한국어 말하기 평가에서 원어민과 비원어민 채점자의 채점 경향 비교

김지영†
이화여자대학교 언어교육원

Comparison of Rating Tendencies between Native and Non-native Speakers in Korean Speaking Test

Jee Young Kim†
Ewha Language Center

ABSTRACT

The purpose of this study was to compare the rating tendencies of Korean and Chinese raters in a Korean Speaking Test. For this purpose, graduate students majoring in Korean education were trained, and an individual rating process was performed. The results of the rating were then analyzed using the multi-faceted Rasch model, focusing on rating consistency, severity, and bias. The results of the analysis indicate that rating severity differed among raters even in the same group, and rating consistency was either an overfit or a misfit in the Chinese rating (CR) group. The results also showed that the CR group tended to be more tolerant of assessment evaluation items with less difficulty than the Korean rating (KR) group, and to score assessment evaluation items with higher difficulty more strictly. In the analysis of criteria, the KR group scored more strictly than the CR group on the three criteria except vocabulary and grammar, and organization. In contrast, there was a statistically significant tendency toward the opposite for "organization". The two groups differed in judging the proficiency of test takers, with differences between the test takers in cases where proficiency is relatively high and when specific languages are reflected in the test taker's pronunciation or intonation, some of which were analyzed as significant bias. Finally, the use of the evaluation scale showed that the CR group had a wider distribution, but the reliability of using the zero scale was low, and the difference between the mean scores of each scale was not uniform. This group also tended toward the middle point rather than toward the peak.

Keywords: Korean speaking test, Korean rater, Chinese rater, rating tendency

† Corresponding author: gb9802@hanmail.net

https://doi.org/10.30961/lr.2019.55.1.179

1. 서 론

본 연구의 목적은 한국어 말하기 평가에서 한국인 채점자와 중국인 채점자의 채점 경향을 비교하여 원어민 채점자와 비원어민 채점자의 채점특성에 차이가 있는지 알아보고자 하는 데에 있다. 이를 위해 한국어교육을 전공하는 대학원생을 대상으로 채점자 훈련을 실시하고 중·고급 수준 한국어 학습자의 음성 파일에 대한 개별 채점을 진행한다. 그리고 그 결과를 채점 엄격성과 일관성, 그리고 평가문항과 평가구인, 평가척도에 대한 엄격성 및 편향성 등을 중심으로 분석해 보고자 한다.

수행평가에서 채점은 평가의 신뢰도와 직결된다. 신뢰도는 일관된 평가 결과가 도출되어야 확보될 수 있으므로 채점자가 채점 기준을 얼마나 일정하게 적용하는가의 문제가 영향을 미칠 수 있기 때문이다. 그리고 교육 현장에서 학습자의 수행평가는 담당 교사에 의해 이루어지는데 최근 국내외에서 한국어교육을 전공하는 외국인의 수는 꾸준히 증가하는 추세이다. 이는 곧 자격을 갖춘 비원어민 한국어교원이 교육 현장에 배치되어 학습자들의 수행평가를 담당할 수 있다는 것을 의미한다. 여기서 비원어민 평가자의 평가 결과를 신뢰할 수 있는가 하는 문제가 발생할 수 있다.

외국어 수행평가에서 원어민과 비원어민 채점자의 채점 결과를 비교・분석한 연구들은 결과가 동일하지는 않으나 두 집단 간에는 분명한 차이가 보고되었다. 이는 양적인 분석을 통해서 전반적인 채점 엄격성에서 나타나기도 하였는데, 주로 평가구인에 대한 채점에서 서로 다른 경향을 보였고(Shi 2001; Kim 2009; Yu 2010; Lee and Chae 2012; Kang and Ahn, 2012) 평가문항에 따라 다르게 나타나기도 하였으며(Kim 2011), 채점자에게 익숙한 언어를 모국어로 사용하는 수험자에 따라 차이를 보이기도 하였다(Carey et al. 2011; Winke et al. 2012). 아울러 질적인 분석을 통해서는 쓰기 평가의 경우 원어민 채점자는 서양의 작문 관례에 기반을 두는 반면, 한국인 채점자의 경우 한국의 수사학적 관례를 적용하여 채점하고(Lee and Chae 2012), 말하기 평가에서는 발화 태도나 몸짓에 대해 원어민 채점자와 중국인 채점자가 서로 다른 프레임을 적용하는 경향(Gui 2012)이 관찰되기도 하였다.

한국어교육에서 채점자 관련 연구는 한국어교원과 일반인(강석한, 안현기 2014), 세부 전공이 다른 한국인 대학원생(이향 2013), 한국인 대학원생과 비원어민 대학원생(Kim 2016; 원미진, 김지영 2017) 등 주로 한국인을 대상으로 연구되어 왔다. 비원어민 대학원생을 대상으로 한 연구의 경우는 특정 언어권이 아닌 한국어를 모국어로 하지 않는 외국인을 대상으로 하여 채점 결과의 적합성 여부를 확인하였다. 최근 비원어민 한국어교원이 증가하고 있고, 국외 한국어교육의 외연도 확장되고 있으므로 보다 경험적으로 특정 언어권의 비원어민 채점자의 채점 특성을 파악하여 신뢰할 만한 비원어민 채점자 양성에 대비할 필요가 있다. 이에 본 연구에서는 한국어교육을 전공하는 중국인 대학원생을 대상으로 이들과 한국인 채점자들의 채점 결과에 대한 양적인 분석을 통해 두 집단의 채점 엄격성과 일관성, 평가문항 및 평가구인에 대한 엄격성과 편향성, 평가

척도의 사용 양상과 적합성 등을 비교하여 집단 간 공통점과 차이점을 살펴보고자 한다.

2. 선행연구

외국어 말하기 평가에서 다국면 라쉬모형을 적용하여 원어민과 비원어민의 채점 경향을 분석한 연구를 살펴보면, 신동일(2001)은 MATE 말하기 평가에서 원어민 채점자 3명과 한국인 채점자 3명의 채점 적합도를 분석하였는데, 두 집단 모두에서 과적합하거나 부적합한 경향을 보이는 채점자들이 도출되었다. 그리고 백현영, 양병곤(2011)은 중학교 한국인 영어교사 5명과 원어민 보조교사 3명의 채점 경향을 비교하였다. 평가문항은 서로 다른 유형의 8개 문항으로 구성하였고, 평가구인은 정확성, 이해, 유창성, 발음 등 4개에 대하여 1~7점 척도를 사용하였다. 채점 결과, 원어민 보조교사들은 모두 적합한 범위 내의 채점 경향을 보였으나, 한국인 영어교사들 중에서는 부적합하거나 과적합한 경향이 발견되었다. Kang and Ahn(2012)에서는 원어민 채점자와 한국인 채점자 각각 5명의 채점 경향을 분석하였는데 채점 적합도 분석에서 집단별로 1명의 채점자에게서 부적합 경향이 분석되었다. 전반적인 엄격성은 원어민 채점자들이 다소 높게 나타났으나 평가구인에 대한 엄격성에서는 한국인 채점자는 문법을 엄격하게 채점하고 원어민 채점자는 담화 응집성을 엄격하게 채점하는 등의 차이를 보였다. 평가과제에 대한 엄격성에서도 의견제시 문항에 대해서 한국인 채점자는 엄격하게, 원어민 채점자는 상대적으로 관대하게 채점한 것으로 나타났다.

Carey et al.(2011)과 Winke et al.(2012)은 외국어 말하기 평가에서 채점자 효과를 분석하였다. Carey et al.(2011)은 발음에 대한 채점은 평가자가 비원어민 영어 악센트에 얼마나 많이 노출되었는가에 따라 영향을 받기 쉽다는 가설을 세우고, 말하기 평가에서 발음 요소가 채점자 간 신뢰도에 영향을 줄 수 있는가를 연구했다. 채점대상은 중국어, 한국어, 인도 영어를 모국어로 하는 영어 학습자의 음성 자료였고, 채점은 5개 지역의 IELTS 채점자 99명이 담당했다. 채점자들은 수험자의 중간언어에 장시간 노출되었거나, 그렇지 않거나, 거의 노출되지 않은 상태였다. 채점 결과, 수험자의 중간언어에 장기간 노출된 채점자들이 발음 구인에서 높은 점수를 주고, 그렇지 않은 다른 두 부류의 채점자들은 낮은 점수를 준 것으로 나타났다. Winke et al.(2012)는 수험자의 발음과 억양에 대한 채점자의 친숙함(accent familiarity)이 채점 편향을 유발하는가를 조사하였다. 이 연구에서는 ETS에서 스페인어, 한국어, 중국어(보통화) 각각을 모국어로 하는 수험자 24명, 총 72명의 iBT TOEFL 음성 파일을 제공 받았고 127명의 채점자들이 채점에 참여했다. 결과는 제2언어로 스페인어, 한국어, 중국어(보통화) 등을 학습한 100명의 채점자들을 중심으로 분석되었는데, *t*-검정과 편향분석 결과, 제2언어로 스페인어와 중국어를 학습한 채점자들이 해당 언어를 모국어로 하는 수험자에게 보다 관대한 점수를 부여한 것으로 나타났으며, 이는 통계적으로 유의했다.

외국어 말하기 평가에서 다국면 라쉬모형이 아닌 분산분석을 통해 집단 간 차이를 분석한 연구를 살펴보면, Gui(2012)는 영어 원어민 채점자들과 중국인 채점자들이 말하기 대회에 참여한 참가자들을 평가할 때의 차이점을 비교하였다. 총점에서 두 집단 사이에 유의미한 차이는 없었으나, 평가구인에 대한 채점 점수는 유의미한 차이를 보였다. 그리고 채점 일관성 분석 결과는 원어민 채점자 집단의 신뢰도 지수가 높았고, 채점 점수의 범위도 더 넓은 것으로 나타났다. 또한 두 전달력에 대한 채점자들의 코멘트와 사후 인터뷰에서 중국인 채점자들은 참가자들이 대체로 편안한 태도를 보였고 청중들과 적절히 눈을 맞추며 발표했다고 긍정적으로 응답한 반면, 원어민 채점자들은 대부분의 참가자들이 긴장하는 모습을 보였고 손이 떨렸으며 준비한 노트를 자주 참고했고 준비한 내용을 로봇처럼 발화했다고 부정적으로 평가했음을 언급하며, 이는 구두 수행에 대한 미국과 중국의 문화적인 프레임의 차이 때문으로 볼 수 있다고 주장했다. Yu(2010)는 영어 원어민 채점자들과 한국인 채점자들이 ESPT 샘플 120개를 정확성, 유창성, 발음, 어휘 등 4개의 구인과 6단계 척도로 채점한 결과를 분석했다. 각 집단의 크롬바흐 알파 계수는 .94~.97로 채점 일관성이 높게 나타났으며, 집단 내 채점자 간 일치도를 확인하기 위한 피어슨 적률 상관계수 평균은 원어민 채점자 집단이 비원어민 채점자 집단보다 약간 높았다. 평가구인에 대한 엄격성은 비원어민 채점자들이 더 엄격한 경향을 보였는데 두 집단의 평균 차이는 숙달도가 낮을수록 높았으며 이는 통계적으로도 유의했다.

한국어 말하기 평가에서 한국인과 외국인 채점자의 채점 경향을 분석한 연구를 살펴보면, Kim(2016)은 한국어교육전공 대학원생을 대상으로 채점자 훈련을 실시하고, 원어민 채점자와 비원어민 채점자의 채점 경향을 훈련 전과 후로 나누어 비교했다. 총괄적 채점과 분석적 채점을 병행하였으며 총괄적 채점은 11점 척도를 사용하였고, 분석적 채점의 평가구인은 정확성, 범위, 유창성, 내용, 조직 등 5개였다. 채점자 훈련을 실시하기 전 채점 결과는 집단 간 엄격성 차이가 크고 통계적으로도 유의미했으며 유창성을 제외한 모든 구인에서 원어민 집단이 비원어민 집단보다 더 엄격한 경향을 보였다. 그러나 채점자 훈련 후에는 반대로 비원어민 집단의 엄격성이 더 높게 나타났으나 통계적으로 유의미하지 않았다. 원미진, 김지영(2017)에서도 한국어교육을 전공하는 대학원생을 대상으로 채점자 훈련을 실시하고 수험자 30명에 대한 채점을 진행한 뒤 대학원생들을 한국어교육 유경험자 집단, 한국어교육 무경험자 집단, 비원어민 집단으로 나누어 채점 엄격성과 적합성 등을 비교하였다. 채점 엄격성은 한국어교육 유경험자 집단이 상대적으로 더 엄격한 것으로 분석되었으며 채점 일관성은 모든 집단에서 부적합・과적합 경향이 나타난 채점자들이 발견되었다. 평가문항에 대한 엄격성은 한국인 유경험자 집단과 외국인 집단이 난이도가 높아질수록 엄격해지는 경향을 보인 반면, 한국인 무경험자 집단은 반대의 경향을 나타냈다. 평가구인에 대한 엄격성은 전반적 능력과 조직은 세 집단이 비슷한 엄격성을 보였으나 어휘 및 문법, 발음, 내용 구인에 대해서는 집단 간 차이가 컸다. 그리고 평가척도 사용에 있어서는 한국어교육 유경험자 집단과 무경험자 집단은 척도 간의 거리는 일정하지 않지만 척도가 높아질수록 요구되는 측정치가 증가

하는 경향을 보였으나 비원어민 집단은 척도에 따라 측정치가 일정하게 증가하지 않았고 척도 간의 차이도 다른 두 집단에 비해 조금 더 컸다.

이와 같은 선행연구를 통해 채점 일관성이나 엄격성은 연구에 따라 일관된 결과가 나타나지 않았으나 평가문항이나 평가구인, 그리고 평가척도와 관련해서는 원어민 채점자와 비원어민 채점자 사이에 유의미한 차이가 있다는 것을 알 수 있다. 또한 수험자의 모국어나 중간언어에 대한 채점자의 친숙함, 수험자의 발화 태도 등도 채점 결과에 영향을 미칠 수 있다는 것을 보았다. 이에 본 연구에서는 한국인 채점자와 중국인 채점자의 채점 경향을 비교해 보고자 한다.

3. 연구대상 및 연구방법

3.1. 연구도구

본 연구는 한국어 말하기 숙달도 평가 개발 연구를 진행하는 과정에서 한국어교육 전공 대학원생을 대상으로 실시한 채점자 훈련 내용을 대상으로 하였다.[1] 말하기 평가는 준직접 방식으로 진행되었고 평가에 사용된 평가문항은 경험 말하기, 그림 묘사하기, 대안 제시하기, 그래프 설명하기, 의견 말하기 등 7개의 문항이며, 문항의 난이도에 따라 1, 2번은 초급 수준, 3, 4, 7번은 중급 수준, 5, 6번은 고급 수준의 문항으로 구성되었다. 박동호 외(2012)에서는 한국어능력시험 말하기 평가문항으로 따라 말하기, 단순 질문에 답하기, 경험이나 계획 말하기, 그림 묘사하기, 도표/그래프 설명하기, 의견 말하기, 설득/대안 제시하기 등의 문항 유형과 각각의 유형에 대한 변형 문항을 초·중·고급 등 등급에 따라 제시하였는데 본 연구에 사용된 7개의 문항은 이와 유사하다.

평가구인은 박동호 외(2012:79-83)에서 제시한 전반적 능력, 어휘 및 문법, 발음, 내용, 조직 등 5개를 따랐다. 전반적 능력은 수험자의 언어 사용 능력에 대한 전체적인 인상을 의미하는 것으로 수험자의 실제 발화 행위를 통해 관찰되는 전체적인 말하기 능력을 측정하기 위한 구인이다. 어휘 및 문법은 수험자가 과제수행에서 동원할 수 있는 어휘와 문법의 범위 및 정확성을 측정하기 위한 구인이며, 발음은 과제수행 과정에서 전하고자 하는 바를 효과적으로 전달하는가를 측정하기 위한 것이다. 그리고 내용은 발화 내용의 깊이와 폭을 의미하는 것으로 과제를 올바르게 수용하여 과제에 부합하게 적절한 담화를 구성할 수 있는가를 측정한다. 마지막으로 조직은 말의 시작과 전개, 마무리를 명확하게 할 수 있고, 담화 내용을 일관되고 적절하게 구성하여 표현하는가를 측정하기 위한 구인이다. 7개의 평가문항에 대해 5개의 평가구인에 따라 채점이 이루어졌으며 이때의 평가척도는 모든 문항에 0점에서 4점까지 5단계 척도를 사용하였다.

1) 본 연구는 제4차 한국어능력시험 말하기 평가 개발 연구(2017) 과정에서 채점 신뢰도를 높이기 위해 대학원생을 대상으로 수행한 채점자 워크숍의 결과를 연구진의 허락을 받아 사용하였다.

3.2. 연구대상

한국 대학에서 수학 중인 외국인 유학생의 국적은 다양한데 그중 중국인 유학생 수는 과반수 이상을 차지할 정도로 많다.[2] 한국어교육을 전공하는 학생들의 정확한 통계치를 확인할 수는 없으나, 현재 각 대학에는 매학기 다수의 중국인 유학생들이 한국어교육 전공 관련 학위를 취득하기 위해 꾸준히 진학하고 있으며, 이러한 유학생의 증가에 따라 일부 대학에서는 국어국문학이나 한국어교육을 전공한 중국인 교수를 채용하여 해당 유학생들의 한국어교육을 담당하도록 하고 있기도 하다. 뿐만 아니라 2014년을 기준으로 한국어교원 자격증을 취득한 외국인 중 81.3%인 765명이 중국 국적자이며(김가람 2016:5), 중국 내에 한국어학과가 개설된 대학은 4년제 대학 123개, 2~3년제 대학 144개 등 모두 267개로(김향란 2019:40) 이미 자격을 갖춘 중국인 교원이나 추후 학위와 자격증을 취득할 유학생들이 본국으로 돌아가 각 대학의 교원으로 채용되면 한국어 학습자들의 성취도 및 숙달도 평가에 참여하게 될 가능성이 높다. 이에 본 연구에서는 비원어민 채점자로 중국인 유학생을 선정하였다.

본 연구의 연구대상인 채점자는 모두 한국어교육을 전공하는 대학원생 10명으로 한국인 채점자 4명과 중국인 채점자 6명이다. 채점에 참여한 한국인 채점자 4명은 교육경력 2년 이상의 현직 한국어교원으로 채점에 참여할 당시에 박사과정 중이거나 박사과정을 수료한 상태였다. 그리고 중국인 채점자 중 3명은 한국에서 석사과정을 졸업한 후 박사과정에 진학하였고, 3명은 석사과정에서 수학 중이었으며, 한국어 숙달도는 박사과정의 경우 한국어능력시험 6급, 석사과정의 경우 4~5급을 받은 수준이었다.[3] 본 연구의 목적이 두 채점 집단의 채점 경향을 비교하는 데에 있으므로 분석 결과를 기술할 때 한국인 채점자는 KR, 중국인 채점자는 CR로 표기한다. 두 집단의 채점자 정보와 개별 채점자 기호는 다음과 같다.

표 1. 채점자 정보

한국인 채점자(KR)		중국인 채점자(CR)	
채점자 기호	학적	채점자 기호	학적
KR1	박사과정	CR1	박사과정
KR2	박사과정	CR2	박사과정
KR3	박사수료	CR3	박사과정
KR4	박사수료	CR4	석사과정
		CR5	석사과정
		CR6	석사과정

2) 통계청의 유학생 현황 자료(http://kosis.kr)에 따르면 2017년을 기준으로 전체 외국인 유학생의 수는 135,087명이며, 그중 중국인 유학생은 66,674명으로 전체 유학생의 68% 이상을 차지하고 있다.

3) 한국어능력시험(TOPIK)의 등급은 1급부터 6급까지 6개로 나뉘어 있으며 1, 2급은 초급, 3, 4급은 중급, 5, 6급은 고급 수준에 해당된다. 대부분의 대학부설 한국어교육기관에서도 이와 동일하게 초급인 1급부터 최고급인 6급까지의 한국어 교육 과정이 개설되어 있다.

10명의 채점자들이 점수를 부여한 수험자는 총 18명으로 남자가 3명, 여자가 15명이고, 국적은 중국 9명, 일본 2명, 브라질, 영국, 대만, 홍콩, 캐나다, 말레이시아, 한국(재일교포) 각 1명 등으로 다양했으나, 중국어권 학습자가 가장 많았다. 그리고 말하기평가에 참여할 당시 수험자들은 대부분 대학부설 한국어교육기관에서 한국어를 학습하고 있었으며, 중급인 3급과 4급이 각각 3명, 6명, 고급인 5급과 6급이 각각 3명, 6명 등으로 초급은 없고 모두 중급 이상의 학습자였다.[4] 이상의 수험자 정보는 표 2와 같다.

표 2. 수험자 정보

번호	성별	한국어 등급	국적	번호	성별	한국어 등급	국적
1	여	5	중국	10	남	3	중국
2	여	4	중국	11	여	5	중국
3	여	6	중국	12	남	4	중국
4	남	6	중국	13	여	4	한국
5	여	6	브라질	14	여	4	일본
6	여	6	영국	15	여	6	중국
7	여	4	말레이시아	16	여	3	홍콩
8	여	4	일본	17	여	6	중국
9	여	5	대만	18	여	3	캐나다

3.3. 연구절차

말하기 평가를 위한 채점자 훈련과 채점은 하루 동안 서울 소재 A 대학 컴퓨터실에서 진행되었다. 일정은 오전과 오후로 나누어 오전에는 평가문항과 평가구인, 평가척도 등 채점 요소와 기준에 대해 설명하고, 중·고급 학습자 각 2명의 샘플파일을 하나씩 듣고 채점자들이 개별적으로 점수를 매기도록 했다. 그리고 전체적으로 점수를 맞추면서 그러한 점수를 부여한 근거와 세부 기준에 대해 함께 논의하는 시간을 가졌다. 그리고 오후에는 개별적으로 수험자 18명의 샘플파일 126개를 들으면서 채점 기준에 맞춰 채점을 하도록 하였다.[5] 채점 분량은 채점자 피로도와 관련되어 채점 신뢰도에 영향을 미칠 수 있는데(김지영 2018:121), 채점자 훈련 전 연구자가 채점을 해보았을 때 수험자 1명이 7개의 문항에 응답한 음성파일을 최소 1번 이상 들어야 했으므로 한 명당 최소 10분

4) 제4차 한국어능력시험 말하기 평가 개발 연구(2017)에서는 3급 이상의 학습자를 대상으로 말하기 모의시험을 진행했다.

5) 개별 채점자들은 인터넷 채점 사이트에 접속하여 수험자의 음성 파일을 듣고 점수를 부여하였다. 채점 사이트의 음성 파일은 표 3에 제시된 수험자 정보의 순서와 같으나, 그 순서는 무작위로 배열되었다.

의 시간이 소요되었다. 이에 1시간 동안 수험자 6명의 음성 파일을 채점하고, 20분의 쉬는 시간을 갖는 방식으로 총 4시간여 동안 채점을 할 수 있도록 설계하였다.

채점 결과는 다국면 라쉬 모형을 적용하여 개발된 FACETS 프로그램(Version 3.71.4)을 사용하였다(Linacre 2014). 라쉬 모형은 문항반응이론의 한 모형으로 평가문항에 대한 피험자의 응답을 확률함수로 나타냄으로써 문항 난이도에 따라 피험자들의 점수가 달라지거나 피험자 집단에 따라 문항 난이도가 달라지는 것을 일반화하는 방법을 제시한 것이다(McNamara 1996:209-210). 그리고 FACETS은 피험자의 언어능력, 평가문항 및 평가구인의 난이도, 채점자의 엄격성 및 일관성 등에 관한 정보를 추정하여 제공해 주는 컴퓨터 프로그램으로(장소영, 신동일 2009: 12) 측정 국면의 오차, 적합도 등의 세부 정보도 상세히 제시한다는 장점이 있다(신동일, 설현수 2005: 194). 본 연구를 위한 다국면 라쉬 모형의 공식은 그림 1과 같다(Linacre 2014:13). 이 공식이 제공하는 값은 n이라는 응시자가 과제 m의 평가요소 i에 대해서 채점자 j에게 하나의 등급점수 Pnmij(k-1)에서 보다 높은 등급 점수 Pnmijk 값을 얻을 확률을 log(로그)의 값으로 변환한 것이다. 즉, 실제 채점 과정에서 받은 점수인 원점수가 아닌 보다 객관적인 변환 점수를 구하기 위해 측정치(measurement scale) 값을 로그 지수로 표기하고, 수험자 n의 능력(Bn)에서부터 과제 m의 난이도(Am)와 평가요소 i의 난이도(Di), 채점자 j의 엄격성(Cj), 그리고 어떤 등급 점수(k−1)에서 그 다음 높은 점수 k로 향상시키는 데에 걸리는 어려움(Fk)을 모두 고려한 값이다(김지영 2018:47). 본 연구에서는 이 프로그램을 이용하여 채점자 집단의 엄격성, 일관성, 편향성, 적합도 등을 분석하여 그 결과를 비교하고자 한다.

$$\log\left(\frac{\text{Pnmijk}}{\text{Pnmij(k}-1)}\right) = \text{Bn} - \text{Am} - \text{Di} - \text{Cj} - \text{Fk}$$

그림 1. 본 연구의 라쉬 모형 공식

4. 연구결과

다음의 그림 2는 FACETS 프로그램으로 분석한 전체 국면을 보여주고, 각 국면에 대한 주요 정보를 요약하여 제공하는 단면 분포도이다. 여기서 첫 번째 칸(measure)은 logit 측정치로 본 연구의 결과는 +4~−2 logit 사이에 분포하고 있다. 두 번째와 세 번째 칸은 수험자에 대한 분석 결과로 18명의 수험자들은 0~4 logit 사이에 위치해 있음을 볼 수 있다. 18명의 수험자 중 11번이 숙달도가 가장 높고, 18번이 가장 낮은 수험자로 나타났으며, 1~2 logit 사이에 가장 많은 수험자가 분포해 있다. 네 번째 칸은 채점자에 대한 분석 결과로 채점자들은 0을 중심으로 −1~1 logit 사이에 분포해 있으며, [KP4]

가 가장 엄격하고 [KP2]가 가장 관대한 채점자로 분석되었다. 그리고 다섯 번째 칸은 평가문항에 대한 분석 결과로 문항의 난이도는 −2~2 logit 사이에 분포하고 있으며, 1번이 가장 쉽고, 5번과 6번이 가장 어려운 문항이었음을 보여주고 있다. 여섯 번째 칸은 평가구인에 대한 분석 결과로 채점자와 유사하게 0 logit을 중심으로 분포해 있으며, 내용(C)이 가장 어렵고, 발음(P)과 어휘(V)가 비교적 가장 쉬운 구인으로 나타났다. 마지막의 Scale 칸은 본 연구에서 사용한 평가척도 0~4점의 사용을 보여주고 있는데, 주로 2점과 3점이 사용되었음을 알 수 있다. 이상의 내용들은 각 평가요소 분석 결과를 보면서 구체적으로 살펴보도록 하겠다.

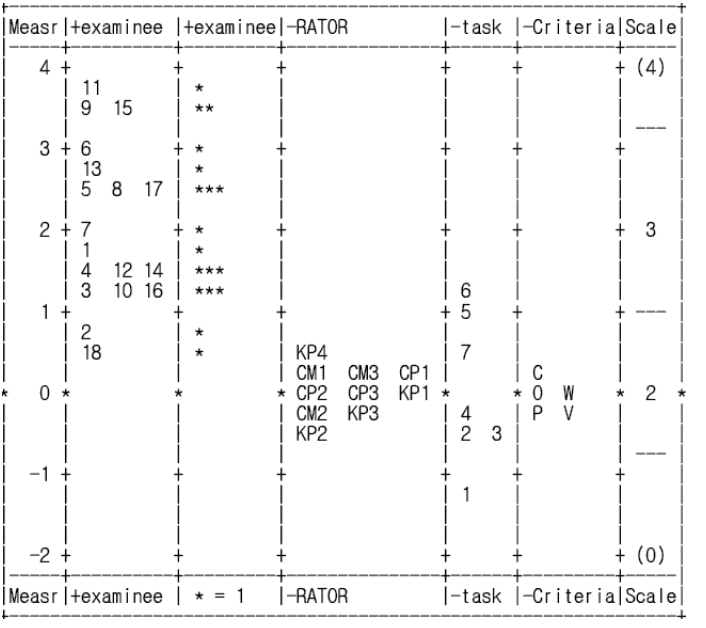

그림 2. FACETS 단면 분포도

4.1. 채점자별 채점 엄격성과 일관성

응시자 18명에 대한 채점 자료를 다국면 라쉬모형을 이용하여 분석한 결과, 채점자 10명의 채점 엄격성은 −.52~.39 logit으로 가장 관대한 채점자인 [KR2]와 가장 엄격한 채점자인 [KR4] 간에는 .91 logit 차이가 났다. 그리고 채점자들의 분리 지수(separation index)는 4.36으로 이는 10명의 채점자들이 통계적으로 다른 약 4개의 엄격성으로 분리될 수 있음을 의미한다(Winke et al. 2012:241) 이와 같은 분포는 χ^2=195.3(p<.01), 분리신뢰도 .00으로 채점자 간 엄격성 차이는 통계적으로 유의했다. 채점자 간 일치도는 [(실제일치도−기대일치도)/(100−기대일치도)]로 계산했을 때 그 값이 0.08로[6] 0에 가

까워 모형이 예측한 값에 부합하는 것으로 나타났다. 전체 채점자들 중 채점을 가장 엄격하게 한 채점자는 [KR4]였고, 가장 관대하게 채점한 채점자는 [KR2]로 모두 한국인 채점자들이었다. 그러나 채점자 집단의 평균 측정치는 KR 집단의 경우 −.08, CR 집단은 .05로 차이가 크지 않으나, 중국인 채점자들이 조금 더 엄격하게 채점한 결과를 보였다. 이는 Yu(2010)와는 동일하나 Kang and Ahn(2012), Kim(2016)과는 다른 결과이다.

표 3. 채점자별 채점 엄격성 및 적합도

채점자	측정치 (logit)	오차	내적합 평균제곱	내적합 표준화값
KR1	.02	.06	1.01	.1
KR2	−.52	.06	1.02	.4
KR3	−.19	.06	1.08	1.4
KR4	.39	.06	1.05	.9
CR1	.15	.06	1.07	1.2
CR2	−.04	.06	.93	−1.2
CR3	−.02	.06	.81	−3.5
CR4	.31	.06	.88	−2.3
CR5	−.33	.06	1.12	2.1
CR6	.22	.06	.89	−2.0

채점자 내 일관성 판정은 내적합 평균제곱값이 0.5~1.5나 0.75~1.3 이내에 있는 경우에 적합한 것으로 본다(장소영, 신동일, 2009:81). 그리고 여러 연구들에서는 내적합 평균제곱값에 내적합 표준화값을 함께 적용하여 적합성 여부를 판정하고 있는데(신동일 2001; 박종임 2013; 이영식 2014; 원미진, 김지영 2017 등), 본 연구에서도 내적합 평균제곱값 0.5~1.5와 내적합 표준화값 −2~+2의 범위를 함께 적용하여 적합성 여부를 검토해 보고자 한다.

채점자 국면에 대한 분석 결과를 보여주는 표 3을 보면 채점자들의 내적합 평균제곱값은 .81~1.12 loigt 사이에 분포하고 있어 모든 채점자들의 채점 결과가 적합한 범위 내에 있었다. 그러나 내적합 표준화값을 함께 살펴보면, KR 집단은 .1~1.4 logit 사이로 모두 적합한 범위 내에 분포하여 채점자들이 본인의 엄격성을 그대로 유지하여 채점하고 있었다. 이와 달리 CR 집단의 경우 [CR1], [CR2]는 내적합 표준화값이 각각 1.2,

6) 분석 결과, 실제일치도는 47.8%였고, 기대일치도는 43.2%였으므로 이를 공식에 대입하여 계산하면 [(47.8−43.2)/100−43.2]]=0.08이 된다.

−1,2 logit으로 적합한 범위 내에 있었으나 [CR3], [CR4], [CR6]은 각각 −3.5, −2.3, −2.0 logit으로 점수 부여에 변별력이 부족한 과적합 경향을 보였고, [CR5]은 2.1 logit으로 채점 기준이 일관되게 적용되지 않은 부적합 경향을 보여 일부 중국인 채점자의 채점 일관성에 다소 문제가 있음을 알 수 있었다.

4.2. KR 집단과 CR 집단의 채점 경향 비교

4.2.1 평가문항의 측정치와 채점 편향성

본 연구에 사용된 일곱 개의 평가문항은 1, 2번은 초급 수준, 3, 4, 7번은 중급 수준, 그리고 5, 6번은 고급 수준의 과제로 구성되었고, 4번과 7번은 문항 유형은 동일하나 다른 주제로 제시되었다. 집단별로 평가문항의 난이도에 따른 측정치를 살펴보면, 1번부터 7번까지의 문항에 대해 KR 집단은 −1.16~1.18 logit, CR 집단은 −1.37~1.25 logit으로 CR 집단의 측정치 범위가 .28 logit 더 넓게 나타났다. 또한 평가문항의 분리지수(separation index)는 KR 집단은 9.46, CR 집단은 13.47이고 분리신뢰도는 모두 .00이었으며 이와 같은 분포는 KR 집단의 χ^2=640.3, CR 집단의 χ^2=1251.5(p<.01)로 통계적으로 유의미했다. 이를 통해 KR 집단은 평가문항이 통계적으로 약 9개의 난이도로 분리되고 CR 집단은 약 13개로 분리될 수 있음을 알 수 있다.

각 평가문항에 대한 집단 간 측정치를 살펴보면, 4번은 모두 −.18 logit으로 동일했고, 5번과 6번은 각각 .95, .93 logit, 1.18, 1.25 logit으로 .02, .07 logit 차이가 나 유사한 경향을 보였다. 그러나 1번과 3번은 각각 −1.16, −1.37 logit, −.46, −.73 logit으로 .21, .27 logit 차이를 보였고, 7번은 .20, .74 logit으로 가장 많은 .54 logit의 차이를 보였다. 두 집단 간 측정치에 차이가 있는 문항이 비교적 난이도가 낮은 문항인 것으로 미루어 난이도가 높은 문항보다는 비교적 쉬운 문항에 대해 집단 간 차이가 있을 수 있음을 알 수 있다. 그리고 문항별 난이도에 따른 측정치 순서를 살펴보면, KR 집단의 경우 1 < 2 < 3 < 4 < 7 < 5 < 6의 순으로 문항의 출제 난이도와 동일한 배열 순서를 보였다. 이와 달리 CR 집단은 3번이 2번보다 난이도가 낮게 책정되어 1 < 3 < 2 < 4 < 7 < 5 < 6의 배열 순서를 보였다. 그리고 4, 5번 문항을 제외하고 난이도가 낮은 문항은 KR 집단에 비해 측정치가 더 낮고 난이도가 높은 문항은 측정치가 더 높은 경향을 보였다. 그리고 문항별 적합도를 살펴보면, 내적합 평균제곱값의 경우 두 집단 모두 .5~1.5 logit 사이에 분포하여 해당 난이도로 일관되게 평가된 것으로 보이나 내적합 표준화값을 기준으로 판단하면 KR 집단은 5번 문항은 과적합, 7번 문항은 부적합 경향이 있고, CR 집단은 4, 5번 문항은 과적합, 7번 문항은 부적합 경향이 있는 것으로 분석되었다.

채점자 집단과 평가문항 간의 상호작용 분석에서 통계적으로 유의미한 편향성을 보인 것은 7번이었다. 앞서 두 문항은 측정치에서도 다른 문항에 비해 차이가 컸는데 KR

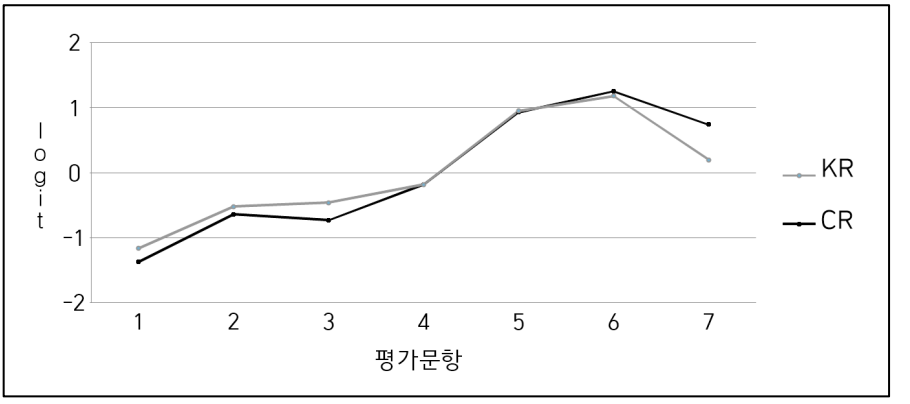

그림 3. 채점자 집단 간 평가문항의 측정치(logit) 비교

표 4. 채점자 집단 간 평가문항의 측정치 및 적합도 비교

문항	측정치 (logit)		오차		내적합 평균제곱		내적합 표준화값	
	KR	CR	KR	CR	KR	CR	KR	CR
1	−1.16	−1.37	.09	.08	.93	1.01	−.8	.2
2	−.52	−.64	.08	.07	1.08	.99	1.0	1.5
3	−.46	−.73	.08	.07	1.07	1.09	.9	.0
4	−.18	−.18	.08	.07	.88	.82	−1.6	−3.2
5	.95	.93	.07	.06	1.20	.74	−2.4	−4.8
6	1.18	1.25	.07	.06	.84	.92	−1.0	−1.4
7	.20	.74	.08	.06	.93	1.31	2.5	4.8

집단은 모형이 기대한 것보다 점수가 높았고, CR 집단은 점수가 낮았다.[7] 편향 크기는 KR 집단이 .29, CR 집단이 −.19 logit이었으며 모두 $p<.01$ 수준에서 유의했다. 4번과 7번 문항은 문항 유형이 동일했는데, 4번은 집단 간 측정치에 차이가 없고 7번은 통계적으로 유의미한 차이가 있는 것으로 미루어 문항 유형으로 인한 차이보다는 친숙하지 않은 주제에 대한 수험자 발화를 판정하는 과정에서 집단 간 차이가 발생한 것으로 보인다. 그림 4는 각 평가문항에 대한 두 집단의 t 값을 그래프로 제시한 것으로 이 그림에서 볼 수 있는 바와 같이 6번과 7번을 제외하고 KR 집단은 모형이 예측한 점수보다 낮은 점수를, CR 집단은 높은 점수를 부여한 것으로 나타났으나 통계적으로 유의미하지는 않았다.

7) 표 6에서 볼 수 있듯이 편향분석에서는 관찰치가 기대치보다 높으면 편향 크기와 t값은 양수로 표기되고, 그 반대의 경우는 음수로 표기된다. 이를 바탕으로 t 값이 양수가 되면 모형이 예상한 것보다 높은 점수가 부여된 것으로, 양수가 되면 반대로 평가된 것으로 해석할 수 있다.

표 5. 채점자 집단과 평가문항과의 상호작용 분석 결과

채점자	평가문항	관찰치	기대치	편향 크기	*t*	*p*
KR	7	1,052	1,000.52	.29	3.85	.0001
CR	7	1,418	1,469.38	.19	−3.12	.0019

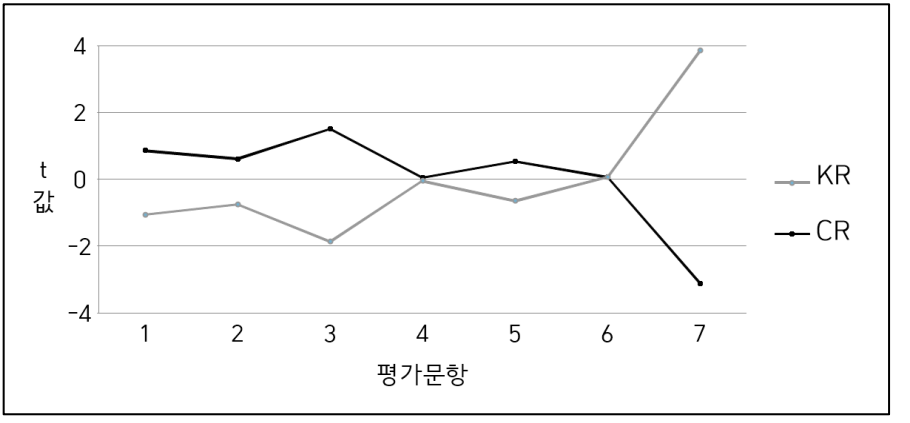

그림 4. 채점자 집단 간 평가문항의 편향성(*t*값) 비교

4.2.2. 평가구인의 측정치와 채점 편향성

본 연구에 사용된 평가구인은 전반적 능력, 어휘 및 문법, 발음, 내용, 조직 등 다섯 개였으며 평가문항과 달리 구인에 따라 엄격성을 달리하거나 가중치를 두지 않았다. 채점자 집단별로 평가구인에 대한 측정치를 살펴보면, 5개의 구인에 대해 KR 집단은 −.31~.43 logit, CR 집단은 −.28~.27 logit의 범위에서 채점을 수행하여 평가문항과 달리 KR 집단의 측정치 범위가 넓게 나타났다. 그리고 평가구인의 분리 지수(separation index)는 KR 집단은 3.77(분리 신뢰도 .93), CR 집단은 3.82(분리신뢰도 .00)였으며, 이와 같은 분포는 KR 집단의 χ^2=77.7, CR 집단의 χ^2=77.8(p<.01)로 통계적으로 유의미했다. 평가문항과 달리 평가구인에 대한 두 집단의 분리지수는 큰 차이가 없어 5개의 평가구인은 두 집단 모두 약 4개로 분리될 수 있음을 알 수 있다.

그리고 구인별 측정치를 살펴보면, KR 집단은 발음, 어휘 및 문법, 조직 구인은 상대적으로 높은 점수가 부여되고, 전반적 능력과 내용 구인에 대해서는 낮은 점수가 부여된 것으로 나타났다. CR 집단은 어휘 및 문법과 발음 구인에 대해서 높은 점수가, 전반적 능력과 조직, 내용 구인에 대해서는 비교적 낮은 점수가 부여되어 KR 집단과 차이를 보였다. 그러나 각 구인에 대한 측정치 차이는 발음 .04, 내용 .05, 어휘 및 문법 .11, 전반적 능력 .11 logit 등으로 .29 logit의 차이를 보인 조직을 제외하고 4개 구인에 대한

측정치 차이는 크지 않았다. 그리고 구인별 적합도를 살펴보면, 내적합 평균제곱값의 경우 두 집단 모두 .5~1.5 logit 사이에 분포하여 해당 측정치로 일관되게 평가된 것으로 보이나 내적합 표준화값을 기준으로 판단하면 KR 집단은 발음과 전반적 능력에 대해서는 과적합, 내용과 조직 구인에 대해서는 부적합 경향이 있고 CR 집단은 발음과 전반적 능력에 대해서는 과적합, 내용은 부적합 경향이 있는 것으로 나타났다. 조직 구인을 제외하면 두 집단 모두 발음과 전반적 능력에 대해서는 과적합, 내용 구인에 대해서는 부적합한 채점 경향을 보였다.

평가구인과의 상호작용 분석에서 통계적으로 유의미한 편향을 보인 것은 집단 간 측정치 차이가 큰 조직이었다. 조직 구인에 대해서 KR 집단은 모형이 예측한 것보다 높은 점수가, CR 집단은 낮은 점수가 부여된 것으로 분석되었다. 편향크기는 KR 집단이 .21, CR 집단이 −.13 logit이었으며 모두 $p<.05$ 수준에서 유의했다. 다른 4개의 구인은 통계적으로 유의미하지 않았으나 어휘 및 문법을 제외하고 다른 세 개의 구인에 대해서 KR 집단은 모형이 기대한 것보다 다소 낮은 점수가, CR 집단은 이와 달리 높은 점수가 매겨진 것으로 나타났다. 평가구인에 대한 측정치에서 CR 집단은 내용과 조직 구인에

표 6. 채점자 집단 간 평가구인의 측정치 및 적합도 비교

평가구인	측정치 (logit)		오차		내적합 평균제곱		내적합 표준화값	
	KR	CR	KR	CR	KR	CR	KR	CR
어휘 및 문법	−.31	−.28	.07	.06	.95	.92	−.8	−1.6
발음	−.14	−.24	.07	.06	.75	.86	−4.2	−2.8
내용	.43	.27	.07	.06	1.28	1.17	4.1	3.1
조직	−.15	.19	.07	.06	1.16	1.10	2.3	1.9
전반적 능력	.18	.05	.07	.06	.7	.84	−3.7	−3.3

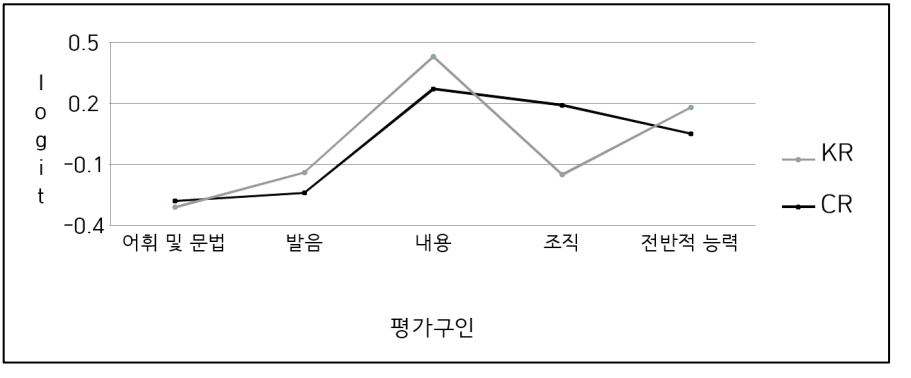

그림 5. 채점자 집단 간 평가구인의 측정치(logit) 비교

낮은 점수를 주고 어휘 및 문법과 발음에는 다소 높은 점수를 주고 있었다. 이와 같은 결과로 볼 때 CR 집단의 경우 수험자 발화의 내용적인 측면과 형식적인 측면을 나누어 내용적인 측면에 대해서는 상대적으로 엄격한 채점 기준을 적용하고 형식적인 측면에 대해서는 관대한 기준으로 점수를 부여한 것으로 보인다.

표 7. 채점자 집단과 평가구인과의 상호작용 분석 결과

채점자	평가구인	관찰값	기대값	편향 크기	*t*	*p*
KR	조직	1,538	1,492.11	.21	3.05	.0024
CR	조직	2,149	2,194.77	-.13	−2.46	.0141

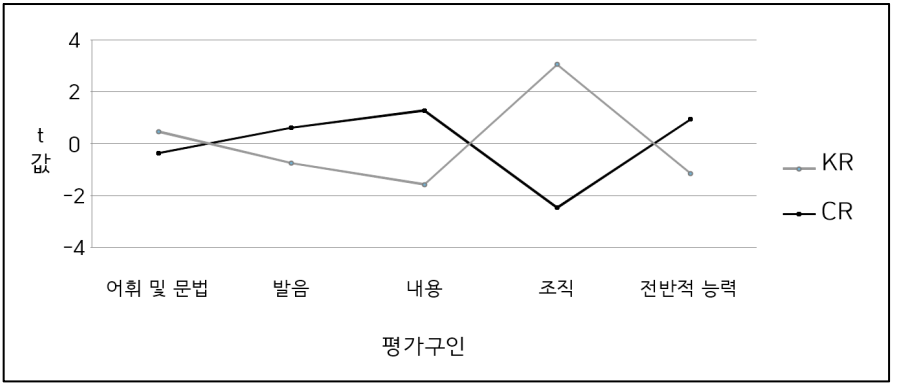

그림 6. 채점자 집단 간 평가구인의 편향성(*t*값) 비교

4.2.3. 수험자의 측정치와 채점 편향성

본 연구에 사용된 126개의 음성 파일은 모두 한국어 학습자 18명이 7개의 문항에 응답한 것으로 앞서 수험자 정보에서 제시한 바와 같이 한국어 학습자의 국적은 중국 9명, 일본 2명, 브라질, 영국, 대만, 홍콩, 캐나다, 말레이시아, 한국(재일교포) 각 1명 등이었고 말하기평가에 참여할 당시의 한국어 능력은 한국어교육기관 등급을 기준으로 3급 3명, 4급 6명, 5급 3명, 6급 6명 등이었다. 채점자 집단별로 수험자에 대한 측정치를 살펴보면, KR 집단은 .67~4.35 logit, CR 집단은 .54~3.76 logit의 범위에 분포하여 KR 집단의 측정치 범위가 넓게 나타났다. 구체적으로는 KR 집단은 수험자의 숙달도가 높은 수험자에 대한 측정치가 CR 집단에 비해 높은 반면 CR 집단은 숙달도가 낮은 수험자에 대한 측정치가 KR 집단에 비해 낮았다. 수험자별 측정치 차이는 .01~.90 logit으로 1, 2, 5, 14, 17번과 같이 두 집단이 거의 동일한 점수를 부여한 수험자가 있는가 하면 10~13번, 그리고 15번과 같이 두 집단에서 매우 다른 점수를 준 수험자도 있었다.

이와 같은 수험자에 대한 분리 지수(separation index)는 KR 집단이 8.24(분리 신뢰도 .99), CR 집단이 8.27(분리신뢰도 .99)이었으며 이와 같은 분포는 KR 집단의 χ^2=1039.8, CR 집단의 χ^2=1171.5(p<.01)로 통계적으로 유의미했다. 즉, 두 집단의 분리 지수는 큰 차이가 없으며 18명의 수험자는 두 집단 모두에서 약 8개의 등급으로 분리될 수 있음을 알 수 있다.

그리고 수험자별 적합도를 살펴보면, 내적합 평균제곱값의 경우 KR 집단은 모두 .5~1.5 logit 사이에 분포하여 해당 측정치로 일관되게 채점되었으나 CR 집단은 17번 응시자의 내적합 평균제곱값이 1.77 logit으로 채점자들이 부여한 점수에 차이가 있는 것으로 나타났다. 그리고 내적합 표준화값을 기준으로 판단하면 KR 집단의 경우, 4번, 6번, 17번 등 한국어능력 등급이 높은 수험자에게는 부적합 경향이 있고 16번, 18번과 같은 한국어능력 등급이 낮은 수험자에게는 과적합 경향이 나타났다. CR 집단도 이와 비슷하게 7번, 14번, 16번, 18번 수험자와 같이 한국어능력 등급이 중급 수준인 수험자들에게는 과적합 경향이, 1번, 15번, 17번과 같이 고급 수준의 수험자에게는 부적합 경향이 분석되었다. 이는 두 집단의 채점자들이 숙달도가 낮은 수험자에 대해서는 비슷하게 낮은 점수를 부여한 반면 숙달도가 높은 수험자에 대해서는 채점자에 따라 상이한 점수를 주었음을 의미한다. 그러나 KR 집단의 경우 수험자의 급수가 3급과 6급으로 숙달도 수준이 상대적으로 아주 낮거나 높은 수험자에 국한되어 있는 데 반해 CR 집단의 경우는 3~6급 수준의 수험자가 모두 포함되어 있어 수험자의 한국어 능력을 판별하는 데에 차이를 보였다.

채점자 집단과 수험자 간의 상호작용 분석에서 통계적으로 유의미한 편향을 보인 사례는 모두 11건이었는데 그 중 5건은 두 집단에서 공통적으로 나타났으며 앞서 살펴본 수험자 측정치 분석에서 집단 간 차이가 큰 수험자들이었다. 수험자의 한국어 능력 등급별로는 3~6급이 고루 나타났고 언어권별로는 중국이 4명, 한국(재일교포)가 1명이

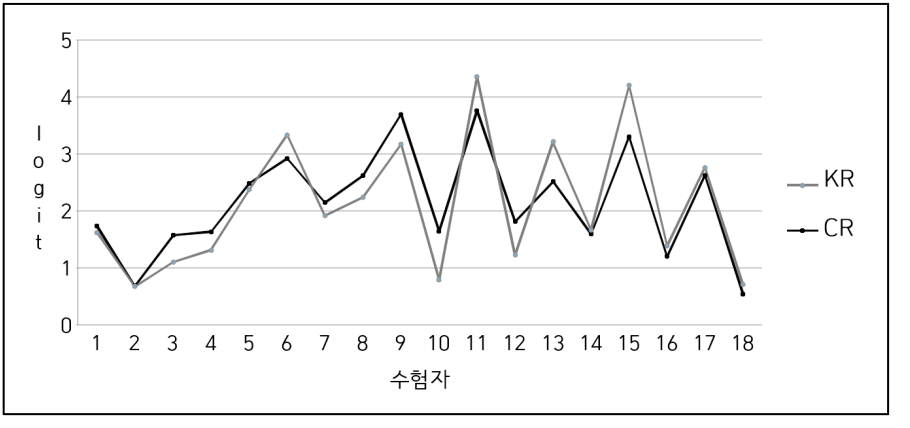

그림 7. 채점자 집단 간 평가구인의 편향성값 측정치(logit) 비교

표 8. 채점자 집단 간 수험자의 측정치 및 적합도 비교

수험자		측정치 (logit)		오차		내적합 평균제곱		내적합 표준화값	
번호(급)	국적	KR	CR	KR	CR	KR	CR	KR	CR
1(5)	중국	1.62	1.73	.12	.10	.80	1.22	−1.8	2.2
2(4)	중국	.67	.68	.12	.10	1.13	.98	1.1	−.1
3(6)	중국	1.10	1.57	.12	.10	1.02	1.14	.2	1.5
4(6)	중국	1.31	1.63	.12	.10	1.31	1.05	2.5	.5
5(6)	브라질	2.37	2.48	.13	.11	1.15	.98	1.2	−.2
6(6)	영국	3.33	2.92	.15	.11	1.54	1.07	3.7	.7
7(4)	말레이시아	1.91	2.14	.12	.10	.95	.81	−.3	−2.1
8(4)	일본	2.24	2.62	.13	.11	.96	.87	−.3	−1.4
9(5)	대만	3.17	3.69	.15	.13	1.03	.82	.3	−1.8
10(3)	중국	.79	1.64	.12	.10	.79	.83	-1.9	−1.9
11(5)	중국	4.35	3.76	.20	.13	.89	1.01	−.6	.1
12(4)	중국	1.22	1.81	.12	.10	.84	.95	−1.4	−.4
13(4)	한국	3.21	2.51	.15	.11	.96	.89	−.2	−1.2
14(4)	일본	1.66	1.59	.12	.10	.84	.77	−1.4	−2.7
15(6)	중국	4.20	3.30	.19	.12	.87	1.33	−.8	3.0
16(3)	홍콩	1.39	1.20	.12	.10	.72	.74	-2.6	−3.0
17(6)	중국	2.76	2.62	.14	.11	1.47	1.77	3.5	6.7
18(3)	캐나다	.71	.54	.12	.10	.72	.58	−2.7	−5.2

었다. 5건의 사례 중 CR 집단에서 모형이 기대한 점수보다 높은 점수를 받은 수험자는 10번과 12번이었고 11, 13, 15번 수험자는 모형이 예측한 점수보다 낮은 점수를 받았다. 이를 수험자의 한국어교육기관 등급을 기준으로 살펴보면 CR 집단은 13번 수험자를 제외하고 등급이 낮은 10번, 12번 수험자에게 후한 점수를 준 반면 급수가 높은 11번, 15번 수험자에게는 박한 점수를 준 것으로 해석된다. 그리고 수험자의 언어적 배경을 기준으로 살펴보면 CR 집단은 중국인의 발음과 억양이 강한 10번, 12번 수험자에게는 높은 점수를, 재일교포인 13번, 조선족인 15번 수험자에게는 낮은 점수를 부여한 것으로 보인다. 이와 같은 경향은 KR 집단에서는 상반되게 나타났다. 통계적으로 유의미한 편향으로 분석된 사례가 많지 않아 이와 같은 해석을 일반화할 수는 없으나 수험자의 발음이나 억양이 채점자에게 익숙할 때 보다 높은 점수를 받을 수 있다는 연구 결과를 보고한 Carey et al.(2011), Winke et al.(2012)과 같이 한국어교육에서도 동일한 언어권의 수험

자를 평가할 때 비원어민 채점자의 모국어 배경이나 원어민 채점자가 학습한 제2, 제3언어가 채점에 어떠한 영향을 미치는가에 대한 후속 연구가 필요할 것으로 생각된다.

표 9. 채점자 집단과 응시자와의 상호작용 분석 결과

응시자	채점자	관찰값	기대값	편향 크기	*t*	*p*
6	KR	492	475.97	.34	2.25	.0258
10	KR	334	369.08	−.48	−4.11	.0001
	CR	576	540.89	.32	3.35	.0010
11	KR	527	511.70	.50	2.57	.0112
	CR	746	761.27	−.26	−2.05	.0419
12	KR	366	388.15	-.31	−2.63	.0094
	CR	592	569.82	.21	2.14	.0335
13	KR	487	461.10	.51	3.47	.0007
	CR	649	674.85	−.29	−2.80	.0056
15	KR	523	498.93	.69	3.71	.0003
	CR	717	741.04	−.35	−2.98	.0033

4.2.4. 평가척도 활용 양상

본 연구에서는 0점부터 4점까지 5단계 척도를 사용했다. 집단별로 평가척도 사용 횟수를 살펴보면, KR 집단은 0점 7회, 1점 149회, 2점 544회, 3점 995회, 4점 825회 등으로 3점을 가장 많이 사용했고 다음으로 4점, 2점, 1점, 0점의 순으로 점수를 부여했다. CR 집단도 0점 6회, 1점 188회, 2점 906회, 3점 1,616회, 4점 1,062회 등으로 KR 집단과 동일하게 3점, 4점, 2점, 1점, 0점의 순으로 점수를 부여했다. 그러나 사용 비율을 살펴보면, 두 집단 모두 가장 많이 사용한 3점은 각각 39%, 42%으로 유사한 반면 2점과 4점에 대해서는 KR 집단이 각각 22%와 33%를 사용하고 CR 집단이 25%, 28%를 사용하여 2점은 CR 집단이, 4점은 KR 집단이 더 많이 사용한 것으로 나타났다. 그리고 평가척도의 중간 점수인 2점과 3점의 비율을 계산했을 때는 KR 집단이 61%를, CR 집단이 67%를 사용하고 있어 CR 집단이 극단의 점수를 피하고 중간 점수를 더 많이 부여했음을 알 수 있다.

그리고 평가척도별 평균 측정치를 살펴보면, KR 집단은 척도에 대한 평균 측정치가 점수가 높아질수록 일관되게 조금씩 증가하는 경향을 보인다. 즉, 0점에서 1점이 되는 데에는 .68 logit이 필요하고 1점에서 2점이 되는 데에는 .73 logit, 2점에서 3점이 되는 데에는 1.02 logit, 3점에서 4점이 되는 데에는 1.16 logit 등이 필요하여 비율이 균일하

표 10. 채점자 집단 간 평가척도 적합도 비교

채점자	평가 척도	사용 횟수	사용비율 (%)	평균 측정치	기대 측정치	외적합 평균제곱	임계값
KR	0	7	0	－.40	－.25	.9	
	1	149	6	.28	.28	1.1	－3.06
	2	544	22	1.01	1.03	1.0	－.66
	3	995	39	2.03	2.02	1.1	.90
	4	825	33	3.29	3.30	1.0	2.82
CR	0	6	0	1.32	－.18	1.9	
	1	188	5	.30*	.36	.9	－3.37
	2	906	25	1.05	1.16	.9	-.83
	3	1,616	42	2.29	2.17	1.1	1.07
	4	1,062	28	3.15	3.24	1.1	3.13

지는 않으나 척도가 높아질수록 조금씩 더 높은 측정치를 필요로 했다. 그러나 CR 집단은 척도 간 차이가 들쑥날쑥하여 0점에서 1점이 될 때에는 −1.02 logit만큼 떨어지고 1점에서 2점이 되는 데에는 .75 logit, 2점에서 3점이 되는 데에는 1.24 logit, 3점에서 4점이 되는 데에는 .86 logit 등이 필요하여 척도가 높아질 때에 필요한 측정치가 일관된 경향을 보이지 않았다. 특히 0점에 대한 평균 측정치는 1.32 logit으로 2점(1.05)보다 더 높게 나타났다.

평가척도 사용의 적합성을 확인할 수 있는 외적합 평균제곱값은 1.0을 기준으로 신뢰성 정도를 판단한다(장소영, 신동일, 2009:84). KR 집단의 경우 각 척도가 .8~1.1 사이에 분포하여 척도 사용이 신뢰할 만한 것으로 해석할 수 있다. CR 집단은 1~4점은 .9~1.1에 분포하여 척도 사용이 적절하였으나 0점은 1.9로 평균 측정치가 높았던 만큼 신뢰도는 낮은 것으로 분석되었다. 그리고 각 평가척도가 서로 교차하는 지점을 수치로 나타낸 임계값은 KR 집단은 −2.82~2.79logit, CR 집단은 −3.37~3.13 logit으로 CR 집단의 범위가 더 넓었다.

5. 결 론

본 연구는 한국어교육을 전공하는 대학원생을 대상으로 채점자 훈련을 실시하고 한국인 채점자들과 중국인 채점자들의 채점 결과를 분석하여 집단 간 채점 경향의 차이를 비교하기 위해 수행되었다. 개별 채점자의 채점 엄격성과 일관성 분석 결과, 가장 엄격

하거나 관대하게 채점한 채점자는 한국인 채점자들이었으나 집단의 채점 엄격성 평균치는 중국인 채점자 집단이 조금 더 높았다. 채점자 내 일관성은 내적합 평균제곱값만을 기준으로 했을 때 모든 채점자들이 일관되게 채점 기준을 적용하는 경향을 보였으나 내적합 표준화값을 함께 적용했을 때는 CR 집단 중에 부적합하거나 과적합한 경향의 채점자들이 있었다. 구체적으로는 채점에 참여한 박사과정생 중 2명은 일관되게 채점하는 것으로 나타났으나 1명은 과적합한 경향을 보였고 석사과정생 3명 중 2명은 과적합, 1명은 부적합한 경향이 있는 것으로 분석되었다. 여기서 과적합한 결과는 비슷한 점수를 많이 부여했다는 것을, 부적합한 경향은 점수를 부여하는 기준이 일정하지 않았음을 의미하므로 중국인 채점자들 중 일부에서 일관된 채점 기준 적용에 문제가 있었음을 알 수 있다.

두 집단의 채점 경향을 분석한 결과에서 평가문항에 대한 측정치는 3번과 7번 문항을 제외하면 두 집단이 유사했으나 분리지수는 CR 집단이 높았다. 그리고 측정치 차이가 컸던 7번 문항은 통계적으로 유의미한 편향이 분석되었는데 7번은 4번과 동일한 유형의 문항으로 난이도가 중급 수준으로 출제되었으나 CR 집단에서는 고급 수준의 문항인 5번과 4번이 비슷한 난이도로 측정되었다. 그리고 평가구인에 대한 분석 결과에서 두 집단의 분리지수는 비슷했으나 조직과 어휘 및 문법 구인은 KR 집단에서 CR 집단보다 난이도가 높게 나타났고, 다른 3개의 구인은 상대적으로 난이도가 낮은 것으로 분석되었으며 조직에 대해서는 통계적으로도 유의미한 편향이 분석되었다. 특히 CR 집단은 발화의 형식적 요소인 어휘 및 문법, 발음 구인의 난이도는 대체로 낮게, 발화의 내용적 요소인 내용과 조직 구인의 난이도는 높게 분석되어 채점에 이분법적인 경향을 보였다. 수험자에 대한 분석 결과에서 분리지수는 집단 간 차이가 적었다. 적합도는 KR 집단의 경우 일부 숙달도가 상대적으로 낮거나 높은 경우에 과적합, 부적합 경향이 분석되었으나 CR 집단의 경우는 3, 4급에 해당되는 일부 수험자에게서 과적합 경향이, 5, 6급에 해당되는 일부 수험자에게서는 부적합 경향이 나타나 중급과 고급 수준에 대한 판별은 KR 집단과 유사하나 세부 등급에 대한 판별에는 차이를 보였다. 그리고 통계적으로 유의미한 편향이 분석된 수험자 중 일부의 경우 CR 집단은 중국인의 발음과 억양이 강한 수험자에게는 높은 점수를, 재일교포나 조선족인 수험자에게는 낮은 점수를 부여하였는데 이와 같은 경향은 KR 집단에서는 상반되게 나타났다. 마지막으로 평가척도 활용 양상에서 CR 집단은 0점 척도 사용의 신뢰도가 낮게 나타났고 각 척도의 평균 측정치 차이가 균일하지 않았으며 KR 집단에 비해 최고점보다는 중간 점수 사용을 선호하는 경향을 보였다.

이상의 결과를 통해 KR 집단과 CR 집단은 평가문항과 평가구인에 대한 결과 분석에서 일부를 제외하고 유사한 측정치를 보였다는 점에서 집단 간 채점 경향의 유사성을 찾을 수 있었다. 유의미한 편향이 분석된 일부 문항과 구인을 통해서는 선행연구 결과와 동일하게 두 집단이 서로 다른 기준을 적용하여 채점하는 부분이 있음을 알 수 있었으나 연구마다 평가에 사용하는 평가문항이나 구인이 달라 공통된 특정 문항 유형이나 구인

을 도출할 수는 없었다. 채점 경향의 차이점과 관련하여 두 집단은 수험자의 숙달도 판정에서 차이를 보였는데 수험자 숙달도가 상대적으로 높은 경우와 수험자의 발음이나 억양에 특정 언어가 반영되어 있을 때 집단 간 차이가 컸으며 일부는 유의미한 편향으로 분석되었다. 아울러 CR 집단의 경우 채점 일관성이 부족한 일부 채점자들이 있었고 평가척도 활용에 있어서도 2점이나 3점과 같은 중간 점수 사용에 대한 비율이 높았으며 0점에 대한 신뢰도 수치가 낮게 나타나 채점 일관성과 평가척도 적용에 대한 훈련이 더 필요할 것으로 보인다. 이는 CR 집단의 경우 모두 대학원생으로 아직 현장에서의 교육 경험이나 평가 경험이 부족한 데에서 하나의 원인을 찾을 수 있을 것이다.

본 연구는 앞으로 한국어교육 현장에서 활동하게 될 한국어교육 전공 대학원생 중 한국인과 특정 국적의 외국인을 대상으로 채점자 훈련을 실시하고 그 결과를 분석하여 두 집단의 채점 경향을 비교하였다는 점에서 연구의 의의를 갖는다. 그러나 그 결과를 일반화하기에는 각 집단의 채점자 수가 적고 양적인 방법만을 사용하여 그러한 경향의 원인을 밝히지 못했다는 점에서 한계가 있다. 또한 본 연구에서는 한국인 대학원생의 경우 한국어교육기관에 종사하는 한국어교원을 대상으로 하였으나 중국인 대학원생의 경우는 정식 한국어교육 경험이 없었으므로 향후 현직에 종사하는 비원어민 채점자를 대상으로 연구를 수행한다면 보다 의미있는 결과를 얻을 수 있을 것으로 생각된다. 최근 한국어교육을 전공하는 외국인 유학생이 점차 증가하고 있고 원어민 채점자와 비원어민 채점자 간에 채점 경향의 차이가 있다는 연구 결과가 꾸준히 보고되고 있는 만큼 여러 국적의 한국어교육 전공자나 비원어민 한국어 교원을 대상으로 한 후속 연구가 이루어진다면 원어민과 비원어민 채점자의 채점 특성을 파악할 수 있어 신뢰도 높은 채점자를 양성하는 데에 유익한 자료가 될 것이다.

References

강석한, 안현기. (2014). “외국인 한국어 말하기 시험의 평가자 요소가 채점에 미치는 영향”, 『이중언어학』 55, 1-29.

김가람. (2016). “중국인 한국어교원에 대한 이론적 고찰”, 『한국어교육』 27(1), 1-20.

김지영. (2018). 『한국어 말하기 평가 채점자의 채점 경향 연구』, 박사학위논문, 연세대학교, 서울.

김향란. (2019). 『중국 내 대학 한국어학 전공자를 위한 기본 교재 개발 연구』, 박사학위논문, 상명대학교, 서울.

박동호, 김유미, 김현정, 신동일, 우창현, 이영식, 조수진, 지현숙. (2012). 『한국어능력시험의 CBT/IBT 기반 말하기 평가를 위한 문항 유형 개발』, 국립국제교육원.

박종임. (2013). 『국어교사의 쓰기 평가 특성 연구』. 박사학위 논문, 한국교원대학교, 충북.

백현영, 양병곤. (2011). “중학교 영어교사의 말하기평가 채점경향 분석”, 『언어과학』 18(4), 77-99.

신동일. (2001). “채점 경향 분석을 위한 Rasch 측정모형 적용 연구”, 『Foreign Language

Education』 8(1), 249-272.
신동일, 설현수. (2005). “NEW FACETS을 활용한 채점자료 분석방법”, 『Foreign Languages Education』 12(2), 191-211.
원미진, 강현화, 김미옥, 김성숙, 김현정. (2017). 『제4차 한국어능력시험 말하기 평가 개발 연구』, 국립국제교육원.
원미진, 김지영. (2017). “한국어 말하기 평가 개발을 위한 채점 경향 분석 연구”, 『외국어로서의 한국어교육』, 47, 169-192.
이영식. (2014). “다국면 Rasch 측정의 Facets 프로그램을 활용한 영어 작문 평가의 원어민 채점 검증”, 『영어어문교육』, 20(1), 475-496.
이향. (2013). “한국어 말하기 평가의 발음 영역 채점에서의 채점자 특성에 따른 채점 경향 연구: 한국어 교육 경험과 전공을 중심으로”, 『외국어로서의 한국어교육』 39, 213-245.
장소영, 신동일. (2009). 『언어교육평가 연구를 위한 FACETS 프로그램: 기초 과정편』, 서울: 글로벌콘텐츠.
Carey, M. D., Mannell, R. H. and Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interview?. *Language Testing* 28(2), 201-219.
John M. Linacre. (2014). A *User's Guide to FACETS Rasch-Model Computer program*. Winsteps.
Kang, Seokhan and Hyunkee Ahn. (2012). A comparative study on criteria and tasks in Korean English Speaking Assessment by Native and Non-native raters. *Language Research* 48(2), 241-262.
Kim, Y. -H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing* 26(2), 187-217.
Kim, Hyunah. (2016). Comparing native and non-native rater assessments of Korean Oral Proficiency: A FACETS analysis. *Korean Language Education Research* 51(5), 84-113.
Kim, Hyun Jung. (2011). Investigating rater behavior across diverse English speaking tasks. *Foreign Language Education* 18(2), 99-125.
Lee, Seongyong and Chae, Hohsung. (2012). Rating of Korean students' L2 writing: similarities and differences between native and non-native raters. *The Journal of Curriculum and Instruction Studies* 16(3), 629-655.
McNamara, T. F. (1996). *Measuring Second Language Performance*. Pearson Education Ltd.(채선희 외 옮김(2003), 『문항반응이론의 이론과 실제: 외국어 수행평가를 중심으로』, 경기:서현사.)
Gui, Min. (2012). Exploring differences between Chinese and American E.F.L teachers' evaluation of speech performance. *Language Assessment Quarterly* 9(2), 186-203.
Winke, P., S. Gass and Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30(2), 231-252.
Yu, Kyung-Ah. (2010). The effect of raters' language background on English-speaking test ratings across test-takers' oral proficiency levels. *Applied Linguistics* 26(4), 395-419.
Shi, Ling. (2001). Native-and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing* 18(3), 303-325.

김지영
이화여대 언어교육원 한국어교육부
한국어강사
03760 서울시 서대문구 이화여대길 52
전자우편: gb9802@hanmail.net

접수일자 : 2019. 3. 1
수정본 접수 : 2019. 3. 25
게재결정 : 2019. 4. 2