『이중언어학』 제56호(2014.9.30). 二重言語學會

한국어 교사의 쓰기 평가 효능감과 평가 특성 연구*

이 인 혜

Abstract

Lee Inhye. 2014. 9. 30. **Assessment Self-Efficacy and Evaluation Characteristics of Korean Language Teachers in Writing Assessment.** *Bilingual Research 56*, 231-266. In this study, we analyzed writing assessment self-efficacy of Korean language teachers and their evaluation characteristics. Our purpose was to collect information about raters and to gain a basic understanding of Korean writing assessment and training for raters. To do this, we split 39 Korean language teachers into three groups based on experience and asked them to complete a Korean writing assessment self-efficacy examination and to actually evaluate a composition. The results were then analyzed using FACETS, a statistics program which uses the "Many-facet Rasch Model" There was a noticeable difference between the group of teachers with less than two years of experience (A) and the two groups of teachers with more than two years of experience (B, C). There was a larger difference in "general writing assessment self-efficacy" than "practical writing assessment self-efficacy." The three groups showed differences as to their strictness in using evaluation criteria. This creates a connection between writing assessment ability and writing assessment self-efficacy. The three groups also showed differences in suitable consistency. Teachers with more experience showed more suitable consistency. The group of teachers with less than two years of experience (A) showed over-fitted and mis-fitted consistency, causing a statistical difference in "general writing assessment self-efficacy." Because writing assessment self-efficacy and assessment ability affect one another, there needs to be more training for Korean language teachers based on this type of writing

* 이 연구는 제16회 한국어문학 국제학술회의(2014년 7월 28일)에서 발표한 내용을 수정 및 보완한 것임.

assessment self-efficacy and evaluation characteristics.**(Korea University)**

【Key words】 teacher efficacy(교사 효능감), writing assessment self-efficacy (쓰기 평가 효능감), writing assessment(쓰기 평가), rater(평가자), Many-facet Rasch model(다국면 라쉬 모형)

1. 서론

본 연구는 한국어 교사의 쓰기 평가 효능감 및 쓰기 평가 특성을 분석하여 평가자 관련 정보를 구축하고 한국어 쓰기 평가 연구 및 평가자 교육을 위한 기초 자료를 마련하는 데 목적이 있다.

한국어 교육 현장에서 직접 평가로서의 쓰기 평가가 지속적으로 이루어지고 있고, 한국어 능력시험(TOPIK)의 쓰기 영역에서도 직접적 평가의 비중을 늘리는 등, 쓰기 평가의 중요성이 확대되어감에 따라, 평가자의 역할이 매우 중요해지고 있다.

그런데 쓰기 평가의 경우, 평가자에 따라 평가 기준이 다르고 평가자의 주관이 개입될 여지가 많아 평가의 객관성과 신뢰성을 확보하기 어렵다(김정숙, 2010:82). 가치판단은 필연적으로 복잡해질 수밖에 없으며 평가의 해석이 개입되어 판단의 불일치가 발생하는 것이다(McNamara, 1996:167). 평가자의 해석에는 임의성이 존재하며, 그 임의성은 '제멋대로의 변덕스러운 것'이라는 의미보다는 '심사숙고한 재량의 뜻'이라는 의미를 지닌다(Block, 1978; Popham, 1978; 서수현, 2008). 따라서 이러한 임의성을 고려하되 최대한 객관성을 유지하는 방향을 찾기 위해서는 평가자의 특성 중 정의적인 요소에 대해 살필 필요가 있다.

평가 과정에서의 평가자 개인이 갖는 사고과정에는 경력과 같은 배경요인과 함께 쓰기 평가에 대한 효능감, 신념, 인식 등이 개입된다. 비슷한 경력을 가진 평가자라 하더라도 이러한 정의적인 요소가 평가자의 엄격성과 일관성에 영향을 미쳐 다른 평가 양상을 가져올 수 있는 것이다.

이중에서도 목표한 바를 산출하기 위해 필요한 행동과정을 조직화하고 실행할 수 있는 자기 능력에 대한 신념(Bandura, 1997)으로 정의되는 효능감(self-efficacy)[1]은 한국어 교사의 쓰기 평가 특성과 영향을 주고받을 수 있는 중요한 부분이라고 할 수 있다.

사회 학습 이론에서 시작된 효능감은 주로 학생과 관련된 국면에서 연구되어 왔으나, 최근에는 교사 영역으로까지 확대되는 추세에 있다. 교사 효능감은 교사가 인지하고 있는 자기 자신에 대한 평가를 바탕으로 하고 있으므로, 자신에 대한 평가가 긍정적일수록 효능감 수준이 높게 나타난다. 또한 교사 효능감은 실제로 교사가 교육적 의사 결정을 내리고 교육적 행동을 조직하는 데 영향을 미친다. 이러한 교사의 직무 중에서도 쓰기 평가는 매우 숙련된 전문적인 능력을 필요로 하는 영역이라는 점에서 쓰기 평가 활동을 긍정적으로 이끄는 효능감의 역할이 중요하다(박영민, 2010: 183-185).

한국어교육에서의 쓰기 평가는 교사가 학습자의 L2 숙달도를 고려하여야 한다는 점에서 L1 쓰기 평가와는 다른 쓰기 평가 효능감 및 평가 특성을 나타낼 수 있다. 교사의 쓰기 평가는 학습자의 쓰기 숙달도를 가늠하는 잣대이자 환류효과를 제공하는 교육적 자료가 되는 만큼 쓰기 평가자에 대한 연구가 중요하다.

이러한 필요성에 따라 본 연구에서 알아보고자 하는 연구 문제는 다음과 같다. 첫째, 한국어 교사의 쓰기 평가 효능감은 어떠한 양상을 보이는가? 둘째, 한국어 교사의 쓰기 평가 특성은 어떠한가? 셋째, 한국어 교사의 쓰기 평가 효능감과 평가 특성은 관련이 있는가?

이러한 연구 문제를 논의하기 위하여 한국어 교사 39명을 경력에 따라 세 집단으로 나누고, 한국어 쓰기 평가 효능감 검사를 실시하는 한편, 10

1) ‘self-efficacy’는 ‘자기효능감’, ‘효능감’ 등으로 사용되고 있다. 본 연구에서는 ‘효능감’이라는 용어를 사용하며, 쓰기 평가에 대한 효능감을 뜻하는 용어로 ‘쓰기 평가 효능감’을 사용하도록 한다.

편의 작문을 공통적으로 평가하게 한 후 다국면 라쉬 모형 기반 FACETS 프로그램을 활용하여 분석하였다. 본 연구의 결과는 평가자로서의 교사에 대한 이해뿐 아니라 평가 교육에도 도움을 줄 수 있을 것이다.

2. 선행 연구

교사 효능감은 학생의 학업 성취에 영향을 미치는 중요한 변인일 뿐만 아니라 교사들이 학급에서 학생들의 학업 활동을 구조화하는 방식 및 학생들의 지적 능력을 평가하는 방식을 결정한다(Bandura, 1997). 이러한 중요성은 교사의 전문성이 요구되는 쓰기 평가 상황에서 더욱 두드러진다. 그럼에도 불구하고 한국어 교육에서는 평가와 관련된 교사 효능감 연구가 이루어지지 않았으며 이를 실제 교사의 평가 수행과 연결시켜 살펴 본 연구도 없었다. 이와 같이 교사의 평가에 대한 연구가 미비한 상황이며, 교사의 경력에 따른 쓰기 평가 특성을 문항반응이론에 근거한 다국면 라쉬 모형을 통해 구체적으로 살핀 연구도 이루어지지 않았다.

이처럼 한국어교육에서 쓰기 평가에 대한 효능감 및 이를 평가와 관련하여 분석한 연구는 아직 없으나, 교육 경력에 따른 교사 효능감을 분석한 연구(안정민, 2011; 안정민·김재욱, 2011; 정지은, 2012), 교사의 경력에 따라 평가 과정이 차이를 보인다는 연구(김동수, 2014)는 이루어진 바 있다. 본 연구에서는 한국어 교사의 쓰기 평가 효능감과 평가 특성이 어떠한 양상을 보이는지, 쓰기에서 평가 효능감과 평가 특성은 관계가 있는지 살피고자 한다.

인접 학문인 국어교육 분야에서는 쓰기와 관련된 교사의 효능감에 대한 연구가 드물게 이루어졌다(가은아, 2008; 박영민, 2010, 2011). 가은아(2008)에서는 평가자의 정의적 요소에 따라 평가 결과가 달라지는지를 분석하였다. 그러나 이는 교사의 '쓰기 평가 효능감'이 아닌 필자로서의 '쓰기 효능감'에 따른 평가 결과 비교였으며, 고전적 통계기법을 활용하

였기 때문에 평가자의 특성을 구체적으로 살피지는 못하였다.

박영민(2010)에서는 예비 국어 교사의 성별 및 교육 실습 경험 유무에 따른 쓰기 평가 효능감을, 박영민(2011)에서는 현직 국어 교사의 쓰기 평가 효능감을 성별 및 경력 차이에 따라 분석하였다. 이 연구에서는 경력이 증가함에 따라 쓰기 평가 효능감이 증가하는 경향을 보였으나, 통계적으로 유의미한 차이는 없었다고 하였다. 이 두 가지 연구에서는 효능감을 실제 평가와 함께 살피지는 않았으나, 교사의 쓰기 평가 효능감을 본격적으로 연구하였다는 데 의의가 있다.

문항반응이론에 기초한 다국면 라쉬 모형은 말하기와 쓰기 영역 평가 척도 개발이나 평가 경향 분석에 유용하다(Lumley & McNamara, 1995; McNamara, 1996; Weigle, 1998; 장소영·신동일, 2009). 한국어 교육에서는 학문 목적 쓰기 능력 평가 척도 개발 연구(김성숙, 2011), 말하기 평가 관련 연구(이향, 2012, 2013; 강석한·안현기, 2014)에서 다국면 라쉬 모형이 활용되었다. 영어교육 분야(이영식, 1998, 2014; 최인철, 1998, 2000; 신동일, 2001, 2002), 국어교육 분야(이현숙, 2008; 지은림, 2008; 박영민·최숙기, 2010; 박종임·박영민, 2011; 김성숙·유혜령, 2013; 박영민·박종임, 2013)에서는 보다 활발하게 다국면 라쉬 모형을 활용한 연구가 진행된 바 있다. FACETS 프로그램을 통한 분석은 고전적 통계방법보다 구체적인 정보를 제공하므로, 경력에 따른 교사의 평가 특성을 살피는 데에 보다 효과적일 것으로 보았다. 즉, 평가 점수 간의 상관계수를 통해 신뢰도를 분석한 고전적 통계방법은 평가 점수가 어느 정도 일치하는지를 보여주기는 하지만 그러한 평가의 차이가 나타나는 원인을 살피기는 어려웠다. 반면 FACETS 프로그램의 분석 결과는 평가자 개개인의 평가 대상에 대한 엄격성, 특정 평가 범주에 대한 편향성, 평가자 내 일관성 등 보다 구체적인 정보를 제공하여 평가자 그룹별 또는 개인에 맞는 평가자 훈련 연구에 도움을 줄 수 있다.

본 연구에서는 이러한 연구의 논의를 참고로 하여 한국어교사의 경력

에 따른 쓰기 평가 효능감 및 실제 평가 수행에서의 특성을 다국면 라쉬 모형을 이용하여 살핌으로써, 쓰기 평가자로서의 교사에 대한 이해를 돕고 향후 쓰기 평가 관련 연구에 교육적 시사점을 제공하고자 한다.

3. 연구 방법

3.1. 연구 대상 및 검사 도구

본 연구에서는 한국어 교사 39명의 쓰기 평가 효능감 검사지를 분석하여 교사의 쓰기 평가 효능감을 분석하였다. 또한 동일 교사들에게 각각 10편의 작문을 평가하게 하여 평가 특성을 살폈다. 본 연구의 대상이 된 39명은 모두 석사학위 소지자로 서울 소재 대학의 한국어교육기관에 재직하고 있다. 이중 2년 미만 경력[2]의 한국어 교사 13명은 집단 A, 2년 이상~5년 미만 경력의 한국어 교사 13명은 집단 B, 5년 이상 경력의 한국어 교사 13명은 집단 C로 분류하였다.[3]

---

2) 본 연구에 참여한 참가자들은 모두 대학의 한국어 교육 기관 소속으로 주당 12~20 시간 정도의 수업을 하고 있었다. 모든 참가자의 강의 시수가 동일하지는 않으므로 같은 기간의 경력이라고 하더라도 교사마다 투입된 시수, 수업의 종류, 학생 수 등이 다를 수 있다. 또한, 동일한 기간, 동일한 시수를 강의했다고 하더라도 개별적 몰입도, 호기심 등 여러 개인적 변인이 개입될 수 있다. 이와 같은 모든 변인을 파악하기 어려운 면이 있어 본 연구에서는 Kim(2010), 이향(2013) 등에서와 같이 교육 기간에 따라 집단을 구분하였다.

3) Kim(2010)에서는 평가자 그룹을 전문성에 의해 세 그룹으로 구분할 때 1년 이하를 초보(Novice), 2~3년을 중급(Developing), 4~5년 이상을 전문(expertise) 교사로 구분한 바 있다(이향, 2013:223). 이향(2013)에서는 경력 5년을 기준으로 경험이 많은 교사와 적은 교사로 분류하였다. 본 연구에서는 Kim(2010)의 연구와 같이 교사의 경력에 따라 세 집단으로 구분하였다.

한국어 교육 경력이 많아질수록 그만큼 다양한 숙달도의 여러 학습자를 만나게 되고 다양한 작문을 접하게 된다. 또한 쓰기 수업에서의 피드백 경험 및 성취도 평가 경험이 쌓여 숙달도 및 작문의 질에 대한 이해가 높아질 수 있다. 따라서 교육 경력이 평가 효능감에 영향을 미칠 것이라고 보았다.

본 연구는 ‘쓰기 평가 효능감’이라는 보다 세부적인 효능감을 측정하기 위하여 박영민(2010)에서 개발한 검사지를 활용하였다. 이 검사지를 통해 알 수 있는 쓰기 평가 효능감은 두 가지 구성 요인으로 이루어진다. ‘일반적 한국어 쓰기 평가 효능감’은 쓰기 평가를 일반적 차원의 관점에서 평가하고 판단하는 내용을 공유하고 있으며, ‘실행적 한국어 쓰기 평가 효능감’은 쓰기 평가 실행과 관련된 구체적인 사항을 공유하고 있다(박영민, 2010:193). 박영민(2010)에서 개발한 검사지는 L1 쓰기 평가를 대상으로 한 것이었으므로 L2 쓰기 평가 효능감 측정을 위해 연구자가 일부 문항을 수정하였다.[4)]

따라서 새롭게 구성된 검사지의 문항 신뢰도를 확인하고자 Cronbach α를 산출하였다. 그 결과, 17개 문항에 대한 Cronbach α는 .950으로 높은 수준의 검사 신뢰도를 보여 17개 항목을 모두 활용하였다.[5)] 모든 문항은 5점 리커트 척도로 이루어져 있다.

---

4) 박영민(2010:193)의 ‘일반적 한국어 쓰기 평가 효능감’ 부분에서 ‘나는 학생 글을 잘 쓴 글과 잘 못 쓴 글로 잘 변별하여 평가할 수 있다’라는 항목을 본 연구에서는 ‘나는 학생들의 글만으로도 학생의 쓰기 숙달도(초급, 중급, 고급)를 판단할 수 있다’와 ‘나는 학생 글을 학생이 속한 쓰기 숙달도 내에서 ‘상·중·하’로 잘 변별하며 평가할 수 있다’는 두 가지 항목으로 바꾸었다. 이는 모국어 쓰기 평가가 아닌 L2 쓰기 평가인 한국어 쓰기 평가에서는 해당 작문의 숙달도가 초·중·고급인지를 판단할 수 있어야하고 해당 숙달도 내에서 잘 쓴 글인지, 잘 못 쓴 글인지 역시 판단할 수 있어야 하기 때문이다. 또한 박영민(2010:193)에서는 ‘실행적 한국어 쓰기 평가 효능감’ 하위 문항을 ‘단어 선택, 표현, 조직이나 구성, 어법’과 같이 박영민(2010)에서 사용한 평가 범주를 중심으로 구성하였는데, 본 연구에서는 이를 본 연구의 평가 기준에 맞추어 수정하였다.

5) Cronbach α 계수는 0에서 1의 값을 가지며, 높을수록 바람직하나 반드시 몇 점 이상이어야 한다는 기준은 없다. 흔히 0.8~0.9 이상이면 바람직하고 0.6~0.7 이상이면 수용할 만한 것으로 여겨진다. 그러나 0.6보다 작으면 내적일관성을 결여한 것으로 받아들여진다. 이 경우 Cronbach α 계수의 크기를 저해하는 항목들을 제거함으로써 계수값을 크게 할 수 있다. 이러한 항목들은 그 항목과 전체 항목 간의 상관관계가 낮은 항목들이다(이학식·임지훈, 2011:121).

<표 1> 한국어 쓰기 평가 효능감 검사 문항

일반적 한국어 쓰기 평가 효능감	1. 나는 학생 글을 잘 평가(채점)할 수 있다.
	2. 나는 학생들의 쓰기 능력을 정확하게 평가할 수 있다.
	3. 나는 높은 신뢰도 수준을 유지하면서 학생 글을 평가할 수 있다.
	4. 나는 학생들의 글만으로도 학생의 쓰기 숙달도(초급, 중급, 고급)를 판단할 수 있다.
	5. 나는 학생 글을 학생이 속한 쓰기 숙달도 내에서 '상, 중, 하'로 잘 변별하며 평가할 수 있다.
	6. 나는 학생 글을 평가할 때 평가 기준을 일관성 있게 적용할 수 있다.
실행적 한국어 쓰기 평가 효능감	7. 나는 채점의 엄격성(점수를 관대하게 또는 엄격하게 주는 정도)을 일관성 있게 유지하며 학생 글을 평가할 수 있다.
	8. 나는 학생 글에서 '내용 및 과제 수행'과 관련된 (명료성, 통일성, 창의성, 풍부성 등)를 잘 평가할 수 있다.
	9. 나는 학생 글에서 '전개구조(조직, 구성, 담화표지 사용 등)'가 적절한지를 잘 평가할 수 있다.
	10. 나는 학생 글에서 '언어사용(어휘와 문법 등)의 정확성'을 잘 평가할 수 있다.
	11. 나는 학생 글에서 '언어사용(어휘와 문법 등)의 다양성'을 잘 평가할 수 있다.
	12. 나는 학생 글에서 '글의 목적과 기능에 따른 격식'에 맞춰 글을 썼는지 잘 평가할 수 있다.
	13. 나는 학생 글을 평가할 때 쓰기 윤리를 위반했는지(예-쓰기를 할 때 인터넷, 친구의 글을 베끼는 행위 등)를 가려낼 수 있다.
	14. 나는 학생 글을 평가할 때 다른 동료 평가자와 협의할 수 있다.
	15. 나는 평가 영향 요인(예-성별, 나이 등)을 적절히 통제하며 학생 글을 평가할 수 있다.
	16. 나는 마음이 편안한 상태에서 학생 글을 읽고 평가할 수 있다.
	17. 나는 쓰기 평가와 관련하여 어떤 문제가 발생했을 때 그 평가 관련 문제의 원인이 어디에 있는지 파악할 수 있다.

또한 평가자 39인의 평가 특성을 살피기 위한 작문 10편은 '성공의 조건과 그 이유에 대해 600~800자로 쓰시오.'라는 쓰기 과제에 따라 작성된 5급 초반 수준 학생들의 작문이었다. 작문을 작성한 학생은 중국인 5명, 몽골인 2명, 일본인 2명, 말레이시아인 1명이다. 본 연구에서 사용된 쓰기 평가 기준은 한국어능력시험(TOPIK) 평가 기준을 참조하였다.[6)]

<표 2> 분석적 채점 범주 및 채점표

구분	채점 근거	점수 구분
내용 및 과제 수행 (5점)	1) 주어진 과제를 충실히 수행하였는가? 2) 주제와 관련된 내용으로 구성하였는가? 3) 내용을 풍부하고 다양하게 표현하였는가?	① ② ③ ④ ⑤
글의 전개 구조 (5점)	1) 글의 구성이 명확하고 논리적인가? 2) 중심 생각이 잘 구성되어 있는가? 3) 논리 전개에 도움이 되는 담화 표지를 적절하게 사용하여 조직적으로 연결하였는가?	① ② ③ ④ ⑤
언어 사용 (5점)	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?	① ② ③ ④ ⑤

3.2. 연구 절차 및 분석 방법

본 연구에 참가한 39인의 교사는 한국어 쓰기 평가 효능감 검사지의 17문항에 응답하였으며, 10편의 작문을 평가하였다. 평가 결과의 특성을 그대로 살피기 위해 별도의 평가자 훈련이나 협의 과정 없이 실험을 진행하였다. 평가 과정의 실제성을 높이기 위하여 작문 하단에 <표 2>의 평가표를 넣어 곧바로 세 가지 평가 범주의 점수에 체크할 수 있게 하였다. 채점은 한 곳에서 동시에 이루어진 것이 아니라 각 평가자가 개별적으로 진행하였다.

6) 본 연구에서 리커트 척도를 활용한 것은 다국면 라쉬 모형을 통해 채점자의 특성을 분석하기 위해서이다. 또한 본 연구에서 중점을 두는 것은 교사의 평가 경향이므로, 이와 같이 5점 척도로 동일하게 점수를 부여하는 것이 문제가 되지 않으며, 오히려 범주 활용의 차이를 보다 잘 드러내 줄 것으로 보았다. 실제 한국어능력시험의 분석적 기준의 경우(제35회 한국어능력시험 II 54번 문항의 경우) 내용 및 과제수행 12점, 전개구조 12점, 언어사용 13x2와 같이 점수를 배분하고 있다.

세 집단의 한국어 쓰기 평가 효능감 검사 결과를 분석하기 위해서 SPSS 12.0을 활용하여 일원배치 분산분석을 실시하였다. 쓰기 평가 특성 분석을 위해서는 FACETS 프로그램[7]을 활용하였는데, '학생의 작문, 평가자, 평가 범주, 교사 집단'이라는 4개의 국면을 적용하여 분석하였다. 평가 양상은 다국면 라쉬 모형을 통해 나타난 엄격성과 일관성을 중심으로 분석되었다.[8] 구체적인 분석 결과 해석에 앞서 평가 자료가 다국면 라쉬 모형에 적합한지 알아보는 모형 적합도 분석 그래프를 확인한 결과, 교사 39명의 평가 자료는 라쉬 모형에 적합함을 알 수 있었다.[9]

4. 한국어 교사의 쓰기 평가 효능감과 평가 특성

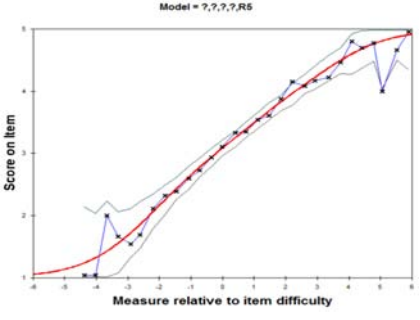
4.1. 한국어 교사의 쓰기 평가 효능감

---

7) 무료로 공개된 FACETS Version No. 3.71.4를 활용하였다.

8) 엄격성이란 평가자가 평가 기준을 엄격하게 적용하여 채점하는 정도를 나타내는 것인데, Linacre(1989:48-49)에서는 이를 전체적 엄격성과 특정 범주에 대한 해석 또는 적용 차이로 인한 엄격성이라는 두 가지 엄격성이 있다고 보았다(Lumley, McNamara, 1995:55). 본 연구에서는 이러한 전체적 엄격성과 범주에 따른 엄격성 모두를 연구 대상으로 하였다.

9) 오른쪽의 그래프는 분석 자료의 모형 적합도를 보여준다. 굵은 곡선인 문항특성곡선을 중심으로 그에 따른 95%신뢰구간이 양쪽 측면에 얇은 선으로 표시된다. 'x'를 연결한 선인 관찰지 곡선이 문항특성곡선을 따라서 얼마나 잘 부합되는지를 시각적으로 판단할 수 있다. 전체적으로 채점 자료가 라쉬 모형에 적합하다면, 즉 관찰된 채점 점수들이 신뢰 구간 안에 적절하게 위치한다면 해당 자료를 활용한 논의는 의미를 갖게 된다(장소영·신동일, 2009:121). 위의 그래프를 보면, 관찰지 곡선이 대체로 신뢰구간 안에 위치하고 있다. 따라서 본 연구의 자료는 라쉬 모형에 적합하다고 볼 수 있다.



세 집단의 한국어 쓰기 평가 효능감의 구성 요인별 평균 및 분산분석 결과는 <표 3>과 같다. 일반적 쓰기 평가 효능감, 실행적 쓰기 평가 효능감 모두 경력에 따라 증가하는 경향을 보였으며, 세 집단 중 가장 뚜렷한 차이를 보이는 집단은 경력 2년 미만의 집단 A였다. 집단 간 평균차이는 A, B 간의 차이가 B, C 간의 차이에 비해 컸다.

일원배치 분산분석의 사후 분석 결과, 집단 B와 C의 차이는 통계적으로 유의하지 않은 데 비해, A와 B, A와 C의 차이는 유의한 것으로 나타났다. 즉, A집단의 쓰기 평가 효능감은 다른 두 집단과 모두 통계적으로 유의한 차이를 보였다. 2년 미만의 집단 A와 2년 이상 경력의 B, C가 뚜렷한 차이를 보였다는 결과를 통해, 개인차는 존재하겠으나, 어느 정도 효능감이 높아진 뒤에는 계속 증가한다기보다는 유지되거나 약간 높아지는 경향이 있음을 알 수 있다.

<표 3> 한국어 쓰기 평가 효능감의 구성 요인별 통계

구분	집단 (명)	평균 (표준편차)	분산분석 집단 간 유의확률	사후 분석 (다중비교)
일반적 한국어 쓰기 평가 효능감	A(13)	2.59 (.60)	.000**	A-B .000**
	B(13)	3.77 (.35)		A-C .000**
	C(13)	3.81 (.58)		B-C .982
	계(39)	3.39		
실행적 한국어 쓰기 평가 효능감	A(13)	3.13 (.33)	.000**	A-B .000**
	B(13)	3.89 (.37)		A-C .000**
	C(13)	3.92 (.53)		B-C .990
	계(39)	3.65		
전체	A(13)	2.94 (.37)	.000**	A-B .000**
	B(13)	3.85 (.34)		A-C .000**
	C(13)	3.87 (.50)		B-C .986
	계(39)	3.56		

P <.05 수준에서 통계적으로 유의미함.

정의적 영역에 속하는 효능감은 인지적 능력처럼 적층적인 특징을 보이지 않는 경향이 있어 경력이 더 많다고 해서 쓰기 평가 효능감이 반드시 높다고는 할 수 없으며, 경력이 많은 교사일수록 쓰기 평가에 대한 부정적인 경험이 더 많아 쓰기 평가 효능감이 오히려 더 낮을 수도 있다(박영민, 2011:139). <표 3>을 보면, 전체적으로 쓰기 평가 효능감은 C집단이 가장 높지만, 표준편차 역시 큰 것을 알 수 있다. 이는 경력이 가장 많은 C집단에 속한 교사들의 쓰기 평가 효능감이 개인차가 컸음을 의미한다. 실제로 C집단에 속한 한 남자 교사는 쓰기 평가 효능감 검사지에 모두 5점으로 답해 가장 높은 효능감을 보인 반면, 한 여자 교사는 경력이 비슷함에도 불구하고 3.24라는 C집단에서 가장 낮은 효능감을 보였다. 이 교사는 특히 일반적 효능감에서 2.50이라는 매우 낮은 효능감을 보였다.[10] 이를 통해 대체로 경력이 쌓이면 쓰기 평가 효능감이 높아지기는 하나, 다양한 평가 경험을 통해 부정적인 평가 경험도 하게 되면서 자신이 쓰기 평가 능력에 대한 자신감을 높이지 못했거나 반성적으로 바라보는 교사가 생기게 되어 집단 안에서 개인차가 커질 수 있음을 알 수 있다.

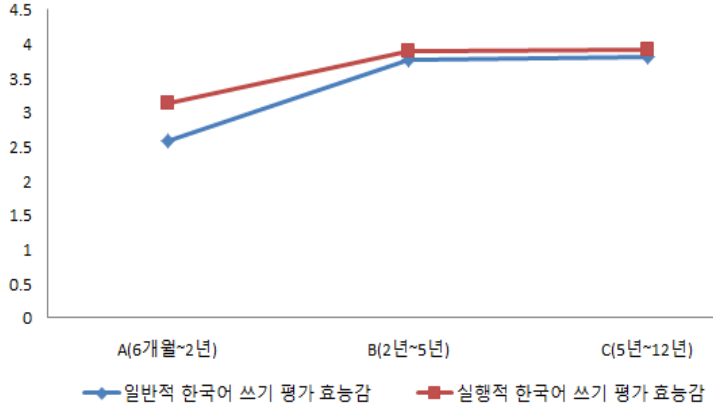
교사의 경력은 시간이 지나 경력이 쌓여야만 바꿀 수 있는 부분이다. 반면, 교사의 효능감은 평가 역량과 연관을 맺으면서도 여러 방법을 통해 향상시킬 수 있는 정의적 요인이다. 평가 효능감이 낮을 경우 자신이 직면한 과제를 실제보다 어렵게 느끼고 스트레스를 받을 확률이 높아지며, 이는 전체적으로 과제 수행 과정을 조율하는 데에 부정적인 영향을

10) C집단에서 가장 낮은 쓰기 평가 효능감을 보인 A교사와 가장 높은 효능감을 보인 B교사는 모두 5년 이상의 경력을 가진 교사이며, 한국어교육학으로 박사과정을 수료하였다는 점도 동일하다. 그러나 두 교사의 쓰기 평가 효능감은 매우 다른 양상을 보였다. 사후 인터뷰에서 A교사는 자신이 경력이 많기 때문에 그만큼 평가를 잘 해야 한다는 걱정이 있으며, 본인의 평가가 학생들에게 끼치는 영향이 크다고 여겨 부담을 갖게 된다고 하였다. 반면 B교사는 자신의 교육 및 평가 경험에 대해 강하게 신뢰하는 경향이 있었으며, 그러한 경험으로 인해 평가 과정의 엄격성 및 신뢰도를 잘 조절할 수 있다고 생각하고 있었다.

끼칠 수 있다. 이에 반해 효능감이 높을 경우 보다 자신감을 갖고 과정 전체를 관리하며, 인지적·정의적 부담을 스스로 통제할 수 있는 여력이 생기게 된다. 평가 과정에서도 효능감은 과제 수행 전반에 영향을 미친다. 쓰기 평가 효능감이 높아지게 되면 평가에 확신과 자신감을 주어 일관성이 높아질 수 있고 이것으로 인한 평가 성공의 경험은 평가 효능감의 상승으로 이어져 자연스럽게 효능감과 평가 수행은 긍정적인 순환을 하게 된다. 반면 평가 과정에서의 부정적인 경험은 효능감의 하락으로 이어질 수 있고, 이는 평가 수행에 다시 부정적으로 작용할 수 있다.

따라서 교사를 대상으로 한 평가 교육을 통해 성공 경험을 높이고, 이를 통해 평가 효능감을 높여 실제 평가 수행에 긍정적으로 작용하도록 도와야 한다. 앞서 언급한 C집단의 교사와 같이 경력이 많음에도 불구하고 쓰기 평가 효능감이 낮은 경우, 한국어 교육 기관 내에서의 평가 정보 공유, 평가 교육 프로그램 및 워크숍을 통해 쓰기 평가 효능감을 높일 기회를 제공할 필요가 있다.

또한 <그림 1>을 통해 쓰기 평가 효능감의 구성 요인별 차이를 알 수 있다. 본 연구에서 한국어 쓰기 평가 효능감은 일반적, 실행적 쓰기 평가 효능감으로 구분되는데, 세 집단 모두 일반적 한국어 쓰기 평가 효능감이 실행적 한국어 쓰기 평가 효능감보다 낮게 나타났다. 쓰기 평가자로서의 숙련도가 가장 낮을 것으로 보이는 A집단은 두 가지 구성 요인에서 모두 가장 낮은 효능감을 보였지만 그 차이는 일반적 한국어 쓰기 평가 효능감에서 더 컸다.



<그림 1> 한국어 쓰기 평가 효능감의 구성요인별 차이 양상

그 이유는 <표 4>에서 쓰기 평가 효능감의 두 가지 구성요인을 검사하기 위한 문항별 일원배치 분산분석 결과를 통해 유추할 수 있다. 일반적 쓰기 평가 효능감을 구성하고 있는 6개 문항은 모두 A집단이 다른 두 집단과 통계적으로 유의미한 차이가 있었던 반면, B와 C 사이에는 유의한 차이가 없어, 일반적 쓰기 평가 효능감 전체의 차이 역시 두드러졌다. 이는 일반적 한국어 쓰기 평가 효능감이 '나는 학생 글을 잘 평가(채점)할 수 있다', '나는 학생들의 쓰기 능력을 정확하게 평가할 수 있다', '나는 높은 신뢰도 수준을 유지하면서 학생 글을 평가할 수 있다', '나는 학생 글을 학생이 속한 쓰기 숙달도 내에서 '상, 중, 하'로 잘 변별하며 평가할 수 있다'와 같이 보다 일반적 차원, 즉 평가 전반에 대한 효능감을 묻고 있기 때문에 경력이 적은 A집단에서는 높은 자신감을 보일 수 없었던 것으로 보인다. 이는 A집단의 효능감이 B, C집단과 확연한 차이를 보이게 하였다.

이와 달리, 실행적 쓰기 평가 효능감을 구성하고 있는 문항들의 경우, A집단이 B, C와 통계적으로 유의한 차이를 보이지 않은 문항도 나타나

일반적 쓰기 평가 효능감과는 다른 양상을 보였다. ‘나는 학생 글을 평가할 때 다른 동료 평가자와 협의할 수 있다(14번)’와 ‘나는 평가 영향 요인을 적절히 통제하며 학생 글을 평가할 수 있다(15번)’에서 세 그룹은 모두 유의미한 차이를 보이지 않았다. ‘나는 학생 글을 평가할 때 다른 동료 평가자와 협의할 수 있다(14번)’라는 문항의 경우 세 집단 모두 효능감 평균이 모두 4점 이상으로 높게 나타나 경력별 차이가 나타나지 않은 것이며, ‘나는 평가 영향 요인을 적절히 통제하며 학생 글을 평가할 수 있다(15번)’ 문항의 경우 A는 3.38, B는 3.85, C는 3.69로 세 집단 모두 중간보다 약간 높은 수준의 비슷한 효능감을 보여 통계적으로 유의미한 차이가 없었던 것이다. 이는 경력과 크게 상관없이 한국어 교사가 평가 시 협의하는 것에 대해서 비교적 호의적임을 보여주며, 글씨, 성별, 연령 등 평가 영향요인을 통제하는 능력에 있어서는 경력이 증가해도 크게 효능감이 향상되지 않았음을 보여준다.

또한 ‘나는 마음이 편안한 상태에서 학생 글을 읽고 평가할 수 있다(16번)’라는 항목에서는 A, B 집단 간의 차이는 있었으나 A와 C, B와 C 사이에는 차이가 없었다. A가 B와 차이를 보이는 것은 다른 문항들의 특성과 비슷하지만, A가 C와 차이를 보이지 않은 점은 특이한 부분이다. 이는 오히려 2년 이상~5년 미만의 경력은 어느 정도 쓰기 평가 경험이 쌓여 초보 평가자일 때보다 심리적으로 안정된 반면, 5년 이상의 경력이 되면 오히려 평가를 잘 해야 한다는 부담 느끼는 부분이 그러한 원인으로 작용한 것으로 추측해 볼 수 있다. 또는 다양한 평가 상황을 접하면서 신뢰도의 문제 등 부정적 경험이 작용하였을 수도 있다.

<표 4> 한국어 쓰기 평가 효능감의 문항별 통계 결과

	문항	A	B	C	집단간 유의 확률(P)	사후분석		
						A-B	A-C	B-C
일반적 한국어 쓰기 평가 효능감	1. 나는 학생 글을 잘 평가(채점)할 수 있다.	2.5	3.6	3.8	.000**	.001**	.000**	0.849
	2. 나는 학생들의 쓰기 능력을 정확하게 평가할 수 있다.	2.5	3.6	3.6	.000**	.001**	.000**	1
	3. 나는 높은 신뢰도 수준을 유지하며 학생 글을 평가할 수 있다.	2.6	3.5	3.7	.000**	.001**	.000**	0.793
	4. 나는 학생들의 글만으로도 학생의 쓰기 숙달도(초급, 중급, 고급)를 판단할 수 있다.	2.9	4.2	4	.000**	.000**	.000**	0.821
	5. 나는 학생 글을 학생이 속한 쓰기 숙달도 내에서 '상, 중, 하'로 잘 변별하며 평가할 수 있다.	2.5	3.9	3.9	.000**	.000**	.000**	1
	6. 나는 학생 글을 평가할 때 평가 기준을 일관성 있게 적용할 수 있다.	2.6	3.9	3.9	.000**	.000**	.000**	0.953
실행적 한국어 쓰기 평가 효능감	7. 나는 채점의 엄격성(점수를 관대하게 또는 엄격하게 주는 정도)을 일관성 있게 유지하며 학생 글을 평가할 수 있다.	2.9	3.7	3.9	.000**	.004**	.001**	0.807
	8. 나는 학생 글에서 '내용 및 과제 수행'(명료성, 통일성, 창의성, 풍부성 등)을 잘 평가할 수 있다.	3	3.9	3.9	.002**	.006**	.006**	1
	9. 나는 학생 글에서 '전개구조(조직, 구성, 담화표지 사용 등)'가 적절한지를 잘 평가할 수 있다.	3.2	4	4.1	.002**	.007**	.003**	0.944
	10. 나는 학생 글에서 '언어사용(어휘와 문법 등)의 정확성'을 잘 평가할 수 있다.	3.2	4.1	4.1	.001**	.004**	.004**	1
	11. 나는 학생 글에서 '언어사용(어휘와 문법 등)의 다양성'을 잘 평가할 수 있다.	2.9	3.5	4	.000**	.015*	.000**	0.082
	12. 나는 학생 글에서 '글의 목적과 기능에 따른 격식'에 맞	3.5	4.2	4.2	.016*	.031*	.031**	1

줘 글을 썼는지 잘 평가할 수 있다.							
13. 나는 학생 글을 평가할 때 쓰기 윤리를 위반했는지(예-쓰기를 할 때 인터넷, 친구의 글을 베끼는 행위 등)를 가려낼 수 있다.	2.5	3.7	3.7	.001**	.002**	.002**	1
14. 나는 학생 글을 평가할 때 다른 동료 평가자와 협의할 수 있다.	4	4.3	4.2	0.357	0.324	0.749	0.749
15. 나는 평가 영향 요인(예-성별, 나이 등)을 적절히 통제하며 학생 글을 평가할 수 있다.	3.4	3.9	3.7	0.279	0.26	0.542	0.856
16. 나는 마음이 편안한 상태에서 학생 글을 읽고 평가할 수 있다.	3.4	4.1	3.9	.034*	.029*	0.19	0.649
17. 나는 쓰기 평가와 관련하여 어떤 문제가 발생했을 때 그 문제의 원인이 어디에 있는지 파악할 수 있다.	2.4	3.6	3.7	.000**	.001**	.000**	0.964

P <.05 수준에서 통계적으로 유의미함.

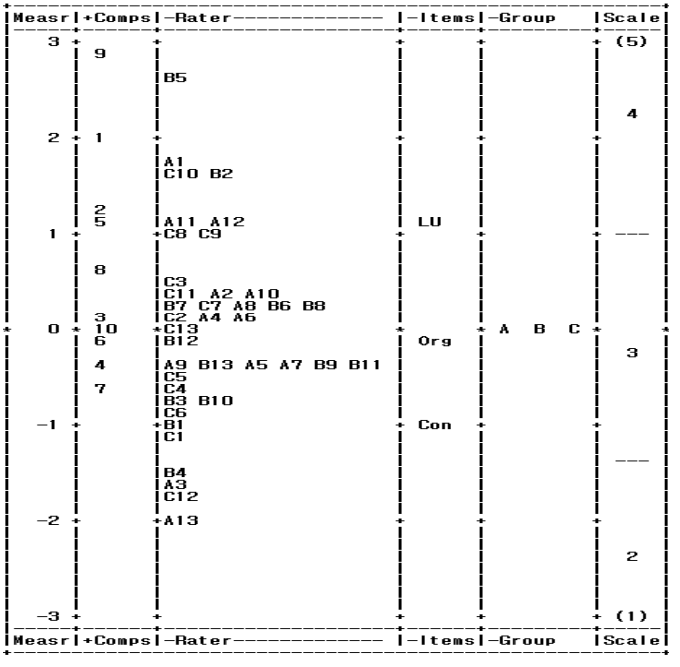
4.2. 한국어 교사의 쓰기 평가 특성

이 장에서는 다국면 라쉬 모형을 활용하여 한국어 교사의 평가 수행이 어떠한 특성을 보이는지 엄격성과 일관성을 중심으로 분석하고, 이를 쓰기 평가 효능감과 함께 살피고자 한다.

4.2.1. 한국어 쓰기 평가의 엄격성과 효능감

먼저 평가 국면의 엄격성을 중심으로 평가 특성을 살펴보도록 하겠다. <그림 2>의 측정 단면 분포도는 작문의 수준 및 엄격성 정보를 보여준다. 가장 왼쪽의 'Measr'는 로짓(logits) 측정 단위로, 각 국면의 평균을

‘0’으로 고정시켜 양수치일 때와 음수치일 때의 상대적인 의미를 나타낸다. 이 로짓 수치를 기준으로 각각의 국면의 정보를 해석할 수 있다(이향, 2013:227). 두 번째 열의 ‘Comps’는 학생들의 작문 10편의 수준을 보여준다. 점수가 높을수록 척도의 위쪽에, 점수가 낮을수록 아래쪽에 위치한다. 따라서 본 연구에서 최고점을 받은 글은 9번 작문이며, 최하점을 받은 글은 7번 작문임을 알 수 있다. 척도에서 가까운 거리에 위치해 있으면 비슷한 수준의 작문이다.



<그림 2> 측정 단면 분포도 (작문, 평가자, 평가 범주, 집단)

다음 칸의 ‘Rater’[11]는 평가자 39명의 엄격성을 나타내준다. 로짓값 0

11) ‘rater’값 앞에 ‘-’ 표시는 점수를 적게 주는, 엄격한 평가자가 보다 위쪽에 위치할 수 있도록 프로그램을 구성한 것이다(박종임, 2013:53).

을 기준으로 위쪽에 위치할수록 엄격한 평가자이고, 아래쪽에 위치할수록 관대한 평가자이다. 분석 결과에 따르면, 본 연구에 참여한 평가자들의 엄격성은 주로 -1~+1로짓 사이에 분포하고 있지만 일부 평가자들은 이를 벗어나 있음을 알 수 있다.[12] B집단의 평가자인 B5는 39명 중 가장 엄격하게 평가하였고, A집단의 평가자인 A13은 가장 관대하게 평가하였다. 로짓값이 0에 가까운, 중간 정도의 엄격성을 보인 평가자는 C13이었다.

<표 5>는 평가자별 엄격성 수치를 나타낸 표이다. 로짓값이 -1~+1을 비교적 크게 벗어난 평가자의 경우 엄격성에서 큰 차이를 보인 평가자로 보고 음영 표시를 하였다. A와 B집단에서는 각 3명, C집단에서는 2명이 이에 해당됨을 알 수 있다.

또한 <표 5>의 'Model, Populn'이라고 표기된 줄에서는 현재 평가 결과에 해당되는 집단(즉 10편의 작문에 대한 39명의 평가)을 전체 모집단으로 보고 산출된 수치를 제공하고, 반면에 그 밑의 'Model, Sample' 줄에서는 현재 채점 결과의 집단을 하나의 샘플 집단으로 보고 전체 모집단을 추정한 수치를 제공한다(이영식, 2014:484). 본 연구에서는 두 가지 모두 신뢰도(reliability)가 .92로 평가자들의 엄격성에는 많은 차이가 있는 채점을 하였고, 분리도(separation)가 3.32, 평가자 층(Strata)의 지수 4.76으로 평가자 간 약 5개 층위로 나뉘는 평가를 하였다. 이는 세 집단에 속한 평가자들 사이에 엄격성이 존재함을 뜻하는 것으로, 평가자 간 일치도가 낮음을 보여주는 것이다.

12) 평가자들이 2로짓 정도의 상대적 차이를 보인다면, 이는 가장 관대한 평가자 대신 가장 엄격한 평가자가 평가한다면 피험자 능력모수의 40% 정도가 감소될 수 있음을 뜻한다. 이는 심각한 결과이기는 하나 희귀한 사례는 아니다. 실제 거의 대부분의 상황에서 이러한 평가 유형이 발생한다(McNamara, 1996:187). 그러나 본 연구에서 B5과 A13은 4.71로짓이라는 큰 차이를 보여, 이 두 평가자는 서로 매우 다른 엄격성을 가지고 같은 작문을 평가하였음을 알 수 있다.

<표 5> 평가자별 엄격성

A집단	엄격성	B집단	엄격성	C집단	엄격성
A1	1.69	B1	-1.04	C1	-1.10
A2	0.39	B2	1.61	C2	0.17
A3	-1.66	B3	-0.78	C3	0.49
A4	0.15	B4	-1.50	C4	-0.66
A5	-0.42	B5	2.66	C5	-0.49
A6	0.07	B6	0.22	C6	-0.83
A7	-0.42	B7	0.30	C7	0.25
A8	0.23	B8	0.22	C8	1.03
A9	-0.34	B9	-0.43	C9	1.03
A10	0.39	B10	-0.78	C10	1.63
A11	1.09	B11	-0.43	C11	0.41
A12	1.09	B12	-0.10	C12	-1.74
A13	-2.05	B13	-0.35	C13	0.01
Model, Populn: RMSE .29 Adj (True) S.D. .96 Separation 3.32 Strata 4.76 Reliability .92					
Model, Sample: RMSE .29 Adj (True) S.D. .97 Separation 3.37 Strata 4.82 Reliability .92					

<그림 2>의 네 번째 열은 평가 범주(items)에 대한 엄격성의 정도를 보여준다. 평가자 39명이 가장 엄격하게 평가한 범주는 언어 사용(LU)이며, 중간 정도의 엄격성을 보인 범주는 전개구조(Org)이고 가장 관대하게 점수를 부여한 범주는 내용 및 과제 수행(Con) 범주이다. 엄격성은 가장 관대하게 평가된 내용 및 과제 수행 범주(-.97)에서부터 가장 엄격하게 평가된 언어 사용 범주(1.10)까지 총 2.07로짓에 분포하고 있다.

다섯 번째 열의 'Group'에는 집단별 엄격성의 정도가 나타나 있는데, 분리도(separation)가 .00, 신뢰도(reliability)가 .89로 나타나 경력별로 구분된 A, B, C 세 집단 간 엄격성의 차이는 없다고 볼 수 있다. 'rater' 열에서 살펴 본 바와 같이 평가자 39명 개인별 엄격성은 전체적으로 차이가 있었지만, 엄격한 평가자와 관대한 평가자가 집단별로 섞여 있어 집단 간 엄격성의 차이는 크지 않았던 것으로 보인다.

또한 이러한 엄격성에 따라 로짓값 '0'을 기준으로 경력과 상관없이 양수의 값을 가지는 엄격한 평가자 집단과 음수의 값을 가지는 관대한

집단으로 나누어 Mann-Whitney 검정으로 쓰기 평가 효능감을 비교했을 때, <표 6>과 같이 통계적으로 유의한 차이가 없었다. 따라서 경력에 따라서 쓰기 평가 효능감에는 차이가 있었으나, 전체적 점수에 대한 엄격성과 효능감은 상관을 보이지 않았다고 볼 수 있다.

<표 6> 평가 엄격성 수준에 따른 쓰기 평가 효능감 검정 유의확률

구분	평가자	일반적 한국어 쓰기 평가 효능감(P)	실행적 한국어 쓰기 평가 효능감(P)
엄격한 평가자 (21명)	A1, A2, A4, A6, A8, A10, A11, A12, B2, B5, B6, B7, B8, C2, C3, C7, C8, C9, C10, C11, C13	.173	.389
관대한 평가자 (18명)	A3, A5, A7, A9, A13, B1, B3, B4, B9, 10, B11, B12, B13, C1, C4, C5, C6, C12		

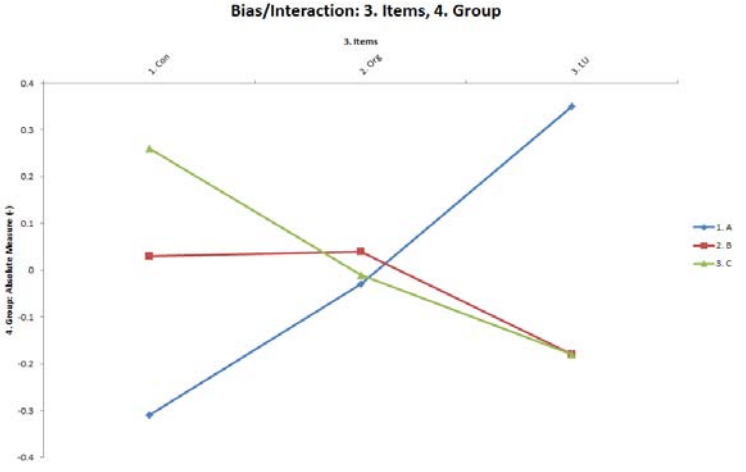
P < .05 수준에서 통계적으로 유의미함.

그런데 주지해야 할 점은, <표 6>에서 집단 구분의 기준으로 삼은 평가의 엄격성은 세 가지 평가 범주별이 아닌, 세 평가 범주의 총합을 기준으로 한 것이라는 점이다. 따라서 총점의 평가 엄격성에 따른 쓰기 평가 효능감에는 차이가 없었더라도, 평가 범주별 엄격성의 차이는 있을 수 있으며, 이러한 평가 범주별 엄격성과 쓰기 평가 효능감이 연관을 맺고 있을 수 있다. 경력별 집단에 따라서도 동일한 평가 범주별 엄격성이 나타나는지는 <그림 3>의 평가자 집단별 평가 범주 편향성 분석 결과를 통해 살펴볼 수 있다.[13]

<그림 3>을 보면, 세 집단의 엄격성이 평가 범주에 따라 다른 것을 알

13) 평가자가 특정 피험자 집단이나 특정 문항에 대해서만 엄격함이나 관대함을 보이는 경우도 있다. 이러한 체계적인 행동 유형은 다국면 라쉬 모형의 편향성 분석을 통해 파악이 가능하다. 이를 상호작용 효과(interaction effect)라고 부르기도 한다(McNamara, 1996:193-194). 본 연구에서는 특정 평가자 집단이 특정 평가 범주에 대해 어떠한 편향성을 보이는지 분석하였다.

수 있다. A집단은 내용 및 과제 수행 범주를 가장 관대하게, 전개구조를 중간 정도로, 언어 사용을 가장 엄격하게 평가하고 있다. 반면, B집단은 내용 및 과제 수행 범주와 전개구조에서 중간 수준의 엄격성을 보이고, 언어 사용은 상대적으로 관대하게 평가하였다. C집단은 A와 정반대의 엄격성의 경향을 보이는데, 오히려 내용 및 과제 수행을 다른 집단에 비해 엄격하게, 언어 사용을 가장 관대하게 평가하였다.



<그림 3> 평가자 집단-평가 범주 간 엄격성의 편향 분석 그래프

이러한 특성은 교사들의 평가 범주에 대한 해석의 차이가 반영된 것으로 볼 수 있다. 본 연구의 평가 대상이 된 작문 10편은 모두 성공의 조건 및 이유에 대해 기술하기는 하였기 때문에 어느 정도 성공적으로 과제를 수행하였다. 따라서 이들 작문 10편의 수준을 상·중·하로 변별하는 것은 5급 수준 작문에서의 내용 및 과제 수행에 대한 판단 능력을 필요로 한다. 경력이 많은 교사들의 경우 내용 및 과제수행 부분에서도 잘 쓴 글과 못 쓴 글의 수준이 이미지화되어 있고 내제화된 구체적인 기준이 있다.

또한 일관성뿐 아니라 내용이 명확하고 적합한지, 풍부함과 다양성을 가지고 있는지도 내적 기준으로 삼고 평가하며, 이는 분석적 평가 범주에 대한 세부적인 기준을 세우게끔 하여 해당 범주에 대한 변별력이 생기게 된다.

반면 경력이 적은 교사의 경우 중급 후반~고급 초반 수준의 작문 평가 경험이 적어 작문 수준에 대한 이미지화가 잘 되어 있지 않고, 내용의 일관성과 같은 하나의 기준에 치우치는 경우가 많다. 그 결과 본 연구에서 평가된 대부분의 작문이 내용 및 과제 수행을 성공적으로 하였다고 보고 후한 점수를 주게 되며, 작문 점수의 변별은 '언어 사용'에서 하게 될 수 있는 것이다.

언어 사용의 경우 A집단이 가장 엄격하게 평가한 범주이다. 평가 경험이 적은 평가자들은 내용 및 과제 수행, 전개 구조에서 비슷한 점수를 주고 언어 사용에서 엄격한 기준을 부여하여 작문의 점수를 변별하는 경향이 있을 수 있다. 또한 언어 사용의 수준을 전체적으로 파악하기보다는 특정 어휘 사용에 치우치거나 오류의 횟수를 세어 평가하는 경향이 평가에 영향을 미치기도 한다. 또한 경력이 적은 평가자들은 다양한 접하지 못해 중간언어에 대한 친숙도가 낮고 언어 사용에 대한 기대치가 높아 언어 사용 범주를 엄격하게 평가할 수 있다.

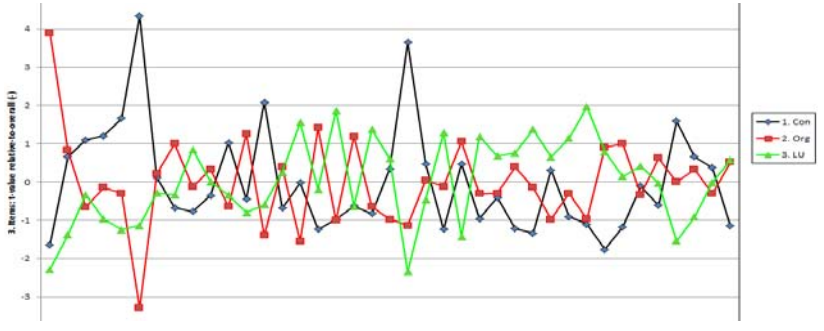
<표 7>은 평가 범주에 따른 집단-평가 범주 간 편향성의 정도를 수치로 보여준다. AM(Absolute Measurement)은 <그림 3>에서 확인한 세 집단의 평가 범주별 엄격성의 편향 수치를 나타내는 것이다. TV(t-value relative-to-all)는 기대 점수와 관찰 점수 간의 차이를 보여주는 그래프로, t 값이 -2~+2 범위 안에 있으면 그 편향성이 통계적으로 유의미지 않고, 이 범위 밖에 있는 경우 다시 평가해야 할 정도로 문제가 될 수 있음을 뜻한다. 따라서 A, B, C 집단 모두 평가 범주별 엄격성에 차이를 보이고 있으나, B와 C 집단은 모두 t 값이 -2~+2 범위 안에 위치하여 그 편향성이 문제가 되지는 않는다. 반면, A집단은 내용 및 과제수행 범주의 t값이

2.31로 통계적으로 유의미하게 관대하였고, 언어 사용 범주의 t값이 -2.48로 통계적으로 유의미하게 엄격하여 두 범주의 사용에서 문제가 있었음을 알 수 있다.

<표 7> 집단-평가 범주 간 편향성의 정도

평가 범주	내용 및 과제 수행(Con)		전개구조(Org)		언어 사용(LU)	
집단	AM(-)	TV(-)	AM(-)	TV(-)	AM(-)	TV(-)
A	-0.31	2.31	-0.03	0.39	0.35	-2.48
B	0.03	-0.5	0.04	-0.56	-0.18	1.04
C	0.26	-1.74	-0.01	0.21	-0.18	1.44

<그림 4>는 그룹별이 아닌, 평가자 39명 개개인의 평가 범주 간 엄격성의 편향 정도(AM)를 보여준다. 가장 왼쪽의 평가자가 A1이며, 그 다음이 A2, 가장 오른쪽의 평가자가 평가자 C13이다. 몇몇 평가자들은 범주별로 엄격성의 정도차가 매우 큰 것을 알 수 있다.



<그림 4> 평가자 별-평가 범주 간 엄격성의 편향 분석 그래프

평가자 39명의 범주별 엄격성의 편향 정도가 평가에 있어 문제가 될 정도인지 아닌지는 <표 8>의 TV값을 통해 알 수 있다. 특히 왼쪽부터 A1, A6, A13, B8은 그 편향성의 정도가 매우 크다. 편향성이 문제가 될

정도인 것으로 나타난 네 평가자의 효능감을 살펴보았을 때, 5점 리커트 척도에서 효능감이 높은 수준으로 볼 수 있는 '3.5'를 기준으로 하여, A1, A6, A13 평가자는 쓰기 평가 효능감이 낮은 집단에 속하며, B8 평가자는 높은 집단에 속해 있었다. 이를 통해 효능감이 낮은 평가자들이 문제가 될 정도의 범주별 편향성을 보이는 경향이 더 있었음을 알 수 있다.[14)]

<표 8> 쓰기 평가 효능감에 따른 집단-평가 범주 간 편향성의 정도

쓰기 평가 효능감	평가자	내용 및 과제 수행	전개 구조	언어 사용
낮은 집단	A1	-1.64	3.89	-2.29
	A6	4.34	-3.3	-1.14
	A13	2.08	-1.4	-0.59
높은 집단	B8	3.65	-1.14	-2.35

쓰기 평가 효능감이 높은 집단과 낮은 집단으로 구분하였을 때 두 집단에 속하는 평가자 및 평가 범주별 엄격성의 평균값, 그리고 두 집단의 엄격성이 통계적으로 유의한 차이를 보이는가는 <표 9>를 통해 확인할 수 있다. 이를 통해 쓰기 평가 효능감 수준에 따라 통계적으로 유의미한 차이를 보인 범주는 '언어 사용' 범주였음을 알 수 있다.[15)] 이는 앞서 언급한 바 있듯이, 쓰기 평가 효능감이 낮은 평가자들이 쓰기 숙달도 및 작문 수준에 대한 이미지화가 부족하고 특히 내용 및 과제수행이나 전개구조에 비해 '언어 사용' 평가에서 보다 자신감을 보이면서 엄격한 채점을 하는 것과 연관된다.

14) 음영 표시된 부분이 특히 문제가 될 정도의 편향성을 보인 범주로 -2와 2 사이에 속한 TV 값이다.

15) 그러나 <표 8>에서 살펴보았듯이, 문제가 될 정도의 범주별 편향을 보인 평가자들은 세 범주에서 모두 나타나고 있었다.

<표 9> 쓰기 평가 효능감 수준에 따른 범주별 엄격성 검정

구분	평가자	평가자 엄격성의 평균		
		내용 및 과제 수행	전개 구조	언어 사용
쓰기 평가 효능감이 낮은 평가자 집단 (17명)	A1, A2, A3, A4, A5, A6, A7, A8, A9, A11, A12, A13, B5, B7, C6, C9, C10	-0.04	0.20	0.55
쓰기 평가 효능감이 높은 평가자 집단 (22명)	A10, B1, B2, B3, B4, B6, B8, B9, B10, B11, B12, B13, C1, C2, C3, C4, C5, C7, C11, C12, C13	-0.04	-0.16	0.44
유의 확률		.777	.453	.039*

P < .05 수준에서 통계적으로 유의미함.

이러한 평가 범주에 따른 엄격성의 편향과 쓰기 평가 효능감은 어떤 관련을 맺고 있는지 문항별로 구체적으로 살펴보도록 하겠다. '내용 및 과제 수행' 범주에 대한 엄격성, 변별력의 차이는 4.1장에서 교사 효능감이 '일반적 한국어 쓰기 평가 효능감'에서 A집단이 낮은 효능감을 보인 것과 관련이 있다. 쓰기 평가 효능감 검사지의 '나는 학생들의 글만으로도 학생의 쓰기 숙달도(초급·중급·고급)를 판단할 수 있다(4번)'와 '나는 학생 글을 학생이 속한 쓰기 숙달도 내에서 상·중·하로 잘 변별하여 평가할 수 있다(5번)'에 대한 응답 평균에서 C집단은 0.1이라는 매우 근소한 차이를 보인 반면, B집단은 0.3, A집단은 0.4로 보다 큰 차이를 보인 것과 연관될 수 있는 것이다. 이는 평가를 위해 작문을 접했을 때, 초보 평가자의 경우 이 작문의 언어적 숙달도가 초·중·고급 중 어느 집단에 속하는지를 판단하는 것보다 해당 숙달도 안에서 잘 쓴 글과 못 쓴 글, 즉 상·중·하 수준으로 변별하는 데에 더 어려움을 느끼고 있다는 것을 보여주기 때문이다. 이는 A집단과 같이 경력이 적은 평가자들을 대상으로 평가 교육을 할 때, 같은 언어 숙달도 내 작문의 상·중·하 수준을 구

별해 낼 수 있도록 하는 교육이 이루어져야 보다 효과적으로 쓰기 평가 효능감을 높일 수 있음을 시사한다.

‘언어 사용’ 범주에 대한 엄격성, 변별력의 차이는 4.1장의 쓰기 평가 효능감을 살핀 부분에서 A집단이 ‘나는 학생 글에서 언어사용(어휘와 문법 등)의 정확성을 잘 평가할 수 있다(10번)’라는 문항과 ‘나는 학생의 글에서 글의 목적과 기능에 따른 격식에 맞춰 글을 썼는지 잘 평가할 수 있다(12번)’ 문항에 비교적 높은 효능감을 보인 것과 연관된다. A집단은 내용 및 과제 수행 범주, 전개 구조를 평가하는 것에 비해 언어 사용 범주를 평가하는 것에 부담을 덜 느끼며, 특히 언어 사용의 다양성보다는 정확성이나 사회언어학적 기능과 같은 보다 객관적인 평가 범주에 자신감을 보이고 있었다. 이는 평가 범주 사용 역량과 맞물려 언어 사용을 다른 범주에 비해 까다롭게 평가하게 하는 원인으로 작용할 수 있다.

지금까지 살펴본 바와 같이 평가자 간에는 엄격성의 차이가 존재하였다. 총점의 집단 간 엄격성의 차이는 없었으나, 평가 범주별 엄격성에 있어서는 집단 간에 뚜렷한 차이가 있었다. 또한 이러한 평가 범주별 엄격성의 차이는 쓰기 평가 효능감과 연관을 맺고 있음을 알 수 있었다.

평가자가 자신이 엄격한 평가자인지, 관대한 평가자인지를 아는 것은 평가 전략의 활용을 돕고 자신을 통제할 수 있다는 느낌을 주어 평가 효능감을 높여줄 수 있다. 평가자의 평가과정을 사고구술로 살핀 김동수(2014:68)에서는 숙련된 평가자가 자신이 관대한 평가자임을 인지하고 있어, 평가자 간 신뢰도를 높이기 위한 전략으로 자신의 엄격성을 조절하는 사례를 보여준 바 있다. 고전적 통계 방법으로 신뢰도를 분석할 경우 평가자 간의 일치 정도는 알 수 있지만 그것이 어디에서 기인하는지, 어느 정도로 엄격성을 조절해야 하는지, 자신의 엄격성은 어느 정도인지 알기 어렵다. 반면 다국면 라쉬 모형을 활용하여 평가 특성을 분석할 경우, 평가자 간 엄격성의 정도를 알 수 있게 된다. 이는 자신의 평가 경향을 파악하는 데 도움을 주고, 이처럼 자신의 평가 특성에 대해 스스로 ‘알고 있

다'는 느낌은 평가 효능감을 높이는 데에도 긍정적 영향을 미친다.

4.2.2. 한국어 쓰기 평가의 내적 일관성과 효능감

여기에서는 평가의 내적 일관성을 중심으로 평가 특성을 살펴보도록 하겠다. 본 연구에서는 적합도(fit)가 0.50~1.50 범위에 위치하면 내적 일관성이 적합한 것으로 설정하였다.[16] 적합도 지수가 0.5보다 작으면 '과적합(overfit)', 1.5보다 크면 '부적합(misfit)'으로 볼 수 있다.

과적합은 평가 결과가 측정 모형에 지나치게 적합한 것으로서, 모든 수험자에 대한 평가 결과가 지나치게 일관되어 관찰 값이 충분히 예측 가능할 때 나타난다. 이런 경우 평가자가 중앙값에 편중된 평가를 하였을 가능성이 높다. 반면 부적합 지수는 평가자들이 평가 범주를 일관성 없이 적용하여 기대 이상으로 편차가 클 때 발생한다. 평가의 내적 일관성 측면에서 볼 때는 과적합보다는 부적합 지수가 더 문제시된다 (Wiseman, 2008; 김성숙, 2011).

본 연구에서는 <표 10>과 같이 39명의 평가자 중에서 30명(77%)이 내적합도와 외적합도에서 모두 0.50~1.50 사이의 값을 보여 적합 일관성을 보였고, 6명(15%)이 부적합 일관성을, 3명(8%)이 과적합 일관성을 보였다. 집단별로 살펴보면, A집단에서는 평가자 A1, A5, A6, A11이 부적합 일관성을 보였으며, 평가자 A2, A7이 과적합 일관성을 보였다. B집단에서는 과적합 일관성을 보인 평가자는 없었고, B2, B11 평가자가 부적합 일관성을 보였다. C집단에서는 부적합 일관성을 보인 평가자는 없었고, C7만이 과적합 일관성을 보였다. A집단은 13명 중 6명(46.15%), B집단

16) 적합도 지수의 상하한 값은 평가의 특성과 편차의 정도에 따라 달리 설정할 수 있다. McNamara(1996)에서는 0.75~1.30의 좁은 범주를 제안하였고 Linacre(1989, 2010)에서는 0.50~1.50의 보다 넓은 범주를 제안한 바 있다. 김성숙(2011), 이영식(2014)에서는 사례수가 200 이하일 때 권장되는 적합도 지수로, z값을 기준으로 분석하는 -2.00<z<+2.00의 범주를 채택한 바 있다.

은 2명(15.38%), C집단은 1명(7.69%)이 일관성에 문제를 보였다. 부적합 일관성을 보인 A, B집단의 평가자 6명은 평가의 일관성, 즉 평가자 내 신뢰도가 낮다고 볼 수 있는 것이며, 과적합 일관성을 보인 A, C집단의 평가자들은 평가한 작문에서 점수 차이가 거의 없어 변별력을 갖지 못했음을 뜻한다.

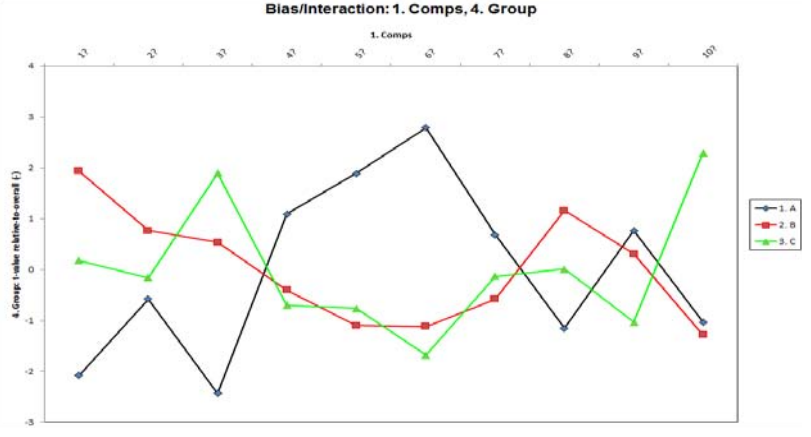
A집단의 경우 50%에 가까운 평가자가 내적 일관성의 적합도에 문제를 보였으나, B, C집단으로 갈수록 내적 일관성에 문제를 보이는 평가자의 비중이 줄어드는 양상을 보였다. 이를 경력이 많을수록 일관성의 적합도 정도가 향상되는 것으로 볼 수도 있으나, 경력이 많은 집단에서도 여전히 적합도에 문제를 보이는 평가자가 있다는 점에서, 경력이 많은 집단 역시 평가자 교육이 필요함을 알 수 있다.

<표 10> 내적 일관성과 적합도 판정

평가자	내적합[17]	외적합[18]	적합도 판정	평가자	내적합	외적합	적합도 판정	평가자	내적합	외적합	적합도 판정
A1	1.82	1.79	부적합	B1	0.98	0.97	적합	C1	0.64	0.64	적합
A2	0.47	0.48	과적합	B2	1.95	1.97	부적합	C2	0.80	0.79	적합
A3	0.76	0.75	적합	B3	1.16	1.14	적합	C3	0.99	1.02	적합
A4	0.70	0.70	적합	B4	0.62	0.73	적합	C4	0.94	0.92	적합
A5	1.93	1.89	부적합	B5	1.12	1.15	적합	C5	0.96	0.94	적합
A6	1.69	1.70	부적합	B6	0.83	0.81	적합	C6	1.21	1.19	적합
A7	0.40	0.41	과적합	B7	1.24	1.24	적합	C7	0.45	0.46	과적합
A8	1.05	1.09	적합	B8	1.17	1.15	적합	C8	0.54	0.54	적합
A9	0.62	0.65	적합	B9	0.52	0.55	적합	C9	1.14	1.12	적합
A10	0.92	0.95	적합	B10	0.82	0.82	적합	C10	1.10	1.08	적합
A11	1.65	1.60	부적합	B11	1.57	1.61	부적합	C11	0.73	0.72	적합
A12	0.82	0.84	적합	B12	0.60	0.61	적합	C12	1.37	1.38	적합
A13	0.97	1.05	적합	B13	0.58	0.60	적합	C13	0.60	0.58	적합

17) 내적 적합도 제곱평균(infit mean square)을 뜻한다.
18) 외적 적합도 제곱평균(outfit mean square)을 뜻한다.

<그림 5>는 작문과 집단 간의 편향성을 분석한 그래프이다. 4.2.1장에서 살펴보았듯이, 특정 항목에 대한 엄격성의 편향성을 나타내는 그래프 중에서 TV(t-value relative-to-all)는 기대 점수와 관찰 점수 간의 차이를 보여주는 그래프로, t 값이 -2~+2 범위 밖에 있으면 그 편향성이 통계적으로 유의미하여 문제가 있는 평가라는 것을 의미한다. <그림 5>를 보면 A집단은 작문 10편 중 세 편의 작문에 대해 과하게 엄격하거나(1, 3번 작문) 관대하게(6번 작문) 평가했음을 알 수 있다. 이러한 작문에 따른 엄격성의 편향성이 A집단의 내적 일관성 적합도의 문제와 연관을 맺고 있을 것으로 보인다.



<그림 5> 작문-평가자 집단 간 편향성 분석 그래프

본 연구에서는 이러한 내적 일관성이라는 평가 특성이 쓰기 평가 효능감과 상관을 갖고 있는지 알아보고자, 평가자 39명 전체를 일관성 '적합' 집단과 '과/부적합' 집단으로 나누고, 쓰기 평가 효능감에 차이가 있는지 Mann-Whitney 검정을 실시하였다. 검정 결과, <표 11>에서 확인할 수 있듯이 '일반적 한국어 쓰기 평가 효능감'과 '실행적 한국어 쓰기 평가 효능

감' 모두 통계적으로 유의미한 차이를 보여, 쓰기 평가 효능감과 내적 일관성의 상관이 있음을 확인하였다. 즉, 일관성이 적합한 집단의 쓰기 평가 효능감이, 일관성이 적합하지 못한 집단의 쓰기 평가 효능감보다 높은 평균을 보이고, 통계적으로도 유의미한 차이를 보였다는 것은 쓰기 평가 효능감과 평가 양상이 상관관계를 갖고 있음을 보여주는 것이다.

<표 11> '적합' - '과/부적합'의 한국어 쓰기 평가 효능감 차이(전체)

구분	일반적 한국어 쓰기 평가 효능감		실행적 한국어 쓰기 평가 효능감	
	평균	유의확률(P)	평균	유의확률(P)
적합(30명)	3.58	.019*	3.75	.037*
과/부적합(9명)	2.74		3.29	

P < .05 수준에서 통계적으로 유의미함.

이와 함께 동일 집단 내에서 내적 일관성에 따른 효능감 역시 차이를 보이는지도 분석하였다. 즉, 2년 미만이라는 비슷한 경력을 가진 A집단 내에서 적합한 일관성을 보인 7명과 부적합 또는 과적합 일관성을 보인 6명의 한국어 쓰기 평가 효능감이 통계적으로 유의한 차이를 보이는지를 살폈다.

이를 위해 A집단 내에서 적합한 일관성을 보인 집단은 '적합(A)', 부적합 또는 과적합 일관성을 보인 집단은 '과/부적합(A)' 집단으로 설정하고, Mann-Whitney 검정을 통해 평가의 일관성 경향에 따라 쓰기 평가 효능감의 차이가 통계적으로 유의한지 알아보았다. 그 결과, <표 12>와 같이 '실행적 한국어 쓰기 평가 효능감'은 통계적으로 유의미한 차이를 보이지 않았으나, '일반적 한국어 쓰기 평가 효능감'은 통계적으로 유의미한 차이를 보여 쓰기 평가 효능감과 평가 일관성의 상관이 일부 확인되었다.

<표 12> ‘적합’ - ‘과/부적합’의 한국어 쓰기 평가 효능감 차이(집단 A)

구분	일반적 한국어 쓰기 평가 효능감		실행적 한국어 쓰기 평가 효능감	
	평균	유의확률(P)	평균	유의확률(P)
과/부적합(A) (6명)	2.22	.038*	3.05	.387
적합(A) (7명)	2.89		3.20	

P < .05 수준에서 통계적으로 유의미함.

효능감은 자신감과 밀접한 연관을 맺는 만큼, 평가에 대한 효능감이 낮을 경우 높은 점수나 낮은 점수를 주기를 두려워하는 중앙집중화 현상으로 이어져 과적합 양상을 보이게 하거나, 자신의 이전 평가에 대한 의심을 하며, 다른 엄격성을 적용하여 일관성 없는 평가를 하는 부적합 일관성으로 나타날 수 있다.

자신감 감소는 평가자의 점수 결정에서의 혼란을 불러올 것이고, 이는 평가자의 일관성 하락을 초래하기 쉽다. 또한 쓰기 평가 과정은 읽은 글을 자세하게 회상해내는 능력과 함께 글에 대한 분석, 판단 등 높은 인지적 부담을 요구하는데 확신을 갖지 못하고 고민하는 과정은 평가에 대한 집중력을 저해하는 결과를 가져올 수 있다(박종임, 2013:89). 따라서 이러한 내적 일관성의 향상이 필요한 평가자의 경우, 쓰기 평가 훈련과 함께 평가 효능감을 높이는 프로그램을 진행하여 자신의 평가에 대해 보다 확신을 갖도록 도울 필요가 있다.

5. 결론

본 연구는 한국어 쓰기 평가 연구를 위한 기초 정보를 제공하고자 경력에 따른 한국어 교사의 쓰기 평가 효능감을 알아보고 다국면 라쉬 모형을 활용하여 경력에 따른 평가의 특성 또한 분석하였다.

세 집단의 쓰기 평가 효능감은 경력 2년 미만의 집단(A)과 2년 이상의 집단(A, B)이 통계적으로 유의한 차이를 보였다. 이중 실행적 쓰기 평가 효능감에 비해 일반적 쓰기 평가 효능감에서 큰 차이를 보였다. 또한 세 집단은 평가 범주 사용의 엄격성에서 차이를 보였으며, 이는 쓰기 평가 효능감과 연관을 맺고 있었다. 특히, 쓰기 평가 효능감이 높은 집단과 낮은 집단의 엄격성의 편향 정도는 '내용 및 과제 수행', '전개 구조' 범주에서는 통계적으로 유의미한 차이가 없었으나, '언어 사용' 범주에 있어서는 통계적으로 유의미한 차이를 보였다. 또한 일관성 적합도가 적합한 집단과 그렇지 못한 집단의 쓰기 평가 효능감 역시 통계적으로 유의한 차이를 보였다. 이를 통해 평가 효능감과 평가 능력은 상호작용하므로, 앞으로 한국어 교사의 쓰기 평가 효능감과 평가 특성을 고려한 평가자 교육이 이루어질 필요가 있음을 확인할 수 있었다.

본 연구는 소수의 평가자를 대상으로 하였고 평가한 작문의 장르 및 과제, 숙달도 역시 제한적이었기에 연구의 결과를 일반화하는 데에는 한계가 있다. 그러나 한국어교육에서 본 연구와 같이 쓰기 평가 효능감과 평가 특성을 살핀 연구가 없었다는 점, 교사 대상 평가자 교육에서 이를 함께 고려할 수 있다는 점에서 의의를 찾을 수 있을 것이다.

<참고 문헌>

가은아(2008). 국어교사의 쓰기 효능감 및 쓰기 신념이 학생 쓰기 평가에 미치는 영향, 한국교원대학교 석사학위논문.

강석한·안현기(2014). 외국인 한국어 말하기 시험의 평가자 요소가 채점에 미치는 영향, <이중언어학> 55호, 이중언어학회. 1쪽～29쪽.

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

김성숙(2011). 학문 목적 한국어 쓰기 능력에 대한 분석적 평가의 일반화가능도 검증, <한국어교육> 22-3호, 국제한국어교육학회. 81쪽～99쪽.

김성숙·유혜령(2013). 다국면 라쉬 모형을 적용한 논술문 평가 사례 보고, <대학작문> 6호, 대학작문학회. 133쪽～157쪽.

김정숙(2010). 한국어 쓰기 능력 평가 방안-종합적 채점과 분석적 채점 결과를 중심으로, <이중언어학> 43호, 이중언어학회. 493쪽～523쪽.

박영민(2010). 예비국어교사의 쓰기 평가 효능감 분석, <청람어문교육> 42호, 청람어문교육학회. 181쪽～207쪽.

박영민(2011). 국어교사의 쓰기 평가 효능감 분석, <청람어문교육> 44호, 청람어문교육학회. 121쪽～146쪽.

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

박영민·최숙기(2010). Rasch 모형을 활용한 국어교사의 쓰기 평가 특성 분석-중학생 설명문 쓰기 평가를 중심으로, <국어교육학연구> 37호, 국어교육학회. 367쪽～391쪽.

박종임(2013). 국어교사의 쓰기 평가 특성 연구. 한국교원대학교 박사학위논문.

박종임·박영민(2011). Rasch 모형을 활용한 국어교사의 채점 일관성 변화 양상 및 원인 분석-중학생 서사문 채점을 중심으로, <우리어문연구> 39호, 우리어문학회. 301쪽～335쪽.

서수현(2012). 쓰기 평가 협의 과정에 나타난 쓰기 평가자의 인식 연구, <國語敎育學硏究> 44호, 국어교육학회. 335쪽～367쪽.

신동일(2001). 채점 경향 분석을 위한 Rasch 측정모형 적용연구, <Foreign Language Education> 8-1호, 한국외국어교육학회. 249쪽～272쪽.

신동일(2002). Rasch 모형을 이용한 고등학교 영어과 말하기 및 쓰기 능력 등급 기술표 개발, <영어교육> 57-4호, 한국영어교육학회. 469쪽～499쪽.

안정민(2011). 한국어 교사 교육 능력 향상을 위한 교사 효능감 연구, 한국외국어대학교 석사학위논문.

안정민·김재욱(2011). 한국어 교사의 교사 효능감 연구, <이중언어학> 45호, 이중언어학회. 109쪽～132쪽.

이영식(1998). 영어작문평가의 채점신뢰도에 대한 분석, <영어교육> 53-1호, 한국영어교육학회. 179쪽～200쪽.

이영식(2014). 다국면 Rasch 측정의 Facets 프로그램을 활용한 영어 작문 평가의 원어민 채점 검증, <영어어문교육> 20-1호, 한국영어어문교육학회. 475쪽～496쪽.

이학식·임지훈(2011).『SPSS 18.0 매뉴얼』. 서울: 집현재.

이향(2012). 한국어 말하기 수행평가의 발음 범주 채점 방식에 따른 채점 신뢰도 분석-다국면 라쉬 모형을 활용하여, <외국어로서의 한국어교육> 37호, 연세대학교 언어연구교육원 한국어학당. 325쪽～355쪽.

이향(2013). 한국어 말하기 평가의 발음 영역 채점에서의 채점자 특성에 따른 채점 경향 연구-한국어 교육 경험과 전공을 중심으로, <외국어로서의 한국어교육> 39호, 연세대학교 언어연구교육원 한국어학당. 213쪽～245쪽.

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

정지은(2012). 한국어 교사의 교사효능감에 관한 연구 - 교사의 배경 변인과 직무환경의 영향을 중심으로. 연세대학교 교육대학원 석사학위논문.

최인철(1998). Test of Oral proficiency(TOP)의 개발 연구, <어학연구> 34-1호, 서울대학교 어학연구소. 245쪽～289쪽.

최인철(2000). 영어 의사소통능력의 모의 구술 면접시험 방식 양상 타당성 검증, <응용영어학> 16호, 응용언어학회. 215쪽～246쪽.

Bandura, A.(1986). The explanatory and predictive scope of self-efficacy theory. *Journal of Social and Clinical Psychology*, 4, 359-373.

Bandura, A.(1997). *Self-Efficacy: The Exercise of Control*. New York: W.H Freeman. 김의철 외 역.(1999). 자기효능감과 인간행동. 서울: 교육과학사.

Bernardin, H. J., & Villanova, P.(2005). Research streams in rater self-efficacy, *Group & Organization Management*, 30(1), 61-88.

Block, J. H.(1978). Standards and criteria: A response. Journal of Educational Measurement, 15. 291-295.

Lavelle, E.(1996). Generalizability theory and many-facet Rasch measurement. In Engleland, G & Wilson, M. (Eds.), *Objective measurement: Theory into Practice*, 3. 85-98. Norwood, NJ: Ablex.

Lavelle, E.(2006). Teachers' self-efficacy for writing. *Electronic Journal of Research in Educational Psychology*, 4(1), 73-84.

Linacre, J. M.(2014). *A user's guide to FACETS: Rasch-Model Computer Programs. Program Manual 3.71.4.*

Linacre, J. M.(2014). *FACETS: Rasch measurement computer program(Version No. 3.71.4),* [Computer software]. Chicago, IL: winstepss.com

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

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

Popham, W. J.(1978). *Criterion-referenced Measurement*. Englewood Cliffs, NJ: Prentice-Hall, INC.

Tziner, A.(1999). The relationship between distal and proximal factors and the use of political considerations in performance appraisal. *Journal of Business &*

Psychology, 14, 217-231.
Tziner, A., & Murphy, K. R.(1999). Additional evidence of attitudinal influences in performance appraisal, *Journal of Business and Psychology*, 13(3), 407-419.
Weigle, S. C.(1998). Using FACETS to model rater effects. *Language Testing*, 15, 263-287.

이인혜(Lee Inhye)
고려대학교 대학원 국어국문학과
136-701 서울시 성북구 안암동 1가
전화번호: 1-612-961-6213
전자우편: heyday0817@korea.ac.kr

접수일자: 2014년 7월 31일
심사(수정)일자: 2014년 8월 20일
게재확정: 2014년 9월 17일