 매우 관대한 채점을 하고 있다고 해석할 수 있다.

사전 채점을 분석한 결과 7명의 채점자에게서 총 9회의 유의미한 편향성이 관찰되었는데, 피드백을 받은 후의 채점에서는 사전 채점에서 발견되었던 편향성이 모두 제거된 것으로 나타났다. 이는 매우 고무적인 결과라고 할 수 있는데, 채점자들이 본인이 특정 범주에 심각하게 엄격

하거나 관대하다는 것을 인지하고 채점하는 것이 채점의 편향성 감소에 매우 긍정적인 영향을 줄 수 있다는 것을 시사한다.

한편, 사후 채점에서 편향성을 보인 R14과 R22의 경우를 살펴보면, 이 두 명의 채점자는 사후 채점에서 기존에 가지고 있었던 편향성은 사라졌지만 사전 채점에서 관대하게 채점하던 범주에서 오히려 엄격한 편향성을 나타내었는데, 이는 본인의 채점 경향을 지나치게 의식하여 채점한 것에서 기인한 결과로 보인다. 이들은 '내용 및 과제 수행'과 '언어 사용' 범주 사이에서 이러한 모습을 나타내고 있는데, 이 두 범주는 채점자들이 흔히 상반된 채점 경향을 가지고 있는 범주들이다(Schaefer, 2008; 강민석, 2015). 예를 들어, 언어 사용 범주에 매우 엄격한 채점자들은 상대적으로 내용 및 과제 수행 범주에는 아주 관대할 가능성이 높다는 것이다. 따라서 R14과 R22의 경우에는 특정 범주의 편향성을 상당히 의식하여 조정하다 보니 상대되는 범주에서 반대의 경향이 짙어진 것으로 해석할 수 있겠다.

5. 결론

본 연구는 한국어 쓰기 평가에서 채점자 개인의 채점 특성에 대한 피드백이 채점에 긍정적인 영향을 미칠 수 있는지를 검증하고자 하였다. 피드백에는 개별 채점자의 엄격성, 일관성, 편향성에 대한 자세한 정보가 시각 자료와 함께 제공되어, 채점자들이 본인의 채점 특성을 면밀하게 인지할 수 있도록 하였다. 피드백은 채점 결과 안내지를 통해 채점자들에게 개별적으로 이루어졌으며, 그 효과를 알아보기 위해 22명의 채점자들이 피드백 이전과 이후에 각각 수행한 채점의 결과를 다국면 라쉬 모형을 활용하여 분석·비교하였다.

본 연구의 결과를 요약하면 다음과 같다. 먼저, 채점 특성에 대한 피드백은 채점자들의 엄격성 차이의 정도를 감소시키는 효과가 있는 것으로

나타났다. 또한, 피드백을 받은 후 일관성에 문제가 있던 채점자 4명 중 3명이 사후 채점에서 일관성을 확보한 모습이 나타났으며 기존에 일관된 채점을 하던 나머지 채점자들 역시 일관성을 유지하는 모습을 보여, 채점 특성에 대한 피드백이 일관성 향상에도 긍정적인 영향을 주는 것을 알 수 있었다. 마지막으로 채점 특성 피드백은 편향성의 제거에도 상당히 긍정적인 효과가 있었는데, 사전 채점에서 나타났던 평가 범주에 대한 채점자들의 편향성이 사후 채점에서는 모두 제거된 모습을 보였다. 종합하여 보면, 채점 특성에 대한 피드백을 통해 채점자가 스스로 자신의 채점 경향을 인지하는 것은 채점에 매우 긍정적인 효과가 있는 것으로 판단된다.

물론 본 연구의 표본이 크지 않기 때문에 연구의 결과를 일반화하는 데에는 한계가 있으나, 한국어 쓰기 평가 분야에서 구체적 채점 개선 방안의 효과에 대한 실증적인 연구가 전무한 현 시점에서 본 연구는 실현 가능한 하나의 방안을 제시하고 그 효과를 검증하였다는 의의를 갖는다고 하겠다.

특히 본 연구에서 제안한 채점 특성 피드백은 개별 채점자들의 채점에 대한 상세하고 객관적인 분석 자료를 제공할 수 있을 뿐만 아니라 어떤 환경에서든 동일한 쓰기 결과물에 대한 채점 결과가 수집되면 분석과 피드백이 가능하기 때문에 집체 교육 형태의 채점자 훈련의 한계들을 보완하는 방법이 될 수 있다고 하겠다.

본고는 기초연구로서의 성격을 가지기 때문에 채점 특성 피드백이 채점자들에게 미치는 전반적인 영향에만 주목하였고, 채점자들의 배경 변인이나 피드백의 횟수 등에 따라 나타나는 세부적인 효과를 설명하는 데까지는 미치지 못하였다. 추후에는 보다 다양한 변인들을 고려한 연구를 진행하여 논의의 깊이를 더하여 가고자 한다.

<참고 문헌>

강민석(2015). 다국면 라쉬 모형을 통한 한국어 쓰기 평가에서의 채점 경향 연구, 고려대학교 석사학위논문.

김동수(2014). 사고 구술법을 통한 한국어 쓰기 평가자의 숙련도별 평가 과정 연구, 고려대학교 석사학위논문.

김정숙(2010). 한국어 쓰기 능력 평가 방안: 종합적 채점과 분석적 채점 결과를 중심으로, <이중언어학> 43호, 이중언어학회. 81쪽~99쪽.

박영민(2012). 예비국어교사의 중학생 논설문 평가에서 발견되는 엄격성 및 일관성의 특성, <국어교육학연구> 43호, 국어교육학회. 253쪽~283쪽.

이인혜(2012). 한국어 쓰기 평가의 채점 방식에 따른 채점자 신뢰도 연구: 종합적 채점 및 분석적 채점을 중심으로, 고려대학교 석사학위논문.

이인혜(2014). 한국어 교사의 쓰기 평가 효능감과 평가 특성 연구, <이중언어학> 56호, 이중언어학회. 231쪽~266쪽.

이정화(2010). 쓰기 채점의 신뢰도 확보 방안: 고급 한국어 학습자 주관식 쓰기 문항 채점 실험을 바탕으로, <어문연구> 38(2)호, 한국어문교육연구회. 597쪽~522쪽.

장소영·신동일(2009). 언어교육평가 연구를 위한 FACETS 프로그램. 서울:글로벌 콘텐츠.

최연희(2002). 채점자 훈련이 영어 작문 채점에 미치는 효과 연구: Facets 분석을 통한 신뢰도 변화 분석, <응용 언어학> 18(1)호, 한국응용언어학회. 257쪽~292쪽.

최윤곤·허유리(2013). 한국어 교사의 채점자 훈련 모형 개발, <한국언어문화학> 10(2)호, 국제한국언어문화학회. 295쪽~316쪽.

최은지(2013). 한국어 교사의 작문 평가 과정에 대한 사고 구술 연구, <우리어문연구> 47호, 우리어문학회. 273쪽~300쪽.

허유리(2014). 한국어 쓰기 평가에서 채점자의 신뢰도 향상을 위한 방안 연구, 동국대학교 석사학위논문.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of educational measurement(5th ed.)*. Prentice Hall.

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

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement: analyzing and evaluating rater-mediated assessments*. NewYork: Peter Lang.

Elder, C., Knoch, U., Barkhuizen, G. & Randow, J. V. (2005). Individual Feedback to Enhance Rater Training: Does It Work?, Language Assessment Quarterly,

2(3), pp.175~196.

Hamp-Lyons, L. & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts, Language Learning, 41(3), pp.337~373.

Linacre, J. M. (1994). *Many-Facet Rasch Measurement(2nd Ed)*. Chicago: Mesa Press.

Linacre, J. M. (2014). *Facets computer program for many-facet Rasch measurement*, version 3.71.4. Beaverton, Oregon: Winsteps.com.

Lumley, T. & McNamara, T. (1995). Rater characteristics and rater bias: implications for training, Language Testing, 12(1), pp.54~71.

McNamara, T. (1990) Item Response Theory and the validation of an ESP test for health professionals, Language Testing, 7(1), pp.52~76.

McNamara, T. (1996). *Measuring second language performance*, UK: Longman.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment, Language Testing, 25(4), pp.465~493.

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, pp.27~33.

Tyndall, B. & Kenyon, D. (1996). *Validation of a new holistic rating scale using Rasch multifaceted analysis*. Cumming, A. & Berwick, R. (Eds.). Validation in language testing. Multilingual Matters.

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

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values, Rasch Measurement Transactions, 8(3), pp.369~386.

〈부록 1〉 채점 특성 피드백 예시

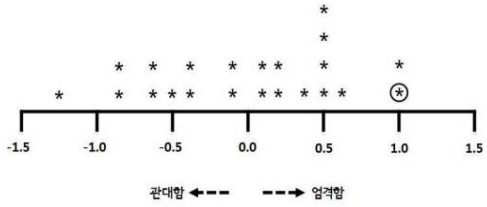
채점 결과 분석 안내

◆ 채점 결과는 다국면 라쉬 모형에 근거한 FACETS 프로그램을 활용하여 분석되었습니다. 분석 결과는 뒷장에 그래프와 함께 상세히 제시되어 있습니다.

◆ 결과 분석에서 제공될 내용은 채점의 엄격성, 일관성, 편향성 정보입니다. 아래는 엄격성, 일관성, 편향성의 개념과 해석에 대한 안내입니다.

엄격성 severity	엄격성(severity)은 채점자가 가지고 있는 고유의 평가 기준이 다른 채점자들과 비교하였을 때 어느 정도 엄격한지, 혹은 관대한지를 의미합니다. 쓰기 평가에서 동일한 학습자의 수행이 서로 다른 점수로 측정되는 가장 큰 원인은 채점자들이 가지고 있는 엄격성의 차이 때문입니다. 따라서 채점자의 엄격성은 채점자 간 신뢰도와 관련이 깊습니다. 상대적으로 높은 엄격성을 가지고 있는 채점자는 지속적으로 낮은 점수를 부여할 것이고, 낮은 엄격성을 가진 채점자는 다른 채점자와 동일한 글을 채점하면서도 높은 점수를 줄 것입니다.
일관성 consistency	일관성(consistency)은 채점의 전 과정에서 채점자가 엄격성을 일관되게 유지하는 정도를 의미합니다. 다시 말해, 채점자가 얼마나 일관된 기준을 가지고 채점했는지를 나타낸다고 할 수 있습니다. 따라서, 일관성은 채점자 내 신뢰도와 일치하는 개념입니다. 일관성은 채점에 있어서 가장 중요한 개념이며, 채점자가 상대적으로 높거나 낮은 엄격성을 가지고 있더라도 일관성을 잘 유지한다면 큰 문제가 되지 않습니다. 그러나 일관성에 문제를 보인다면 채점자가 학습자들의 쓰기 능력을 올바르게 추정하고 있다고 할 수 없습니다. 일관성의 문제는 두 가지 양상으로 나타날 수 있습니다. • 일관성이 부족한 경우: 일관성이 부족한 채점자는 채점 경향 자체를 예측하는 것이 불가능한 채점자를 의미한다. 즉, 채점을 하는 글마다 다른 엄격성을 적용하는 등 채점자 내 신뢰도가 낮은 경우입니다. (2쪽에서 일관성 지수가 1.4를 초과하는 경우입니다.) • 일관성이 지나치게 높은 경우: 지나치게 예측 가능한 채점을 하는 것을 의미합니다. 이러한 채점자들은 채점자 내 신뢰도에는 문제가 없지만, 주로 대부분의 쓰기 결과물에 대해서 엇비슷한 점수를 주거나, 인상 평가와 같이 잘 쓴 글에는 비슷하게 높은 점수를 주고 그렇지 못한 글에는 비슷하게 낮은 점수를 주는 경향이 있습니다. 일관성이 지나치게 높은 채점자의 가장 큰 문제는 비슷한 점수대에 대한 변별력이 떨어진다는 것이며, 또한 이들은 한두 가지 특징에 이끌려 전체 점수를 주게 되는 등과 같이 분석적 채점에 적합하지 않는 모습을 나타낸다. (2쪽에서 일관성 지수가 0.6미만인 경우입니다.)
편향성 bias	편향성(bias)이란 채점자가 특정한 과제나 특정 학습자 집단, 또는 특정 채점 범주에 대해서만 지속적으로 엄격하거나 관대한 채점을 하는 경향을 말합니다. 분석적 채점 방식의 쓰기 평가에서 채점자들의 편향성은 특정한 평가 범주에 대하여 나타나는 경우가 많습니다. 채점자들은 어느 한 평가 범주에만 상대적으로 유독 엄격하거나 관대하게 채점하는 경향을 보일 수 있습니다. 이러한 경향은 흔하게 나타나는 것으로서, 보통 채점자가 스스로 중요하다고 인식하는 범주에 상대적으로 높은 엄격성을 적용하는 것으로 나타납니다. 그러나 채점 결과에 영향을 줄만큼 유의미한 편향성을 보이는 경우에는 특정한 학습자들의 쓰기 능력, 또는 학습자들의 쓰기 능력의 일부분의 측정 결과를 왜곡하는 문제가 발생할 수 있습니다.

◆ **채점 엄격성 분석 결과**



위의 그래프는 선생님의 채점 **엄격성** 정도를 나타내는 그래프입니다. 보시는 바와 같이 0을 기준으로 왼쪽으로 멀어질수록 상대적으로 관대한 채점을, 오른쪽으로 멀어질수록 엄격한 채점을 하고 있음을 나타냅니다. 그래프의 ⊛ 표시는 선생님의 엄격성 정도입니다.

선생님의 엄격성 수치는 0.96이며, 다른 채점자들과 비교하였을 때 **매우 엄격한 편**입니다(가장 엄격한 채점자를 기준으로 22명 중 2번째에 해당합니다).

채점자들 중 중앙에 위치한 채점자는 한 개의 작문에 평균적으로 13.7점을 주고 있습니다. 선생님께서는 한 개의 작문에 평균적으로 10.7점을 주어, 3.1점 정도의 차이를 보이고 있습니다.

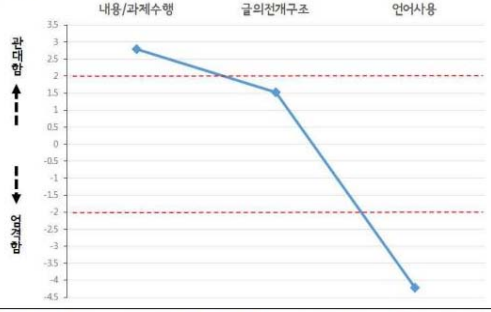
◆ **채점 일관성 분석 결과**

선생님의 채점 **일관성**은 1.47점으로 일관성이 부족한 편입니다. 일관성의 유지를 위한 노력이 필요해 보입니다. 아래에 일관성 유지를 위해 자주 사용되는 채점 전략이 제시되어 있습니다.

(0.6~1.4가 허용 범주이며, 1.00에 근접할수록 일관성이 높다고 해석됩니다.)

- 채점을 시작하기 전에 전체 작문을 종합적으로 간단히 살펴본 후 채점
- 범주마다 채점 세부 기준을 설정한 후 이를 환기하면서 채점
- 일정량의 샘플을 먼저 가채점한 후, 상중하의 기준을 마련하여 채점

◆ **채점 편향성 분석 결과**

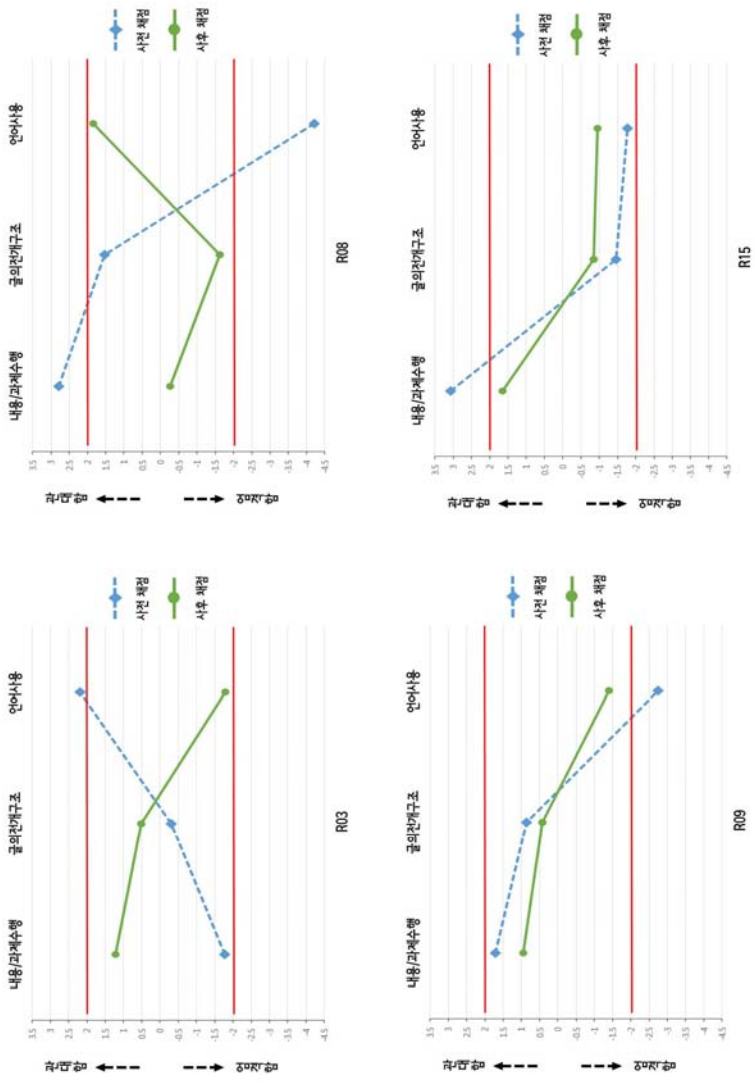


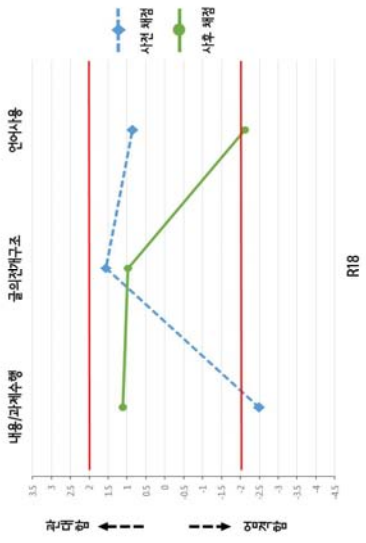
위의 그래프는 각 채점 범주에 대한 선생님의 채점 **편향성** 정도를 나타내는 그래프입니다. 보시는 바와 같이 0을 기준으로 점이 위쪽에 위치하면 관대함을, 아래쪽에 위치하면 엄격함을 나타냅니다.

2점과 -2점을 기준으로 그어진 점선을 넘어서는 편향성은 채점 결과에 영향을 줄 만큼 특정 범주를 엄격하거나 관대하게 채점하고 있다는 것을 의미합니다.

선생님께서는 내용/과제수행 범주를 다른 범주에 비해 관대하게 채점하고 있습니다. 언어사용 범주는 상대적으로 매우 엄격하게 채점하고 있습니다. 이러한 편향성은 채점 결과에 유의미한 영향을 미칠 수 있는 정도이므로, 해당 범주에 대한 기준을 조정해야 할 필요가 있어 보입니다.

〈부록 2〉 사전 채점에서 편향성을 보인 채점자의 사전-사후 비교




R18
내용/과제수행
글의전개구조
언어사용
사전 채점
사후 채점
R16
내용/과제수행
글의전개구조
언어사용
사전 채점
사후 채점
R21
내용/과제수행
글의전개구조
언어사용
사전 채점
사후 채점

강민석(Kang Min-suk)
고려대학교 국어국문학과
서울시 성북구 안암로145 고려대학교(136-701)
전화번호: 010-3067-6650
전자우편: hatnim0423@korea.ac.kr

심재경(Sim Je-kyoung)
고려대학교 국어국문학과
서울시 성북구 안암로145 고려대학교(136-701)
전화번호: 010-4872-5345
전자우편: jaee893@gmail.com

접수일자: 2015년 10월 20일
심사(수정)일자: 2015년 12월 19일
게재확정: 2015년 12월 23일