『이중언어학』 제63호(2016.6.30). 二重言語學會

한국어 말하기 평가에서 고정 오류 연구

강 석 한

Abstract

Kang, Seok-Han. 2016. 6. 30. **A study of the constant errors in criterion-referenced examinations of Korean speaking**. *Bilingual Research 63*, 1-22. Raters' constant errors, represented mostly as leniency errors and halo effect, have been examined for the L2 Korean speaking assessment by native Korean speakers. Fifteen Korean raters assessed twenty-seven native English, Japanese, and Chinese learners of Korean with three levels of beginning, intermediate, and advanced groups. FACETS program with multi-facets Rasch model has been used to analyze the data. The results are as follows: (1) test-takers' mother tongue has little impact on Korean raters' rating patterns in a whole, but they tend to grade leniently to English learners of Korean, (2) test-takers' levelling by L2 Korean proficiency show statically significant influence on raters' ratings. The study finds that Korean raters could have meaningfully constant errors on their assessment, even though they don't agree with committing constant errors in the survey. In details, they have more leniency grading to English learners of Korean, while they have halo effects to the beginning stages of L2 Korean learners. The result implies that Korean raters should carry out the authentic speaking assessment training before the real tests. **(Konkuk University)**

【Key words】 Korean speaking assessment(한국어 평가), raters(평가자), criteria(기준), tasks(과업), constant errors(고정 오류), halo effect(후광 효과), criteria-based assessment(절대 평가), bias(편견), leniency errors(관대성 오류)

1. 서론

현재 외국인을 대상으로 하는 대학부설 한국어 교육현장에서는 자체 교육평가안을 이용하여 학생들의 한국어 능력에 대한 평가가 활발히 이루어지고 있다. 보통은 읽기, 듣기, 쓰기, 말하기의 전 영역에 대한 평가가 실시되는데, 이를 근거로 반 배정 및 상급반 진학, 기타 수료 여부가 결정된다. 이중 말하기 영역은 평가 환경으로 인하여, 한 두 평가자에 의한 절대 평가가 주를 이루고 있다.

절대 평가는 상대적 등급 비율이 존재하는 상대평가와 달리, 평가자의 평가준거에 의해 피평가자의 수행능력을 자의적으로 판단하기 때문에, 평가의 '공정성'(fairness)은 매우 중요한 가치 항목이 된다. 미국 국립교육평가위원회(National Council on Measurement in Education, 1999)는 '공정성은 편견의 부재, 평가 과정에서 피평가자에 대한 공정한 처리, 각기 다른 집단에 속한 피평가자의 평가 결과 공정성, 측정내용에 접근하는 학습기회에 대한 동등성으로 구성된다'고 하였다. Kunan(2008)도 공정성은 타당성, 편견의 부재, 과정 행정 사회적 결과등의 동등성이 고려될 때 공정한 평가가 이루어진다고 보았다. 기존 연구들을 살펴보면 평가의 공정성은 단일 요소로 이루어지는 것이 아니라, 평가 과정 이행에서 발생하는 많은 요소가 개입되어 있고, 이 요소는 각 문화권이나 사회가 평가자의 평가에 영향을 미치는 평가외적인 측면뿐만 아니라, 평가내용 및 방법같은 내적요소에 의하여 좌우된다고 할 수 있다.

절대평가에서 공정한 평가를 저해하는 주요한 요소로 평가자의 오류를 둘 수 있다. Murphy와 Balzer(1989)는 평가자의 평가오류가 시험 결과의 신뢰성에 커다란 영향을 미칠 수 있다고 주장하면서, 이들이 지닌 고정오류가 평가의 공정성에 흠집을 낼 수 있다고 지적했다. Bachman과 Palmer(1996)은 이러한 고정오류는 총체적 채점 방식에 가장 잘 나타나며, 따라서 분석적 방식을 권장하고 있다. 이 연구들에서는 절대 평가에

서 고정 오류가 존재할 수 있으며, 공정한 평가를 위해서는 고정오류 인자를 찾아내는 것과 함께 평가 방식의 개선을 요구하고 있다.

이상적인 평가 방식은 해당 평가가 신뢰성(reliability)과 타당도(validity)를 확보하고 있다는 의미이다(신상근, 2010). 타당도란 관련된 요소들을 포함하는 평가가 정당화될 수 있는가 하는 것이며, 신뢰성이란 표본과 수험자 수행능력을 일관성 있고 정확하게 평가하는 것을 의미한다. 결국, 표현영역에서의 수험자들의 언어 표현 능력의 공정한 평가를 위하여 높은 수준의 신뢰도와 타당도를 확보하는 것은 매우 중요하다. 그러나 이런 공정성을 담보하는 신뢰도와 타당도는 평가과정이나 평가자가 지니는 고유의 오류인자에 인하여 훼손될 수 있다.

평가 오류란 평가자가 피평가자의 표현 능력을 평가할 때 체계적으로 발생하는 오류로, 편견이나 주관적, 혹은 외부의 영향을 받지 않는 이상적인 평가 결과와 실제 평가자의 부정확한 판단과정에 의하여 도출되는 평가결과와의 차이를 의미한다(Pulakos, 1984). 이런 신뢰도에서 중요한 내적 일관성(consistency)을 무너뜨리는 고정 오류중에 대표적인 것이 관대화 오류와 후광효과이다. 관대화 오류는 평가오류에 관한 기존의 언어평가 연구들에서 가장 많이 다루어졌으며, 피평가자들의 능력이나 성과를 실제보다 더 높게 혹은 더 낮게 일관되게 평가하는 것을 의미한다. 후광효과란 평가자가 피평가자의 언어 능력에 대한 인상을 근거로 모든 표현 영역의 구인에 대하여 전반적으로 좋음 혹은 나쁨으로 일관되게 평가하는 경향을 말한다. 일반적으로 후광효과는 평가자가 피평가자의 언어능력이외의 당사자 및 외부 특성에 대하여 특정 요소에 비중을 두고 평가할 때 평균 이상으로 과도하게 발생한다. 이러한 후광효과가 발생하면 평가도구의 평가 변수간 척도 변별력을 떨어뜨리는 결과를 낳고, 이는 평가의 신뢰도와 타당도에 부정적인 영향을 미치게 된다. 본 연구는 외국어로서의 한국어 말하기영역에서 절대 평가에 대한 고정 오류 인자를 후광효과와 관대화 오류 위주로 찾아내어 평가 결과의 공정성을 담보

할 수 있는 교육적 함의를 제시하려고 한다.

2. 선행 연구

고정오류에 대한 연구는 한국어 교육평가분야에서는 찾아보기 힘들지만, 해외에서는 1960년대 이후로 언어평가 영역에서 꾸준히 연구되어져 오고 있다. 현재까지 알려진 평가오류는 대비 효과(contrast effect), 첫인상 효과(first impression effect), 후광효과(halo effect), 유사성 오류(similar-to-me error), 관대화 오류(leniency error), 최신오류(recent error), 논리적 오류(logical error) 등이 있다(김명소, 이광진, 2008). 이중에 가장 대표적인 것은 고정오류(constant errors)에 속하는 관대화 오류와 후광효과인데, 이 오류들은 평가되는 변수들간의 잘못된 유사성이 반영될 경우 발생한다.

관대화 오류는 평가오류에 관한 기존의 언어 평가 연구들에서 가장 많이 축적되어 왔다. Bardor, Yeager, Klingsporn(l982)은 관대화 오류란 평가시 척도의 중위점수로부터 일관되게 낮게 혹은 높게 평가하는 것을 의미한다고 하였다. 따라서 이 경우 피평가자의 능력이나 성과를 실제보다 과소/과장되게 평가하게 된다고 보고하고 있다. 이 측면에서 본다면 관대화 오류는 엄격화 오류(severity errors)를 포함하는 개념이다. Bernardin, Alvares, Cranny(1976)는 평가자의 평가점수 패턴이 평가 척도상에서 중간지점에서 부적인 혹은 양적인 방향으로 일관되게 이동하는 경향이라고 정의한다. 따라서 이런 평가에서는 일관된 오류 반응경향을 파악할 수 있다고 보았다. 결국, 관대화 오류는 다른 평가자들이 부여한 점수에 의해서 도출된 점수보다 과도/과소하게 일관되게 평가된 평가 패턴을 의미한다(Becher, Maris, Hsiao, 2010). 이 같은 관대화 오류는 평가 상황에서 가장 흔히 나타나는 오류로서 평가 결과를 비정상적으로 왜곡시키는 위험을 안고 있으며(Solomon, Lance, 1997), 실제로 수

행평가에 의한 절대평가가 확산되면서 학교의 입장에서는 평가의 변별력저하로 인해, 피평가자간 개인차 측정 능력이 실제적으로 이루어지지 않는 것이 가장 큰 문제점으로 지적되어 왔다(신동일, 2006). 일반적으로 관대화 오류의 원인으로는 자기 보호측면에서 설명된다. 피평가자에게 낮은 혹은 높은 점수를 부여할 경우 발생할 수 있는 피평가자와의 심적인 갈등, 평가 관련자, 혹은 사회에 미치는 영향등을 고려하여 왜곡하여 평가한다고 할 수 있다.

후광효과란 평가자가 피평가자의 측정하려는 능력이외의 단편적인 특징에 의한 인상(impression)을 근거로 모든 표현 영역의 구인에 대하여 전반적으로 좋음 혹은 나쁨으로 평가하는 경향을 말한다. 일반적으로 후광효과는 평가자가 피평가자의 언어 능력외적인 특성으로 인하여 부정적인 요소에 비중을 두고 평가할 때 발생한다. 혹은 역으로 피평가자들이 인상관리를 통하여 평가자에게 긍정적인 인상을 만듦으로서 의도적으로 후광효과를 조장할 수 있다. 이러한 후광효과가 발생하면 평가도구의 평가 변수의 척도 변별력을 떨어뜨리는 결과를 낳고, 이는 평가의 신뢰도와 타당도에 부정적인 영향을 미치게 된다. 즉, 평가 척도들이 중복적인 정보만을 제공해 줌으로써 주어진 평가척도에 대한 타당도를 떨어뜨리게 된다. 학교에서 실시하는 수행평가에서는 과업에 대한 여러 가지 준거들이 제시되어 피평가자의 수행능력이 정확하게 진단하는 것이 중요한데, 후광효과가 높으면 정확한 분석이 어렵게 된다.

우리나라에서는 평가자측면의 고정오류에 대한 연구가 매우 드믄데, 극소수의 연구들도 영어 말하기 평가에 집중하고 있다. 신동일, 장소영(2002)은 MATE 채점자를 대상으로 채점오류와 후광효과의 원인을 분석하였다. 그들은 채점 신뢰도의 심각한 오류를 제공하는 후광효과의 원인으로 총체적 채점방식, 적절치 못한 평가기술표, 일치도 중심의 채점자 교육방법, 채점자의 노력 및 역량부족을 원인으로 지목했다. 이호, 김하나(2009)에서는 영어 쓰기 절대 평가에서 대면방식과 익명평가 방식의

차이에 의하여 오류 인자가 발견된다고 보고하고 있다. 이러한 오류를 감소하는 방법으로 (1) 분리식 채점 방법, (2) 평가 영역과 등급간에 차별적인 정보가 담겨진 준거지향적 기술표 작성, (3) 채점자간 일치된 점수 대신 채점자 개인의 일관성을 강조한 채점자 교육, (4) 채점자 공동체를 통한 협력 작업이나 개인모니터 시스템 구축을 제안하였다.

본 연구는 한국어 말하기평가에서 평가자의 오류를 검증하는 연구이다. 아직 한국어 평가분야에서 말하기평가가 본격적으로 도입되지 않았고, 따라서 평가자 오류에 대한 연구가 축적되어 있지 않았다. 그러나 점차 말하기 평가의 필요성이 제기되고 있으며, 교육현장에서 이 영역의 평가 적용은 점차 확대될 것이라고 예측할 수 있다. 이런 측면에서 평가자 오류 연구는 반드시 필요하다고 보며, 본 연구는 선도연구로서 의미를 갖는다.

3. 연구방법 및 내용

3.1. 평가자 및 피평가자

한국어 평가 자료로 사용된 발화 자료는 27명의 외국인 남성 발화자의 녹음이다. 이들의 국적은 미국인 9인, 일본인 9인, 중국인 9인이다. 이들은 평가 당시에 초급반(1, 2급) 9인(미국인 3인, 일본인 3인, 중국인 3인; 이하 같음), 중급반 9인 (3, 4급), 그리고 고급반 9인(5, 6급)으로 구성되었다. 이들이 다니는 한국어 어학교육원은 모두 6급(단계) 수준으로 되어 있으며, 각 급수의 승급은 언어의 4가지 영역을 모두 측정하여 평균 60점 이상이면 진급을 하였다. 이들의 한국 체류 기간은 평균 3.7년(표준편차 2.6년), 연령은 24.3세(표준편차 2.4세) 이다.

본 연구와 관련하여 말하기 평가에 대한 연구 목적을 사전에 공지하여 승낙을 받았다. 발화자들은 모 대학교 소속 한국어 어학교육센터에 소속된

학생들로서, 그들의 직업은 대학원 학생, 주한 미군 군무원, 상사원, 어학원 강사 등으로 다양한 분야에서 한국 사회와 문화를 접하고 있었다. 이들에 대한 한국어 말하기 평가는 조용한 교실에서 27대의 컴퓨터를 통하여 화면에 문제가 제시되었고, 피험자들은 연구자 지시에 따라서 컴퓨터 마이크에 발화를 하였고, 이를 녹음하였다. 일반적으로 평가 방식에는 인터뷰 방식과 컴퓨터를 이용한 반-직접 평가 방식(semi-directed interview)이 있는데, 반 직접 평가 방식을 취한 이유는 연구에 사용할 시료의 객관성 확보 때문이다.

이 녹음된 발화 자료는, 2차례에 걸쳐서 15인의 한국인 한국어 평가자들에 의하여 평가를 수행하였다. 평가자들은 헤드셋을 통하여 수험자들의 녹음을 들으면서 개별적으로 평가를 하였으며, 사전에 평가지와 기준표가 제시되었다. 평가자들은 대부분 한국어 교육학/국어 교육학/국어국문학 분야의 박사과정(4명) 혹은 학위를 보유(11명)하고 있으며, 대부분 한국어 교사 자격증 2급 이상 소지자들이다. 또한 이들은 대학에서 한국어 교육을 담당하고, 연령이 28 ~ 38세 정도(평균 32.8세, 표준편차 5.7세)되는 한국어 강의 경력이 1 ~ 6년 정도(평균 3.4년, 표준편차 3.3년)되는 초/중견 교/강사들이다. 이들은 남자 3인, 여성 12명으로 구성되어 있으며, 소정의 참여비가 지급되었다.

3.2. 연구방법

이들의 평가 분석은 다면-측정 Rasch 모형에 근거한 FACETS(Lumley & McNamara, 1995; Weigle, 1998; 신동일, 2006)를 바탕으로 수준별, 모국어별로 평가상의 고정 오류가 개입되어 있는지를 분석하였다. 본 연구 기간중에 평가자들에게 2회에 걸친 연구 평가가 이루어졌다. 1차 평가에서는 피평가자들의 모국어 및 수준에 대한 정보를 공지하지 않고, 무작

위로 배열된 피평가자의 발화파일이 제공되었다. 15명의 교강사들이 한 교실에서 헤드셋을 통하여 들리는 소리를 바탕으로 5단계 리커 스케일(1: 매우 나쁨, 5: 한국어 원어민 발화와 유사함)로 한국어 발화 능력을 판단하도록 하였다. 즉, 이 분석은 '총체적 평가'방법으로 '피평가자가 얼마나 한국어 말하기 능력을 보여주고 있는가'를 평가한 점수를 바탕으로 이루어졌다.

1주일 후 동일한 교실에서 피평가자들의 모국어 및 한국어 수준 정보가 제시된 재배열된 동일한 피평가자들의 발화 음성이 평가용으로 다시 제시되었다. 따라서 두 차례에 걸친 평가자료는 동일하지만, 단지 피평가자들의 모국어 및 수준별 정보가 제공되었는가의 여부가 다르다.

이 연구가 종료된 후에 평가자들에게 설문지가 주어졌다. 본 연구에 사용된 설문지는 Foddy(1993)의 제안에 따라, 문장구조가 간결하고 묻고자하는 바가 뚜렷하도록 구성하였다. 설문지는 피평가자의 모국어와 한국어 수준이 평가에 영향을 미쳤는지에 대하여 5단계 리커스케일로 구성하였다(1: 절대 영향을 미치지 않았다, 5: 매우 큰 영향을 미쳤다).

분석과정을 보면, 2회에 걸친 실험 종료후 피평가자 27명의 한국어 발화에 대한 15명의 평가자들의 평가 자료를 FACETS에 입력한 후 분석하였다. 여기에서 도출되는 평가자료는 평가자들의 신뢰도와 타당도 검증을 위하여 문항 반응 이론중 하나인 Rasch 모형에 기반을 둔 통계 프로그램 FACETS를 적용하였다. 이 프로그램은 각 평가자들의 관대함 정도, 그리고 특정 문항이나 평가 영역에 관한 편향적 평가 경향을 확률적으로 추적하여 평가자 평가 특성과 평가지 타당화 연구에 도움을 제공할 수 있다. 이 Rasch 모형에 근거한 FACETS 프로그램은 표현영역 절대평가에서 신뢰성있는 평가 도구로 인정받고 있다 (North & Schneider, 1998; Weigle, 1998; 강석한, 안현기, 2014).

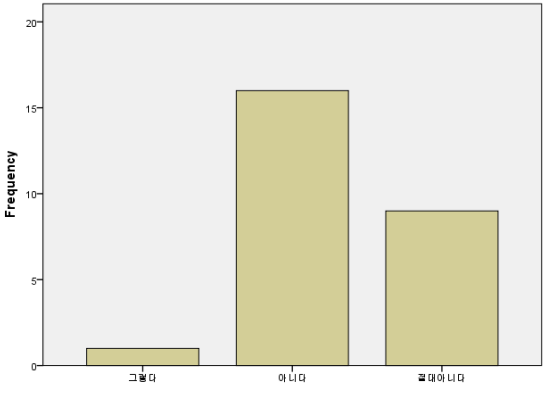
4. 연구 결과

4.1. 설문지 분석

피평가자들은 평가후에 자신의 평가가 피평가자의 정보(모국어 및 한국어 수준)이 평가에 영향을 미쳤는지를 설문조사하였다. 평가자간 신뢰도를 측정한 결과 $r(15) = 0.967$, $p < .0001$로 매우 유의미하게 나타났다. 그 내용은 다음과 같다.

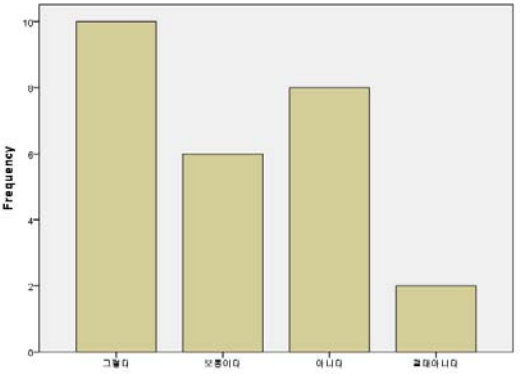
(1) 피평가자의 모국어 정보 공개가 본인의 평가에 영향을 미쳤는가?
(2) 피평가자의 한국어 능력 등급 공개가 평가에 영향을 미쳤는가?

첫 번째 질문인 피평가자의 모국어가 본인의 평가에 영향을 미쳤는가의 질문에 '그렇다'가 3.8%, '아니다'가 61.5%, '절대 아니다'가 34.6%로 95%의 이상의 평가자가 피평가자의 모국어와는 관계없이 평가에 응했다고 응답하고 있다.



<그림 1> 피평가자의 모국어 여부에 의한 평가자의 평가영향

두 번째 항목인 피평가자의 한국어 능력 등급이 평가에 영향을 미쳤는가에 대한 설문을 분석하였다. 그 결과, ‘그렇다’ 38.5%, ‘보통이다’ 23.1%, ‘아니다’ 30.8%, ‘절대 아니다’ 7.7%로 응답을 했다. 이는 피평가자의 수준이 평가에 영향을 미친 정도가 모국어보다 더 많이 나타났다는 점을 설문지는 보여주고 있다. 그럼에도 불구하고 약 반수 이상은 평가자 자신이 공정하게 평가하고 있다고 판단하고 있다.



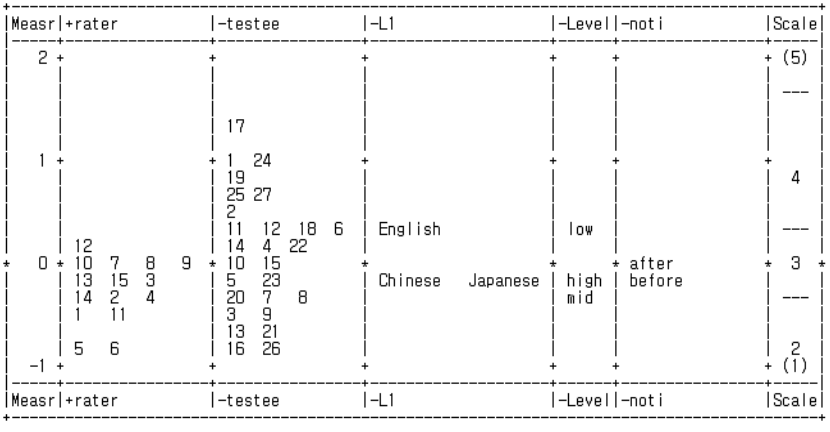
<그림 2> 피평가자의 한국어 능력 급수에 의한 평가자의 평가영향

4.2. 평가 분석

본 사전 연구는 전반적인 분야에서 설정된 5가지 단면(평가자 * 피평가자 * 피평가자 모국어 * 피평가자 수준 * 사전정보 공지여부)의 정보를 분석하였다. 관대화 오류를 측정하는 방법은 평균, 변량분석, 왜도(skewness)등을 이용하는 방법이 있다. 이들중 측정요소의 평균값을 평가척도의 값과 비교하는 방법이 가장 많이 이용되는데, 이 방법은 측정요소의 평균점수가 평가척도의 중간점을 넘으면 관대화로 간주하고 중간점보다 낮으면 비관대화(non-leniency), 즉 엄격화로 간주하는 것이다.

본 연구에서는 평균값과 평균척도의 중간값을 이용하여 분석하였다. <표 1>은 15명의 평가자의 한국어 말하기를 평가한 도표의 전반적인 모습이다.

<표 1> 평가 분포도



위 도표에 의하면 평가자들의 평가는 0.17 logit에서 -0.95 logit으로 비교적 균등한 분포를 보이고 있지만, 전반적으로 관대한 평가가 이루어지고 있다고 볼 수 있다. 그러나 5번과 6번 평가자들은 -0.93 logit에 위치함으로서 다른 평가자들보다 매우 엄격한 평가를 시도하고 있다. 피평가자들의 점수는 대부분 0.54 logit에서 -0.88 logit사이로 비교적 균등한 엄격도(severity)를 보여주고 있으며, 평균 2.73점(5점 만점, 표준편차 1.25점)을 획득하고 있다. 2차례에 걸친 종합 평가(정보공개전과 후)를 전반적으로 살펴보면, 피평가자의 모국어 변수 부분에서 영어 모국어 대상으로 한 한국어 말하기 평가에서 가장 엄격한 점수를 받고 있으며(0.26 logit), 반면에 일본어 모국어(-0.18 logit)와 중국어 모국어(-0.11 logit)들은 상대적으로 비교적 후한 점수를 받고 있음을 알 수 있다. 이는 수치

그대로 해석하면 일본어 및 중국어 학습자들이 한국어 말하기 능력이 영어 출신 화자들보다 더 뛰어나다는 점을 의미한다.

또한, 피평가자들의 수준에 의한 평가에서는 예상대로 초급수준이 0.38 logit으로 상대적으로 엄격한 평가를 받는 반면, 중급은 -0.26 logit, 고급은 -0.13 logit으로 상대적으로 후한 점수를 받고 있다. 모국어와 수준을 제시하기 전과 후를 비교한 항목에서는 0.08 logit과 -0.08 logit사이에 위치함으로서 평균점수 및 점수 분포에는 약간의 차이를 보였다. 한국어 말하기 평가에서 각 평가자들의 평가 형태에 대하여 유사점도 발견되지만, 일부 측정 영역에서 상당히 유의미한 차이점을 나타내고 있다. 다음 도표의 각 측정 항목은 채점 경향을 잘 보여주고 있다.

<표 2> 평가 엄격성 수준과 적합도

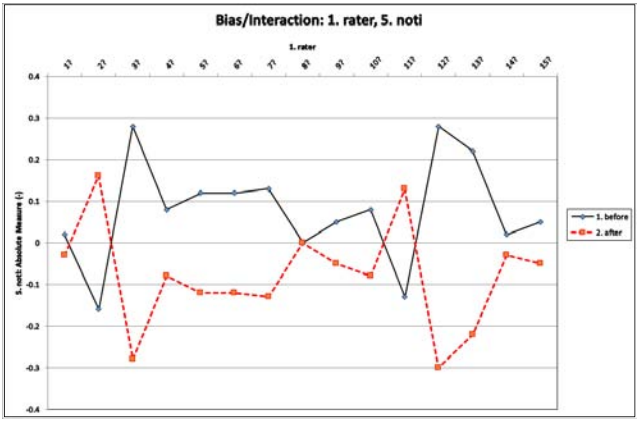
평가자	로짓 (logit)	표준 오차	내적합 제곱평균	내적합 Z 검정	외적합 제곱평균	외적합 Z 검정
1	-0.46	0.17	0.78	-1.0	0.77	-1.0
2	-0.37	0.17	0.83	-0.8	0.80	-1.0
3	-0.16	0.17	0.88	-.07	0.86	-0.6
4	-0.34	0.16	0.95	-0.1	1.10	0.5
5	-0.76	0.17	0.85	-0.6	0.91	-0.3
6	-0.77	0.17	0.85	-0.6	0.78	-0.8
7	-0.03	0.17	0.90	-0.3	0.97	-0.1
8	-0.02	0.16	1.51	2.2	1.55	2.5
9	-0.02	0.16	1.11	0.5	1.09	0.5
10	0.01	0.17	1.01	0.1	1.01	0.0
11	-0.46	0.16	0.93	-0.2	0.88	-0.4
12	0.22	0.11	1.21	1.0	1.17	0.8
13	-0.11	0.17	1.25	1.1	1.17	0.8
14	-0.41	0.16	1.14	0.7	1.05	0.2
15	-0.17	0.16	0.95	-0.1	0.89	-0.4

<표 2>는 평가자들의 개별적인 logit 점수 및 표준 오차, 내적합도 지수, 외적합도 지수를 제시한 것이다. 적합도 통계치 분석은 이론적인 Rasch 모델에 의해 기대되는 점수와 실제 관찰된 점수를 비교하는 것으로 여기서는 내적합도 및 외적합도 지수가 모형 적합도 통계치로 제공된다. 외적합도 지수는 예상할 수 없는 평가형태에 민감한 지수이다. 따라서 소수의 예상치 못한 평가반응 형태로 인하여 부적합으로 판단되는 것을 방지하기 위하여 내적합지수가 주로 사용된다. 일반적으로 1을 기준으로 적합도가 0.75이하는 과적합(overfit), 1.3이상은 부적합(misfit)으로 산정한다(McNamara, 1996). 혹은 좀 더 진보적으로 해석하여 0.5이하를 과적합, 1.5이상을 부적합으로 산정하기도 한다(신동일, 2006). 평가자들의 일관성을 판정할 때는 외적합 값보다는 좀 더 안정적이고 지속성을 지닌 내적합값에 따라 판정한다(최숙기, 2011; 강석한・안현기, 2014).

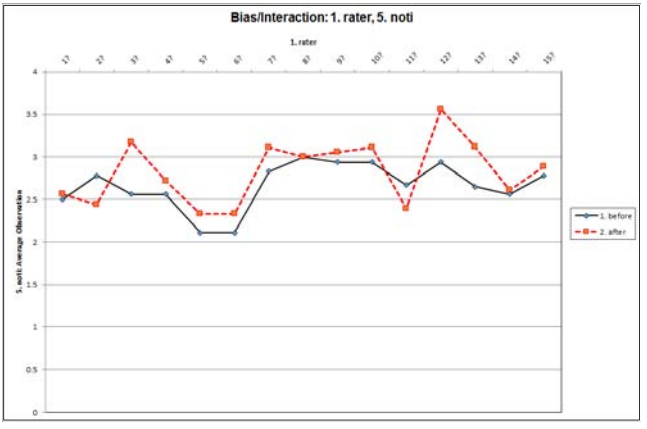
본 연구에서는 내적합제곱 평균이 거의 모든 평가자가 0.75에서 1.3사이에 위치하고 있다. 즉 보수적으로 해석한다고 하더라도 평가자들은 매우 일관된 평가를 하고 있다고 판단된다. 이러한 분포에 대하여 Chi-square 값은 42.1($P < 0.05$), 분리 신뢰도 R=.67로 나타나 통계적으로 평가자들의 측정 오차가 존재함을 보여주고 있다. 평가자 분리 신뢰도는 Cronbach α와 같이 0에서 1의 범위를 가지며, 1에 가까울수록 해당 평가항목들이 평가자의 평가 신뢰도 분포를 잘 설명해주는 것을 의미한다(Wright & Master, 1982). 평가자 분리신뢰도 값 0.67은 평가항목간 난이도차이에 의하여 관찰 변량이 발생하고 있다고 볼 수도 있지만, 어느 정도는 측정 오차가 존재함을 의미한다.

구체적인 고정 오류인자를 찾기 위하여 편파성(bias) 문항 분석이 동원되었다. 이 분석은 평가도구가 측정하고자 하는 구인이외에 다른 이유 때문에 동일한 능력을 가진 수험자가 다른 결과를 얻게 된다면 이는 평가 집단, 준거기준, 적용등에 의하여 점수가 달라진다는 것을 의미한다. 이 분석을 위하여 Mantel-Haenszel 방식과 로지스틱 회귀분석(logistic regression), 혹

은 문항반응 이론을 이용하여 평가자들로부터 추정된 문항모수치나 문항 특성곡선을 이용하여 분석하였다. 즉, 평가의 특정 국면(facet)들 사이에 어떤 상호 작용이 발생하는지를 분석하는 것이다. 이 분석은 서로 다른 로짓값의 차이가 Rasch 모델의 기댓값이 통계적으로 유의할 경우 분석되는 것이다(장소영, 신동일, 2009; 최숙기, 2011; 강석한, 안현기, 2014). 편향도 분석은 한국인 평가 집단의 일부 평가자들이 편향적인 평가를 수행하고 있음을 보여주고 있다. 편향 분석은 평가자 집단과 평가 과업 및 영역간에 잠재적인 상호 작용을 조사하는 것이다(신동일, 2006; Kim, 2009). 편향 분석은 Z 검정에 의하여 판별되는데, 만약 이 검정값이 -2와 2사이에 위치하면 평가자는 편향 없이 평가한다고 판단할 수 있다. 만약 -2이하로 내려가면 과적합 평가 판정이 내려지는데, 이는 이 평가자가 동일 집단의 다른 평가자들에 비하여 관대한 평가를 내린다는 의미이다. 본 연구에서는 모든 평가자들이 -2와 2사이에 분포하고 있어서 전반적으로 평가의 일관성을 유지하고 있다고 할 수 있다.



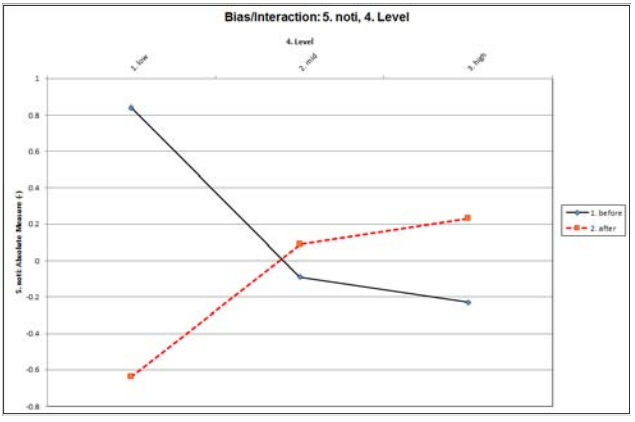
(가) 로짓 함수



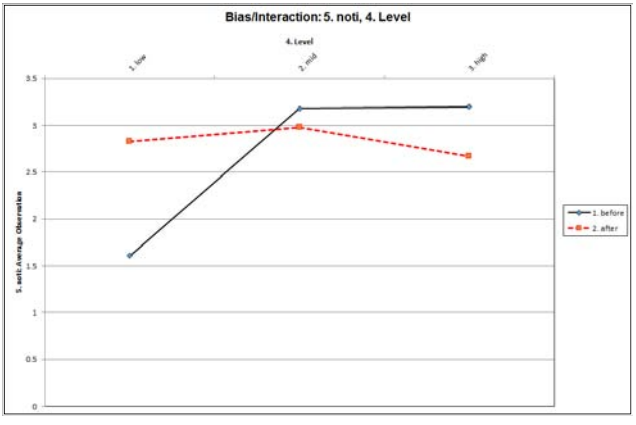
(나) 평가 점수

<그림 3> 정보공개와 평가자와의 편향도 상관관계
(*실선: 정보공개전, 점선: 공개후)

본 실험에서 피평가자의 모국어와 수준별로 정보 공개전의 평가(평균 2.46)와 후의 평가(평균 2.83)가 통계적으로 유의미하게 달랐다(χ^2=4.1, $p < 0.05$). 위 편파성 조사에서도 정보 공개전이 상대적으로 엄격하게 평가하는 반면에 (0.02 logit), 피평가자들의 정보 공개 후에는 상대적으로 관대하게 평가하고 있음(-0.25 logit)을 엿볼 수 있다. 평가자들의 평균점수도 평가전에는 2.663에서 평가 후에는 2.828로 상승하고 있다. 이 평가중 가장 문제가 될 수 있는 평가자는 3번, 12번 평가자로서 이들은 피평가자들의 정보공개에 영향을 상당히 받고 있다. 3번 평가자는 정보공개 전에는 0.29 logit으로 비교적 엄격하게 평가하는 반면, 평가 후에는 -0.30 logit으로 후하게 평가하고 있다. 12번 평가자도 비슷한 경향을 보이고 있다. <그림 4>는 정보공개시점(전과 후)과 각 수준별 평가에 미치는 영향을 살펴보았다.



(가) 로짓 함수



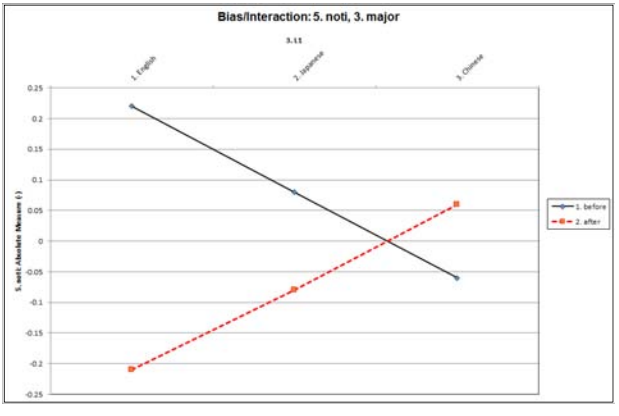
(나) 평가 점수

<그림 4> 정보공개여부와 수준과의 편향도
(*실선: 정보공개전, 점선: 공개후)

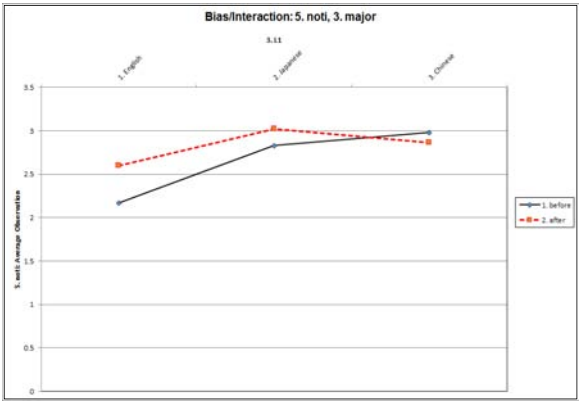
피평가자들의 정보 공개 전 점수(평균 2.459)와 후의 평가점수(평균 2.829)가 수준별로 매우 유의미하게 달랐다 (χ^2=82.3, $p < 0.001$). 초급 수

준의 영어모국어 한국어 학습자들이 정보공개 전에는 0.82 logit 으로 매우 엄격한 평가를 받는 반면에, 수준이 공개된 후에는 -0.62로 급격하게 평가가 관대해지고 있다. 반면에 중급 및 고급 학습자들은 정보공개 전에는 0.12와 0.22로 비교적 엄격한 평가를 받았지만, 수준이 공개된 후에는 -0.13과 -0.21로 비교적 관대한 평가를 받았고, 상대적으로 편차는 초급에 비해 작았다. 한국인 평가자들이 초급자들에게 관대한 평가 경향을 보이는 것은 관대화 오류를 범하고 있음을 보여주는 증거이다.

흥미롭게도 피평가자의 정보공개 전(평균 2.46)과 후의 평가(평균 2.83)가 모국어변수에 의하여 통계적으로 유의미하게 나타나지 않았지만(χ^2= 6.9, $p > 0.05$), 영어모국어 학습자에게는 유의미한 선호도를 나타냈다.



(가) 로짓 함수



(나) 평가 점수

<그림 5> 정보공개여부와 한국어 학습자 모국어와의 편향도
(*실선: 정보공개전, 점선: 공개후)

전반적으로 일본어 모국어 학습자와 중국어 모국어 학습자에 대하여는 정보공개전과 후와 별로 다르게 나타나지 않았지만, 명백히 영어 모국어 학습자에 대하여는 정보공개전이 0.23 logit으로 엄격하게 평가를 하고 있지만, 정보공개 후에는 -0.22 logit으로 평가가 상당히 관대하게 진행되고 있음을 엿볼 수 있다. 이는 한국인 평가자들이 영어 모국어 학습자들에 대하여 매우 우호적인 시각을 지니고 있음을 보여주는 후광효과가 존재함을 의미한다. 본 연구를 통하여 한국인 평가자들에게는 몇 가지 고정오류가 존재함을 살펴볼 수 있었다. 한국인 영어 평가자들에게는 피평가자들의 수준별, 모국어별로 평가행태에 영향을 미친다는 것을 보여준다.

5. 결론

본 연구는 한국어 말하기 평가에서 고정 오류인자를 찾아내어, 좀 더 공정한 평가가 이루어질 수 있도록 평가형식과 방법을 찾으려는 목적을 지니고 있다. 설문조사에서는 평가자들이 대부분 공정하게 평가하고 있다고 답했지만, 편파성 조사에서 정보 공개전이 상대적으로 엄격하게 평가하는 반면에, 정보 공개 후에는 상대적으로 관대하게 평가하고 있었다. 평가자들의 평균점수도 정보 공개 전에는 2.429에서 공개 후에는 2.829로 상승하고 있다(5점 만점).

피평가자의 수준은 평가자의 평가에 매우 큰 영향을 미친다. 초급 수준의 영어모국어 한국어 학습자들이 정보 공개 전에는 0.82 logit으로 매우 엄격한 평가를 받는 반면에, 수준이 공개된 후에는 -0.62로 급격하게 평가가 관대해지고 있다. 반면에 중급 및 고급 학습자들은 정보공개 전에는 0.12와 0.22로 비교적 엄격한 평가를 받았지만, 수준이 공개된 후에는 -0.13과 -0.21로 비교적 관대한 평가를 받았고, 상대적으로 편차는 초급에 비해 작았다. 한국어 평가자들이 초급자들에게 관대한 평가 경향을 보이는 것은 관대화 오류를 범하고 있음을 보여주는 증거이다.

이러한 평가오류를 제거하기 위한 평가자 교육 및 훈련은 매우 중요하다. Bernadin(1978)은 평가 강의 훈련을 통하여 평가자 내부의 토론을 유도한 평가자 훈련 방법이 가장 신뢰성이 있는 평가결과를 제공했다고 주장했다. 또한 후광효과가 발생하는 원인을 평가자 내부적인 요인과 외부적인 요인으로 나누어 설명하면서, 평가자 교육 후에 평가자들이 평가자오류에 대하여 얼마나 이해하고 있는지 설문조사를 실시하였다. 그 결과 오류 유형에 대하여 더 많이 이해하고 있는 평가자일수록 보다 신뢰성있는 평가 결과를 제공한다는 것을 발견했다. 이 연구는 고정 오류 인자 확인은 분명히 오류정도 및 빈도를 감소시킬 수 있으며, 이를 위하여 평가자 훈련이 매우 중요하다는 것을 시사한다. 따라서 차후 연구는 평

가자 훈련교육에 대하여 다루어 보고자 한다.

<참고 문헌>

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

김명소, 이광진(2008). 평가형식 및 척도화 방법이 다면평가에서 관대화 오류 및 후광효과에 미치는 영향, <한국심리학회> 21집, 한국심리학회. 201쪽～224쪽.

이호(2011). 귀납적 이분법 척도를 활용한 예비 영어교사의 영어 말하기 평가 신뢰도와 효용성에 관한연구, <한국교육연구> 27집 3권, 한국교육연구회. 215쪽～237쪽.

이호, 김하나(2009). 영작문에서의 대면평가와 익명평가의 비교연구, <영어영문학>, 9집 3권, 한국영어영문학회. 403쪽～427쪽.

신동일(2006).『한국의 영어 평가학』, 서울: 한국문화사.

신동일, 장소영(2002). 후광효과의 이해를 통한 채점 오류 분석 연구, <외국어교육>, 9집 4권, 외국어교육학회. 215쪽～232쪽.

신상근(2010).『외국어 평가의 이론과 실제』, 서울:한국문화사.

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

최숙기(2011). Rasch 모형을 활용한 요약문 평가 준거 개발및 타당도 분석, 25호, <독서연구>, 한국독서학회. 415쪽～445쪽.

Bachman, F. & Palmer, S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: OUP.

Bardo, J. Yeager, S. & Klingsporn, J. (1982). Preliminary assessment of format-specific central tendency and leniency error insummated rating scales. *Perceptual and Motor Skills*. 54. 227-234.

Bechger, T., Maris, G. & Hsiao, Y. (2010). Detecting halo effects in performance-based examination. *Applied Psychological Measurement*. 34.8. 607-619.

Bernardin, J., Alvares, A. & Cranny, J. (1976). A recomparison of behavioral expectation scales to summated scales. *Journal of Applied Psychology*. 61. 564-570.

Bernardin, J. (1978). Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology*. 63.3. 301-308.

Bretz, D. Milkovich, T. & Read, W. (1992). The current state of performance appraisal

research and practice: Concerns, directions and implications. *Journal of Management.* 18(2). 321-322.

Fisicaro, A. (1988). A re-examination of the relation between halo error and accuracy. *Journal of Applied Psychology*. 73. 239-244.

Jones, L. & Fletcher, C. (2002). Self-assessment in a selection situation: An evaluation of different measurement approaches. *Journal of occupational and organizational Psychology*. 75. 145-161.

Kim, J-T. (2015). Overview of criterion-referenced test. *2015 KATE SIG conference*. 251-254.

Kunnan, A. (2008). *Large-scale language assessments.* New York: Springer Science.

Lee, J. & Kim,J. (2014). The effect of task types and task difficulty on Korean college students' speaking test performance. *2014 Kelta conference handbook.* 31-32.

Lincare, M. (2005). *A user's guide facets: Rasch-model computer program*. Retrieved from www.winsteps.com.

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

Luoma, S. (2004). *Assessing speaking.* Cambridge: Cambridge University Press.

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

Murphy, R. & Balzer, K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*. 74.4. 619-624.

North, B. & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*. 15(2). 217-262.

Pulakos, E. 1984. A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology.* 69(4). 581-588.

Shrock, A. & Coscarelli, C.(2000). *Criterion-referenced test development.* Washington: International society for performance improvement.

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

강석한(Kang Seokhan)
건국대학교 글로컬 캠퍼스 교양교육원
380-701 충북 충주시 중원대로 268
전화번호: 82-43-840-3395
전자우편: kangs45@kku.ac.kr

접수일자: 2016년 4월 20일
심사(수정)일자: 2016년 5월 31일
게재확정: 2016년 6월 15일