『이중언어학』 제55호(2014.6.30). 二重言語學會

외국인 한국어 말하기 시험의 평가자 요소가 채점에 미치는 영향

강석한* · 안현기**

Abstract

Seokhan Kang & Hyunkee Ahn. 2014. 6. 30. **A Study of Korean Raters' Characteristics for Foreign Speakers' Korean Oral Performance.** *Bilingual Research 55*, 1-29. The study investigated the effects of raters' experience on the validity and reliability for foreign speakers' Korean speaking assessment. Twelve professional Korean and twelve non-professional Korean raters assessed twenty-seven Korean learners' oral performance. For the quantitative analysis the FACETS program with multi-facets Rasch model was carried out. The results are as follows: (1) professional raters tended to assess strictly rather than non-professional raters, (2) the strict assessment was realized on the open-ended response task, while the relaxed assessment was found on the opinion response and story-telling response tasks, and (3) both groups evaluated strictly on the grammar section, while they gave higher scores to the questions on the cohesive discourse section. The study also found that some professional raters caused problems of reliability and validity on their assessment. Other non-professional raters, on the other hand, tended to show over-fitted or under-fitted assessment. The result implies that raters should take the assessment training before participating in the real assessment.**(Konkuk University·Seoul National University)**

【Key words】 발화 평가(speaking assessment), 평가자(raters) 제2언어평가(second language assessment), 기준(criteria), 과업(tasks), 영역(area), 한국어 말하기(Korean speaking)

* 제1저자
** 교신저자

1. 서론

현행 한국어 능력을 평가하기 위한 TOPIK, KBS 한국어 능력 시험, ToKL 등의 시험들은 말하기 평가를 시행하지 않고 있다. 이러한 말하기 영역 제외는 당연히 수험자의 한국어 능력을 정확히 나타낸다고 할 수 없다. 근래 들어 한국어 말하기 영역 도입에 대한 논의가 활발한 가운데, 신뢰성과 타당도를 지닌 평가자의 평가 능력에 관심이 쏠리고 있다. 본고는 한국어 평가자의 평가에서 평가자 자질요소가 주관적 평가에 어떤 영향을 미치는지를 살펴보고자 한다.

주로 한국어 말하기 평가 연구는 외국의 말하기 평가 기준을 참고로 하여 한국어교육의 기준을 세우는 연구들로 시작되었다. 노대규(1983)는 FSI(U.S. Department of State Foreign Service Institute)의 등급을 통해 말하기 숙달도 평가의 기준과 방법을 제시했다. 여기서 9개의 등급과 통사 부분의 정확성, 음운 부분의 자연성, 발화 속도의 유창성, 의미 부분의 다양성이라는 평가 범주를 제시했다. 김정숙 외(1993)에서는 ACTFL (American Council on the Foreign Language)의 숙달도 판단 기준을 근거로 한국어 말하기 숙달도 기준을 마련하는데, 한국어의 특수성이 반영된 객관적 말하기 평가 기준의 필요성을 밝히며 Canale과 Swain의 의사소통 모델 중 '전략적 능력'을 제외한 세 능력으로 초급-중급-고급의 각 평가 범주별 특성을 제시했다. 이 연구에서는 학습 단계별로 평가 범주의 상대적 비중이 달라지며 단계별 교육의 필요 하다고 주장했다.

정광 외(1994)는 ACTFL 숙달도 판단 기준을 근거로 숙달도 평가를 위한 등급 기준(초-중-고-최상급)과 말하기 평가 단계(예비-수준 검색-정밀 검사-수합)를 제안했다. 전은주(1997)는 ACTFL의 기준과 평가 방법이 타당성을 갖을 수 있는지를 검토하여 일반적인 한국어 말하기 능력 평가 범주를 제시하였는데, 이런 범주에는 문법 능력, 어휘 능력, 발음 능력, 구성력, 사회언어학적 능력, 의사소통적 전략과 상호작용, 과제 수행력등이 포

함된다. 박성원(2002)은 ACTFL OPI(Oral Proficiency Interview)를 한국어 교육에 적용할 때 나타날 수 있는 문제점을 인터뷰 자료 분석을 통해 찾고, 이를 통해 동료 수험자의 대화 참여와 평가 참여, 다양한 유형의 상호작용 과제 개발의 필요성을 제안했다. 김정숙 외(2007)는 기존의 한국어 말하기 평가 연구들이 ACTFL을 중심으로 표준화된 평가에 초점이 맞춰졌음에 대한 반성적 연구로 말하기 평가 개발 과정을 상세하게 들여다보는 시도를 하였다. 또한 전나영 외(2007)는 기존 영어 교육의 시험 유형을 비교 분석하여, 표준화된 한국어 말하기 시험에 적용 가능한 평가 문항 유형과 기준을 예를 들어 제시했다.

일반적으로 평가자 개인 역량이 평가의 신뢰성과 타당도에 영향을 미친다고 알려져 있다. 일반적인 언어 평가에서의 평가자의 특성에 대한 연구들은 평가자 역량 요소 변수가 신뢰도와 타당도에 상당한 차이를 가지고 온다는 점을 지적하고 있다 (Hadden 1991; Chalhoub-Deville 1995; Luoma, 2004). 기존 연구들을 요약하면, 수험자의 발화를 전부 듣고 그 전체적 인상을 중심으로 하나의 등급으로 채점하는 총괄 평가(holistic marking)에서는 평가자 사이에 유사성이 발견되지만, 말하기 능력의 구성 요소를 하나하나 분석적으로 고려하여 채점하는 형태인 세부 평가(analytic marking) 방법에는 전문가와 비전문가 평가 집단사이에 상당한 차이점이 발견된다고 보고하고 있다.

우선, 총괄평가는 비전문가와 전문평가자 사이에 높은 상관관계를 나타난다고 주장하였다(Luoma, 2004). 이석재·박전규(2003)는 음성학 훈련을 받은 한국인 전문가 그룹, 비전문가 그룹, 원어민 그룹 간의 평가와 자동 인식기를 기반으로 한 기계식 평가 사이의 상관관계를 연구하였다. 연구 결과는, 한국인 평가자 사이의 상관관계가 가장 높고 (r=0.98), 전문가/비전문가 한국인 평가자와 원어민 평가자와의 상관관계 (r=0.92)도 높았으며, 인간 평가와 기계 평가 사이에도 유의한 수준의 상관관계가 있었다(r=0.72). 이 연구의 초점은 기계식 평가의 타당도를 살펴보는 연구

이지만, 부수적으로 전문가와 비전문가의 총괄적인 평가에서도 큰 차이가 없다는 점을 보여준다는 점에서 의의가 있다.

그러나 세부 영역에서는 평가자 역량 변수에 의하여 평가의 신뢰도와 타당도가 달라진다고 보고하고 있다 (Hadden, 1991; Chalhoub-Deville, 1995). 일반적으로 전문가 집단은 비전문가 집단에 비하여 엄격한 평가를 한다고 알려져 있지만, 모든 연구가 이결과를 따르는 것은 아니다. 이는 아마도 평가자의 배경 언어, 샘플수, 평가 방식의 다양성등의 변수에 의하여 결과가 영향을 받는다. Hadden (1991)은 영어 전문가와 비전문가 집단이 중국인 영어 발화에 대하여 평가를 비교 연구하였다. 이 결과에 의하면, 언어 능력 측면에서는 전문가 집단이 더 엄격한 채점을 하는 경향을 보이지만, 이해도, 수용성, 개성, 비언어적 능력 측면에서는 두 집단 사이에서는 큰 차이가 없다고 보고하고 있다. Chalhoub-Deville (1995)은 배경언어와 전문가 요소의 두 매개변수를 이용하여, 아랍인 미국 거주 전문가 집단, 아랍인 미국 거주 비전문가 집단, 그리고 아랍 거주 비전문가 집단을 대상으로 연구하였다. 그 결과 전문가 집단에서는 언어 측면보다는 담화 과업 측면에서 창의성과 정보 수용성에 더 평가 가중치를 두는 경향을 보였다.

아쉽게도 현재까지 한국어 말하기 평가에서 평가자 개인 역량 요소가 어떤 차이를 가져오는지에 대한 연구가 없었다. 지금까지 한국어 말하기 평가는 개인 역량 요소를 간관한 측면이 많았다. 본 연구는 국어 말하기 평가의 신뢰도와 타당도를 검증하기 위하여 채점자 역량에 의한 평가 내용 과업(구조 확정형과 개방식 과업) 요인이 말하기 평가 영역에 어떤 영향을 미치는 지를 살펴볼 것이다. 연구 목적은 우리나라의 한국어 말하기 평가에서, 한국인 평가자에 의한 외국인 한국어 말하기 평가가 신뢰도와 타당도를 담보할 수 있는지를 알아보고자 하는 것이다. 따라서 본 연구는 다음의 두 가지 연구 질문에 대한 답을 구하기 위하여 노력하였다.

1. 각 평가자는 평가 양상이 평가영역별, 과업별로 어떻게 같게/다르게 나타나는가?
2. 전문가 평가자는 비전문가 집단에 비하여 내적 일관성과 엄격성에서 각 영역별로 동일한/다른 특징을 나타내는가?

이 두 가지 질문에 답을 구하기 위하여 반직접 말하기 평가방식을 이용하여 평가를 측정하였다. 이 방식은 여러 단점에도 불구하고 대량으로 동시에 실시할 수 있고, 실제적인 의사소통 속성을 가지고 있으며, 채점 신뢰성을 향상 시킬 수 있고, 비원어민 수험자에 대해서도 타당한 시험으로 검증된 것이다(신동일, 2006).

2. 한국어 말하기 평가 요건

2.1. 말하기 능력의 구성요소

세부 평가는 다양한 세부 영역에 걸쳐서 학습자의 능력을 종합적으로 측정하며, 높은 신뢰도와 구인 타당도를 지닌다. 이런 세부 평가 방식은 학습자의 정보를 자세하게 알려주기에 진단 평가적 성격에 적합하다(Knoch, 2009; 이호·김혜영·이진화, 2011). 세계의 여러 공인 평가들(ACTFL, FCE, IELTS, ASLPR, CPE, CAE)의 세부 채점 척도에서 공통적으로 고려하고 있는 요소들중에서 중요한 것들을 간추려 보면 정확성(accuracy), 어휘 범위(vocabulary range), 적절성(appropricacy), 유창성(fluency), 상호작용(interaction), 발음 명료성(clear pronunciation)등으로 나누어 볼 수 있다.

국내의 한국어 능력 평가는 일반적으로 외국의 평가를 준용하는 경우가 많다. 현재 대표적인 한국어 능력 시험인 TOPIK 이나 KBS 한국어 능력시험에서는 한국어 말하기 영역을 채택하고 있지 않지만, 한국어 교

육기관을 갖고 있는 각 대학은 각기 사정에 맞는 독특한 말하기 평가 영역을 구축하고 있다. 정확성, 유창성, 내용, 준비, 태도 (A 대), 상호작용 능력, 문법, 발음, 유창성 (B 대), 정확성, 유창성, 완성도 (C 대), 정확성, 유창성, 내용, 발표력 (D 대)등인데, 주로 음성 영역(발음, 유창성), 언어 능력 (문법, 어휘), 담화 능력(상호 작용, 완성도), 비 언어능력 (준비, 태도)등을 평가한다.

본 연구에서는 과제 실현 능력, 의사소통 능력, 대화의 적절성, 언어의 풍부함, 발음 명료성, 유창성, 평가자로부터의 독립성등이 언어 말하기 평가에 중요하다는 연구결과를 바탕으로(이영식, 2006; 이은혜·김성아, 2007) 문법의 정확성(grammar usage), 어휘 선정의 적절성(vocabulary usage), 담화 일관성(cohesive discourse), 발음의 명료성(clear pronunciation)과 전반적 인상(holistic impression)을 주요한 연구 주제로 선정하였다. 다음은 구체적인 평가 영역이다.

(1) 발음 영역: 발음 영역에서는 유창성과 발음 명료성을 평가한다. 유창성이란 말을 할 때 더듬거나 망설이지 않고 자연스럽고 자신감 있게 말하는 능력을 지칭한다. 유창성은 문법사용의 정확성까지 포함하는 개념이다. 발음 명료성 영역에서 의미하는 발음에는 개별 분절음 단어의 발음 영역인 분절 요소와 억양이나 강세, 악센트 등의 초분절 요소로 대별된다. 개별적인 분절 요소는 매우 중요하지만, 억양이 포함된 초분절 요소가 더 중요하다고 할 수 있다. 따라서 평가에서 발음 부분은 의미를 실어서 전달하는 매체인 발음이 전체적으로 의미를 효과적으로 전달하느냐 못하느냐에 따라 평가를 한다.

(2) 문법의 정확성: 문법 사용의 정확성을 측정하는 항목이다. 용언의 형태소, 시제, 연결어미, 수식어, 접속부사등을 중점적으로 살핀다. 또한 피동 및 사동 표현과 높임말 표현에 유의한다. 빈도가 높은 접미사와 접두사, 다양한 종결어미, 보조 동사 등을 이해하고 구

사할 수 있다. 상대방에 맞추어 존댓말과 반말뿐만 아니라 극존칭 및 하오체, 하게체를 이해할 수 있다. 문법과 관련해서 발생되는 문제가 거의 없이 문법을 활용할 수 있다.

(3) 어휘선정의 적절성: 한글 어휘의 사용 빈도를 측정하는 항목이다. 정확한 단어나 구문을 찾아내어 큰 어려움 없이 한글로 자기표현을 할수있느냐를 알아보는 것이다. 이 영역에서는 사회 각 분야의 기초적인 어휘를 이해하고 사용할 수 있으며 전문적인 어휘도 설명하면 이해할 수 있다. 또한 신문이나 방송등의 시사관련 어휘를 이해하며 사회생활 관련 어휘를 별 어려움 없이 사용할 수 있다.

(4) 담화 일관성: 전체적인 의사소통 활동이 원활하고 효과적이다. 문장과 문장사이, 단락과 단락 사이에 얼마나 조직적으로 발화가 일어나고 있는가를 평가한다. 의견제시나 아이디어 전개에 연결사의 사용이 적절하여 청자가 이해하는데 문제가 없음을 알아보는 항목이다. 따라서 이 영역에서는 교양적인 내용의 담화가 가능하고 매우 전문적 주제가 아니라면 어떤 주제에 대한 토론이나 의견 교환등이 가능하다.

(5) 전반적 능력: 발화자가 발음, 어휘선정, 담화 일관성을 적절하게 조합하여 과제를 유기적으로 해결한다. 또한 특정 의견을 지지하기 위하여 구체적인 예시나 자료를 활용하고, 추상적인 논제를 펼 수 있는지를 측정한다.

이 다섯 가지 평가영역을 외국인 한국어 수험자의 한국어 말하기 평가에 적용하였다.

2.2. 말하기 평가 과업

말하기 질문 형태는 크게 개방형 종료 과업(open-ended)및 구조 확정형 과업(structured task)으로 크게 나눌 수 있다(Luoma, 2004). 개방형 종

료 과업은 다양한 대답을 수용하는 여지를 주는 방법이고, 반면에 구조 확정형 과업은 시험자가 제시해야할 내용이 정해져 있다. 일반적으로 말하기 시험은 수준에 따라서 이 두 종류의 과업을 적절하게 혼합하는 방식을 취한다. 예를 들어 ELPA 말하기 평가 문항에서는 초급과 중급에서는 구조 확정형이 압도적으로 많으며, 중급과 고급에서는 개방형 종료 과업이 많다. 구체적으로 이 시험의 초급은 주어진 문장을 보고 따라 말하기, 상황을 보여주는 배경이 제시된 질문에 답하기, 중급에서도 자료를 보고 경험이나 계획을 말하기등의 구조 확정형 문제를 유지한다. 그러나 중, 고급에서는 주로 개방형 문제들이 출제되는데, 중급에서는 그림 설명하기가 대표적이며, 고급수준에서는 도표, 그래프 설명하기, 의견 말하기, 설득, 대안 제시하기가 출제된다.

본 연구는 고급반 수험자들의 심도있는 한국어 말하기 능력을 측정하기 위하여 개방형 종료 과업 방식을 채택하였다. 이 개방형 종료 시험 방식은 과제 수행 완성에 여러 가지 방식을 허용하기 때문에, 상대적으로 긴 제시 시간과 광범위한 대답영역을 요구한다. 이 평가 방법에는 묘사, 기술, 지시, 비교, 설명, 정의, 예측등의 과업이 그림이나 지시문, 예시, 주제등을 통하여 수험자의 반응을 평가한다. 예를 들어, Oral Proficiency Interview (e.g., Luoma, 2004) 평가에서는 묘사식 방식을 이용하는데, 피험자는 숙모의 집이나, 조카등에 대한 친숙한 주제에 대하여 묘사하도록 요청을 받는다. 혹은 그림제시를 통하여 설명을 요구할 수도 있고, 주어진 질문에 응답할 수도 있다. 다음은 본 연구에 이용된 평가과업들이다.

(1) 개방형 반응 과업 (Open ended response task)

폐쇄형 질문 과업에 비하여 비교적 긴 시간에 자기 생각을 발표하는 과업이다. 질문은 주로 의문사로 구성된다. 개방형 질문은 선택지나 항목들을 미리 준비하거나 답을 일정한 양으로 제한하지 않고 응답자가 자신의 견해나 태도를 자유롭게 표현할 수 있도록 구성된 질문을 말한다.

다음은 본 연구에 사용된 질문이다.

한국에서 명절때는 무엇을 합니까? (제한 시간 1분)

(2) 지시 반응 문항 (Directed response task)

특정 과제를 글이나 말로 지시하고, 이에 대한 반응을 설명하는 과업이다. 지시 반응 과업은 반드시 질문지에 특정한 (지시된) 질문이나 아이디어, 주제가 제시되며, 수험자를 이를 답변에 방영시켜야 한다. 다음은 본 연구에 사용된 지시문이다.

당신이 가고 싶은 장소를 설명해 주세요.

대답에 다음의 항목이 들어가야 합니다.

어디에 위치하고 있습니까?
왜 이 장소를 선택했습니까?
무엇을 하고 싶습니까?

(제한기간 1분)

(3) 그림 / 사진 설명 과업 (Picture-cued story telling task)

주어진 그림이나 사진을 정해진 시간내에 설명하는 과업이다. 수험자는 주어진 그림이나 사진에 대하여 설명을 한다. 본 연구에서는 주어진 그림에 대하여 설명하도록 요구하였다. 다음은 본 연구에 사용된 지시문이다.

<그림 1> 사진 설명 과업 예시

주어진 사진내용을 설명하시오 (제한시간 1분).

(4) 의견 제시 문항 (Opinion task)

특정 주제에 대하여 2 – 3 분정도 자기 생각을 말하게 한다. 주제는 상당한 깊이를 지니는 주제들이다. 수험자는 이 주제에 대하여 자신의 의견을 명확히 표현하고 그 이유 및 준거를 제시한다. 다음은 본 연구에 사용된 지시문이다.

당신이 알고 있는 가장 똑똑한 사람은 누구인가요? 이 사람에 대하여 설명하고 왜 그/그녀가 똑똑한 사람인지 설명하시오. (제한시간 2분).

3. 연구의 내용 및 방법

3.1. 수험자 및 평가자 정보

한국어 평가 자료로 사용될 음성 자료는 27명의 외국인 발화자(남성 16명, 여성 11명)의 음성 녹음 이다. 이들의 국적은 미국인 8인, 일본인 7인, 중국인 12인이다. 이들은 평가 당시에 한국어 최고급반에 속한 학생들이였고, 이들 중 8명이 한국어 능력 1급 이상 소지자 이었다. 이들이 다니는 한국어 교육원은 모두 6단계 수준으로 되어 있으며, 각 단계의 승급은 언어의 4가지 영역을 모두 측정하여 평균 60점 이상이면 승급을 하였다. 이들의 한국 체류 기간은 평균 3.4 년 (표준편차 2.7 년), 연령은 23.2 세 (표준편차 2.3 세) 이다. 이들은 이미 최고 단계를 수료한 학생들이 많았으며, 단지 한국어 능숙도를 더 연마한다는 어학연수 목적이외에도 사교적인 의도를 동시에 갖고 있었다. 이 학교에서 제시하는 6단계는 다음과 같은 평가기준을 제시하고 있었다.

(1) 기본 학습 목표

사회생활이나 직장에서 필요한 한국어를 이해하며, 고도의 내용의 한국어 구사가 가능하다. 수준 높은 문장이나 텔레비전, 라디오, 강연등의 시사적인 내용을 충분히 이해하고 문장이나 말로 정확히 전달할 수 있으며, 토의 토론에서 자신의 의견을 정확히 표현할 수 있다.

(2) 어휘

대부분의 일상적 어휘와 전문적 어휘를 구사한다. 그 밖의 어휘도 문맥에 의지하거나 사전을 능숙하게 이용하며 해결한다.

(3) 문장

괴팍한 표현이나 지나치게 빠른 말이 아니면 사실상 거의 이해한다.

(4) 발음

정상적인 발화에서 발음과 관련된 문제가 없다.

말하기 평가에 대한 연구 목적을 사전에 공지하여 서면으로 승낙을 받았다. 발화자들은 서울에 위치한 사립 대학교 소속 한국어 교육원에 소속된 학생들로서, 그들의 직업은 대학원 학생이거나 (중국어, 일본어, 영어), 주한 미군 군무원(영어), 상사원 (일본어, 중국어), 어학원 강사(영어, 일본어)등으로 한국 사회와 문화에 비교적 익숙한 고급반 학생들이다. 이들에 대한 한국어 말하기 평가는 조용한 교실에서 30대의 컴퓨터를 통하여 화면에 문제가 제시되었고, 피험자들이 이 지시에 따라서 컴퓨터 마이크에 발화를 하였고, 이를 녹음 및 녹화하였다.

이 녹음된 말하기 자료는, 약 1주일 후 전문가 12인, 비전문가 12인에 의하여 평가를 수행하였다. 이들은 수험자들의 녹음을 들으면서 개별적으로 평가를 하였다. 그들에게는 사전에 평가지와 기준표가 제시되었다. 평가에 참여한 각 평가자들은 직업, 교육연한, 전공, 학력등을 고려하여 전문가 집단과 비전문가 집단으로 구분하였다. 전문가 평가자들은 대부분 국어 국문학 분야의 박사 학위 소지자들이며, 한국어 교사 자격증 2급 이상 소지자들이다. 또한 이들은 대학에서 한국어 교육을 담당하고, 연령이 34세-46세 정도 (평균 38.2세, 표준편차 6.1세)되는 한국어 강의 경력이 3-9년 정도 (평균 6.5년, 표준편차 3.2년)되는 중견 교/강사들이다.

비전문가 집단은 일반 직장인들이며, 평균 연령은 24.3세 (표준편차 2.6세)이다. 이들의 학력은 고졸 5, 대졸 7명이며, 대학졸업 이상 학력인 경우 전원 비어문계 전공자들이다. 이들은 교육에 종사하여 일반 학생들을 가르쳐본 경험이 전혀 없다. 본 연구자가 취미 동호회에서 섭외를 하여 말

하기 평가를 실시하였다. 평가자들에게는 소정의 평가료가 주어졌다.

3.2. 평가 항목

평가 척도는 평가의 신뢰도와 구인 타당도를 확보하는데 중요한 요인이다 (Lane, 2008; 이호, 김혜영, 이진화, 2011). 평가의 신뢰도는 측정의 일관성을 의미하여, 구인 타당도는 시험점수를 토대로 내릴수 있는 의미성 혹은 적절성과 관계가 있다. 일반적으로 대규모 국제 공인 시험인 경우 대부분 9-12단계 평가방법을 채용하는 경향이 있다. 예를 들어, ACTFL-OPI는 10등급, ISLPR은 12등급, MATE 12등급, KBS 한국어 능력은 8등급으로 구성된다. 그러나, 일부 시험에서는 5-6단계 등급을 이용하는데 이는 평가기관의 규모 및 시험 성격에 따라 달라진다. 이런 유형의 시험에는 TOPIK (6등급), G-TELP (5등급), ICAO (6 등급), FCE (6등급)등이 있다. 본 평가는 연구용으로 제작되었기 때문에 5 단계 방법을 채용했다. 평가 방법은 OPIc(2010)을 참고하였다. 점수는 0에서 4점까지 5단계를 제시하였다.

3.3. 자료 분석 방법

본 연구는 집단별로 다른 배경을 지닌 채점자들의 신뢰도와 타당도 검증을 위하여 문항 반응 이론중 하나인 Rasch 모형에 근거한 컴퓨터 프로그램 FACETS를 적용하였다. 이 프로그램은 각 채점자들의 엄격함이나 관대함 정도, 그리고 특정 문항이나 평가 영역에 관한 편향적 채점 경향을 확률적으로 추적하여 채점자 훈련과 시험 타당화 연구에 도움을 제공할 수 있다. 이 Rasch 모형에 근거한 FACETS 프로그램은 영어 말하기 평가 영역에서 신뢰성 있는 평가 도구로 인정받고 있다(North & Schneider, 1998;

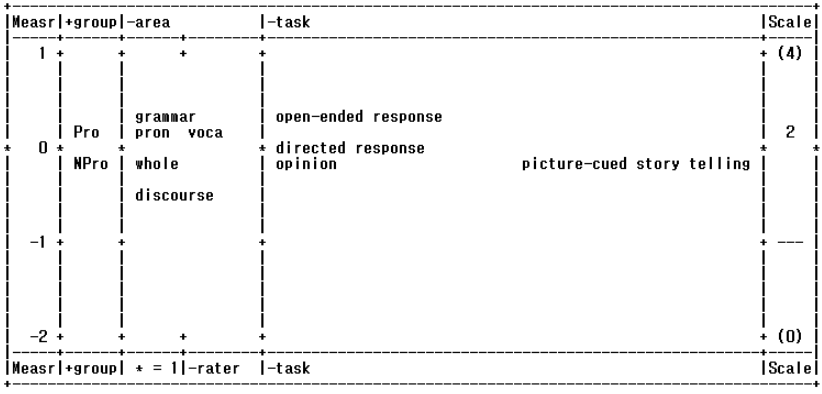
Weigle, 1998; 신동일, 2006; 장소영·신동일, 2009; 이호, 2011).

이 FACETS 프로그램은 다단면 Rasch 모형(multi-facets Rasch model)을 근거로 하는데, 여기서 단면(facet)은 평가 결과에 영향을 미치는 변인을 의미하며, 본 연구에서는 채점자 배경 변인(전문가 여부), 개별적 채점자 변인, 평가 문항 변인, 평가 영역 변인, 수험자 변인 등 다섯 변인이 분석 대상이다. 이 변인이외에도 요인(component)이 분석되었는데, 채점자 요인에는 채점자간 엄격도, 채점자내 일관성, 평가 문항 및 평가영역 타당도등을 분석하였다.

4. 결과 및 토의

4.1. 개별 평가 분석

한국인 영어 말하기 평가에 대한 일반적인 평가 경향을 알아보기 위하여 평가자 분석을 하였다. 영어 평가자의 평가 자료를 FACETS 프로그램으로 분석하여 얻은 logit 값을 도식화하여 제시하면 그림 (1)과 같다.



<그림 2> 집단 * 평가 영역 * 평가 과업

<그림 2>는 이 연구에서 설정한 6국면중 평가집단, 평가 영역, 평가과업을 동일 척도에서 제시한 것이다. 따라서 각 국면의 logit 값을 명시적으로 확인할 수 있다. 둘째 열(+group)은 평가자 집단 분포를 - 전문가 집단과 비전문가 집단 - 나타낸다. 셋째 열(-area)은 평가 영역, 넷째 열(-task)은 평가 과업별 영역별 난도를 나타낸다.

이 연구에 따르면, 한국어 말하기 평가에 참여한 평가자들의 엄격성은 주로 -1 logit과 1 logit 사이에 위치하고 있다. 수험자들의 점수는 0점에서 4점까지 분포되어 있지만, 주로 1.5점에서 2.5점 사이에 중점적으로 집중해 있다는 것을 보여준다. 즉, 대부분의 수험자들은 -1과 1 logit 사이에 해당하는 2점 등급에 위치하는 것으로 나타났다. 전반적으로 하위 등급인 0점 등급과 1점 등급이 거의 없고, 역으로 상위 등급인 3점대와 4점대가 매우 드물다는 것을 보여주고 있다. 세 번째 항목인 평가 영역의 난도 측정에서 문법 부분이 가장 엄격하게 채점되고 있음을 보여주고 있으며(0.33 logit), 반대로 담화 적정성부분에서 가장 관대하게 평가를 하고 있음을 보여주고 있다(-0.36 logit). 네 번째 항목인 평가 과업에 대한 측정에서, 개방형 반응 영역에 대하여 가장 엄격한 평가가 이루어지고 있으며(0.37 logit), 반대로 의견제시 및 그림 묘사나 지시형 설명 문제는 상대적으로 관대한 평가가 주어지고 있다(-0.18 logit).

한국어 말하기 평가에서 비전문가 평가자와 전문가 평가자 집단사이에 평가 형태에 대하여 유사점도 발견되지만, 일부 영역에서 상당히 의미 있는 차이점을 나타내고 있다. 다음 도표의 집단 항목은 전문가 채점자들이 비전문가 채점자들에 비하여 엄격한 채점 경향을 잘 보여주고 있다. 다음 표에서 제시한 것처럼 전문가 집단의 엄격성은 0.21 logit로, 비전문가 집단 채점자 -0.05 logit 보다 상위에 위치하고 있다.

<표 1> 평가 엄격성 수준과 적합도

평가 집단	평가자	로짓 (logit)	표준 오차	내적합 제곱평균	내적합 Z 검정	외적합 제곱평균	외적합 Z 검정
전문가 평가자	P1	0.75	0.10	1.31	1.5	1.13	1.5
	P2	0.50	0.09	0.63	-1.7	0.65	-1.7
	P3	-1.28	0.09	1.63	6.7	1.62	6.6
	P4	0.34	0.09	0.80	-2.5	0.81	-2.4
	P5	0.35	0.09	0.67	-1.7	0.70	-1.8
	P7	0.11	0.09	0.82	1.2	0.72	0.4
	P8	0.32	0.09	0.76	-1.5	0.81	-1.4
	P9	0.33	0.09	0.77	-1.9	0.70	-2.0
	P10	0.12	0.09	0.82	0.8	0.82	0.6
	P11	0.13	0.08	0.81	1.1	0.72	0.5
	P12	0.33	0.09	0.75	-1.4	0.81	-1.5
	평균	0.21	0.09	0.98	0.63	0.98	0.67
비전문가 평가자	N1	0.78	0.10	0.82	-1.8	0.85	-1.8
	N2	0.34	0.10	0.63	-2.9	0.82	-2.3
	N3	0.14	0.10	0.81	-2.0	0.82	-2.2
	N4	0.19	0.09	0.52	0.5	0.92	0.8
	N5	0.16	0.09	0.82	-0.9	0.82	-1.2
	N6	-0.48	0.10	1.55	2.9	0.84	1.1
	N7	-0.56	0.10	1.51	2.3	0.86	2.2
	N8	0.18	0.09	0.72	0.5	0.92	0.8
	N9	0.17	0.09	0.82	-0.8	0.82	-1.1
	N10	-0.32	0.10	0.84	-1.6	0.83	-1.7
	N11	-0.46	0.10	1.41	2.1	0.87	2.1
	N12	0.21	0.09	0.71	0.6	0.91	0.7
	평균	-0.05	0.10	0.97	-0.5	0.98	-0.3

<표 1>은 평가자들의 개별적인 logit 점수 및 표준 오차, 내적합 지수, 외적합 지수를 제시한 것이다. 평가자들의 엄격성은 -1.28logit 부터 0.78 logic까지 포함한다. 이러한 분포에 대하여 Chi-square 값은 523.6 ($P<0.001$), 분리 신뢰도 R=.98로 나타나 통계적으로 평가자들의 평가 엄격성면에서 유의한 편차가 있음을 확인할 수 있다.

전반적으로 전문가 집단이 엄격한 평가를 하는 있으며, 비전문가 평가

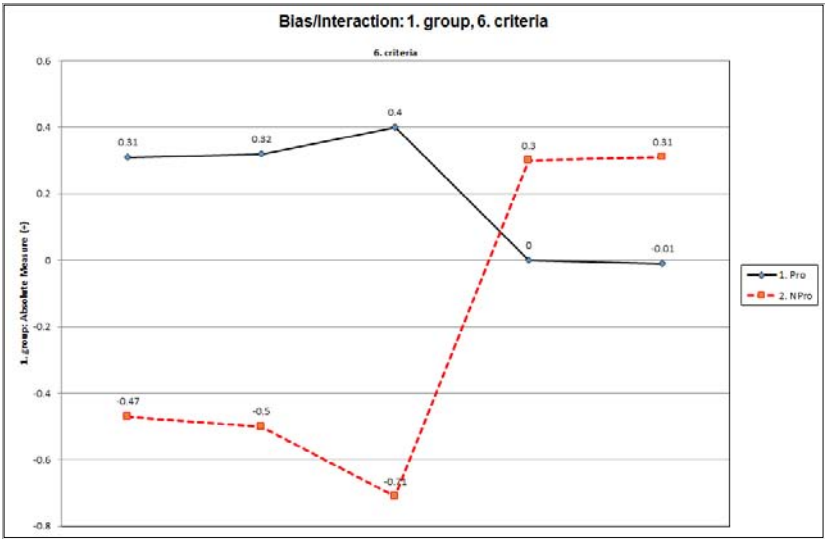
자들은 관대한 평가를 하는 경향을 발견할 수 있다. 그러나 가장 엄격한 평가를 하는 평가자는 비전문가 평가자인 N1이며 (0.78 logit), 전문가 평가자인 P3는 가장 관대한 평가를 하고 있었다(-1.28 logit). 평가자의 내적 일관성을 나타내는 내적합도 분석 결과에 따르면, 전문가 평가자 집단에서도 채점의 일관성에 문제가 있는 채점자가 발견되었다. 즉, 전문가 평가자 P1이 부적합 판정을, 그리고 P2와 P5가 과적합 판정을 받았다. 일반적으로 적합도가 0.75 이하는 과적합(overfit), 1.3 이상은 부적합(misfit)로 산정하였다(McNamara, 1996). 평가자들의 일관성을 판정할 때는 외적합 값보다는 좀 더 안정적인 내적합값에 따라 판정한다(최숙기, 2011). 부적합 판정을 내리는 P1은 우수한 한국어 발화에 대하여 예상치 않게 낮은 점수를 주거나, 혹은 매우 나쁜 한국어 발화에 대하여 기대하지 않는 높은 점수를 주는 경향이 있음을 보여준다. 또한 과적합 판정을 내리는 P2와 P5는 모든 항목에 대하여 일정한 점수를 부여하는 경향이 과도하게 많음을 보여주고 있다. 전문가 집단에서는 3명만이 평가상 문제점을 보이고 있지만 (25%), 비전문가 집단에서는 부적합 판정자 4명 (N6, N4, N8, N12), 과적합 판정자 3명 (N6, N7, N11)으로 평가에서 문제점을 보여준 평가자가 이 집단의 12명중 60%인 7명에 이른다. 즉, 평가자 역량은 평가의 신뢰도와 타당도에 일차적으로 영향을 미친다고 볼 수 있다.

편향도 분석은 두 집단의 일부 평가자들이 편향적인 평가를 수행하고 있음을 보여주고 있다. 편향 분석은 집단과 평가 과업 및 영역간에 잠재적인 상호 작용을 조사하는 것이다(Lincare, 1989; Kim Y-H, 2009). 편향 분석은 Z 검정에 의하여 판별되는데, -2와 2사이에 위치하면 평가자는 편향 없이 평가한다고 볼 수 있다. 만약 －2이하로 내려가면 과적합 평가 판정이 내려지는데, 이는 이 평가자가 집단과 과업 변수에 따라서 관대한 평가를 내린다는 의미이다. 역으로 +2 이상의 부적합 판정이 나오는 경우는, 이 평가자가 다른 과업에 상관없이 엄격한 평가를 내린다는

의미이다. 전문가 집단에서는 2명 (P3가 부적합 판정을 내리고 있으며, P4는 과적합 판정을 내리고 있다)만이 편향도에 문제점을 노출하고 있지만,. 비전문가 평가자에서는 부적합 판정자 2명 (N6, N7), 과적합 판정자 3명(N2, N3, N7)등 모두 5명이 편향적인 평가를 하고 있는 것으로 나타났다.

4.2. 평가 영역의 엄격성 비교

각 집단의 평가영역에 대한 평가 엄격성을 조사하였다.



<그림 3> 전문가 및 비전문가 평가자 집단에 의한 평가 영역 난도

<그림 3>은 4개의 평가 영역에 대한 두 집단간 난도 추정치가 나타나 있다. 한국어 말하기 평가의 준거는 문법, 발음, 어휘, 담화, 전반적 인상

의 5개 영역을 하위 요인으로 세부적으로 구성하였다. 외국인으로서의 외국어 말하기는 구사자가 조직화, 문법, 내용, 화용론, 기능적, 사회언어적 지식을 얼마나 잘 구사하는냐에 달려있기 때문에(Bachman & Palmer, 1996), 이를 바탕으로 5개 영역으로 분류하여 수험자들에게 제시하였다.

이 분석은 두 집단의 평가자들이 각기 다른 평가영역에서 엄격성이 어떻게 나타나는지를 보여준다. 위 표에서 전문가 집단은 -0.01 logit에서 0.40 logit 까지 각 영역별로 비교적 차이가 작은 범위내에서 움직이고 있지만, 반면에 비전문가 평가자 집단은 -0.71 logit에서 0.31 logit 까지 영역별로 폭넓은 범위를 나타내고 있다. 이는 전문가 평가자들이 각 영역별로 비교적 고르게 평가를 시행하고 있는 반면에, 비전문가 집단에서는 영역별로 기준이 애매모호함을 보여주고 있다. 특기할 사항은 비전문가 집단에서는 발음과 문법, 전반적 인상 부분에서 상당히 관대한 평가를 내린다는 점이다.

<표 2>는 평가 영역중 문법, 담화 영역에서 비교적 커다란 평가상의 차이를 보여주고 있다. 비전문가는 문법 영역에서 -0.71 logit으로 매우 후한 평가를 하는 반면에, 전문가는 0.40 logit으로 비교적 엄격한 평가를 하고 있다. 또한 이런 경향은 발음과 전반적 영역에서 모두 나타났다. 그러나 담화와 어휘 부분은 두 집단 모두 비교적 동질적 평가 경향을 보였다. <표 2>는 각 평가 영역의 신뢰도 지수를 보여준다.

<표 2> 평가 영역 엄격성 수준과 적합도

평가 과업	집단	logit	표준 오차	내적합 제곱평균	내적합 Z 검정	외적합 제곱평균	외적합 Z 검정
어휘	전문가	0.31	0.09	1.01	0.00	1.02	0.2
	비전문가	-0.01	0.10	0.75	-3.1	0.76	-3.0
발음	전문가	0.31	0.09	1.21	2.3	1.19	2.2
	비전문가	-0.47	0.10	0.77	-3.0	0.76	-3.0
문법	전문가	0.40	0.09	0.76	3.1	0.75	3.2
	비전문가	-0.71	0.10	1.33	-3.4	1.32	-3.3

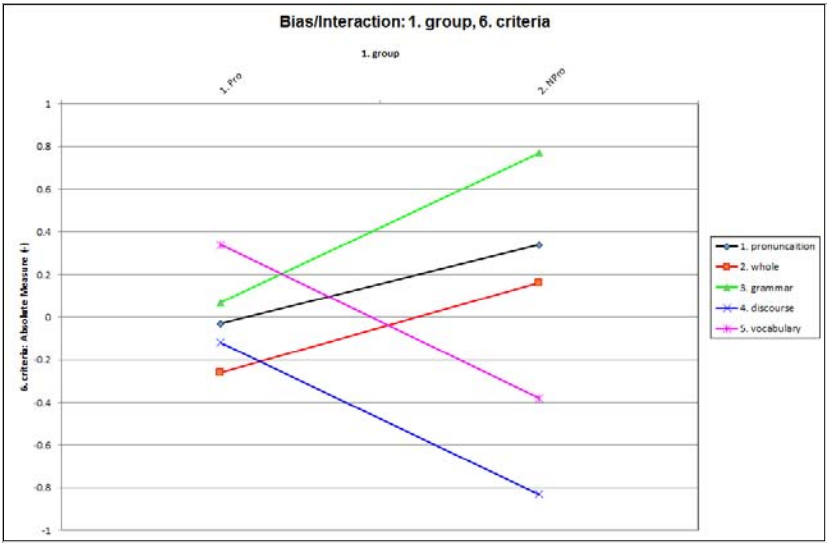
담화	전문가	0.30	0.09	0.98	-0.1	0.98	-0.2
	비전문가	0.00	0.10	1.26	0.7	1.07	0.8
전체	전문가	0.32	0.09	0.99	-0.1	0.98	-0.2
	비전문가	-0.50	0.10	1.05	-0.6	1.04	-0.4

우선 내적합 표준화 점수 Z값을 살펴보면, 담화와 능숙도 영역은 일관적인 평가를 한다고 볼 수 있지만, 전문가 집단은 문법 영역에서, 비전문가 집단은 발음과 문법 영역에서 평가 일관성을 확보하지 못하는 것으로 파악되고 있다. 이 표준화 점수 Z 값은 −2와 2사이에 분포하고 있는 경우 전반적으로 일관성을 유지하고 있다고 보는데, 이는 문제 선정이 매우 적절하게 제시되어 평가가 일관성을 유지하고 있느냐를 나타내는 지표이다(Lincare, 1989; 신동일, 2006).

몇 몇 평가 영역에서 과적합과 부적합이 발생하는 이유는 이 평가 영역 준거에 대하여 두 집단의 평가자들이 해석상의 혼란으로 일관적인 평가를 내리지 못하고 있음을 잘 보여주고 있다. 특히, 발음 영역과 문법 영역은 수험자들의 능력기준을 설정하지 못하고 일관성 없이 점수를 부여하는 것으로 나타났다. 이와는 반대로 담화 영역과 전반적 인상 영역은 평가자들의 평가기준을 확고히 하고 일관적인 평가를 내린다. 어휘 영역에서는 전문가는 일관적인 평가를 내리지만, 비전문가 평가자의 평가는 부적합 판정이 내려지고 있다. 이런 과, 부적합 현상은 각 집단의 평가자들이 한국어 말하기 평가에 평가자 지식 적용에 혼란을 겪는 것을 보여주는 것으로, 전문가 집단에서도 일부 영역에서 문제점을 노출하고 있다.

각 집단의 한국어 말하기 평가의 영역별 평가 준거의 타당함을 살펴보기 위하여 평가집단의 편향 분석을 실시하였다. 편향 분석은 평가의 특정 국면(facet)들 사이에 어떤 상호 작용이 발생하는지를 분석하는 것이다. 이 분석은 서로 다른 logit값의 차이가 Rasch 모델의 기댓값과 상당한 차이가 있을 때, 그 값이 유의할 경우 분석되는 것이다(Lumely and

McNamara, 1995; 장소영·신동일, 2009; 최숙기, 2011).

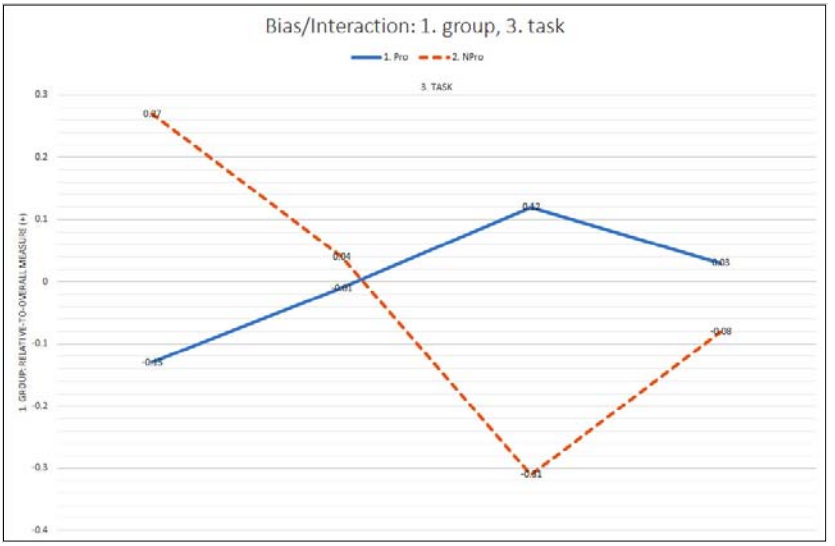


<그림 4> 평가 영역에서 두 집단간 편향도

이 그림은 두 집단의 5개의 평가 영역에 대하여 편향 결과를 그림으로 나타낸 것이다. 두 집단사이에서 어휘 영역과 발음 영역에서 유의할만한 상당한 정도의 편향이 발생하였다. 전문가 집단은 각 영역이 비교적 고른 편향도를 지니고 있지만, 비전문가 집단의 편향도는 상당히 큰 폭을 지니고 있다.

4.3. 평가 과업의 엄격성 비교

각 집단의 평가영역에 대한 평가 엄격성을 조사하였다.



<그림 5> 전문가와 비전문가 평가자 집단에 의한 평가 과업 난도

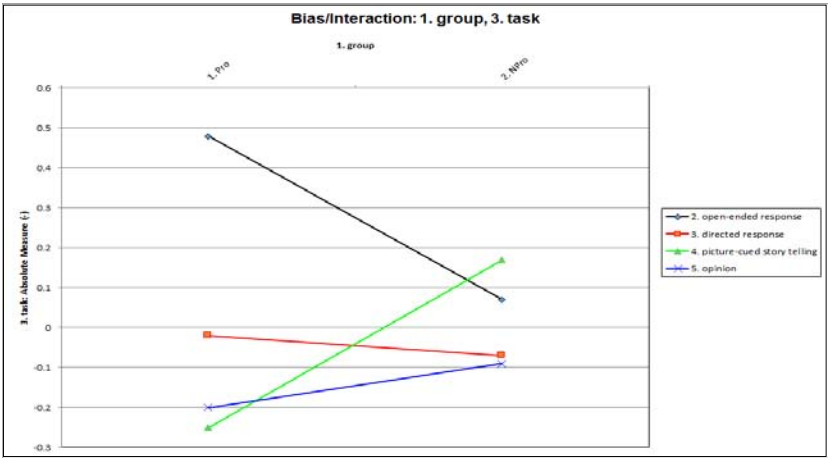
대체적으로 평가 과업별 난도 측정에서 비전문가 평가자와 전문가 평가자는 전혀 다른 평가 형태를 보인다. 위 표에서 전문가집단은 개방형 과업에 엄격한 평가를 하는 반면, 그림 설명 과업은 비교적 후한 점수를 주고 있다. 반면에 비전문가 집단에서는 개방형 과업에 매우 후한 점수를 주고, 그림 설명 과업에서는 매우 엄격한 점수를 주고 있다.

<표 3>은 평가 영역중 의견 제시영역에서 두 집단 간 매우 다른 평가 형태를 보인다. 비전문가는 문법 영역에서 0.10 logit으로 비교적 엄격한 평가를 하는 반면에, 전문가는 -0.22 logit으로 비교적 관대한 평가를 하고 있다. 지시 반응 영역에서도 두 집단 모두 비슷한 평가를 하지만, 비전문가는 -0.25 logit 으로 상대적으로 관대한 점수를 부여하지만, 전문가는 -0.03 logit 으로 비교적 엄격한 평가를 수행하고 있다.

<표 3> 평가 과업별 두 집단간 엄격성 수준과 적합도

평가 과업	집단	logit	표준 오차	내적합도 제곱평균	내적합 Z 검정	외적합 제곱평균	외적합 Z 검정
개방형	전문가	0.37	0.08	0.94	0.7	0.96	0.4
	비전문가	-0.12	0.09	0.32	-0.8	1.17	-0.9
지시 반응	전문가	0.03	0.08	0.94	0.8	0.92	1.0
	비전문가	-0.01	0.09	1.04	-0.5	1.04	-0.5
설명	전문가	0.11	0.08	0.98	1.4	0.89	1.4
	비전문가	-0.31	0.09	0.83	-2.4	0.83	-2.2
의견 제시	전문가	0.09	0.08	1.18	1.3	1.05	1.2
	비전문가	-0.08	0.09	0.93	-0.9	0.92	-1.0

내적합 표준화 점수 Z값을 살펴보면, 거의 대부분 -2와 2사이에 위치하고 있다. 이는 두 집단의 평가자들이 평가 영역에 관한한 전반적으로 일관성을 유지하고 있다고 볼 수 있다. 그러나 비전문가 집단에서는 그림 설명 영역에서 과적합 판정이 나오고 있다. 다음은 이 평가 과업에 대한 타당도를 조사하기 위하여 평가집단에 대한 편향 분석을 실시하였다. 편향 분석은 특정 국면들 사이에 어떤 상호 작용이 있는지를 분석하는 것이다.



<그림 6> 평가 과업에서 두 집단간 편향도

이 그림은 확실히 그림 설명과 개방형 문제 영역 평가에서 두 집단 간 상당한 편향성이 존재함을 잘 보여주고 있다. 전문가 집단에서 개방형 문제의 편향도는 상당히 주목할 만한 수치를 보여주고 있으며, 이는 채점자 훈련에서 개방형 문제 평가에 상당히 관심을 기울어야 함을 의미한다.

4.4. 평가 척도의 집단 간 비교

평가 척도는 평가가 얼마나 적절한지를 나타내는 중요한 기준이다(최숙기, 2011). 척도의 폭이 너무 클 경우, 평가자가 각 척도에 대한 해석에 혼란을 줄 수 있기 때문에 정확한 평가에 어려움을 줄 수 있다. 반면에 척도의 폭이 너무 작을 경우 신뢰도가 부여된 개별 평가 능력을 측정하기가 불가능할 수 있다. 본 평가는 연구용으로 제작되었기 때문에 5 단계 방법을 채용했다.

이 평가 척도가 집단별 말하기 평가에 어떤 특징을 보여주는지를 보여주기 위하여 범주화 통계치와 확률 곡선에 대한 분석을 통하여, 반응 범주별 산출 빈도(%)와 평균 능력 추정치와 경계 모수치에 대한 결과를 제시하였다.

<표 4> 집단별 영어 말하기 척도별 반응빈도 및 Rasch 범주 통계치

척도	집단	범주빈도(%)	외적합지수	평균능력 추정치	경계 모수치
0	전문가	1%	1.9		
	비전문가	1%	1.4		
1	전문가	18%	1.0	-4.02	0.32
	비전문가	30%	1.0	-4.49	0.23
2	전문가	47%	1.0	-0.85	0.08
	비전문가	52%	1.1	-1.11	0.07
3	전문가	27%	0.9	1.55	0.07
	비전문가	16%	0.8	1.74	0.06
4	전문가	5%	0.7	3.52	0.13
	비전문가	1%	1.0	3.85	0.24

척도 범주별 반응 빈도는 두 집단 모두 척도의 중간 등급인 2점 등급에서 가장 높은 것으로 나타났다. 반면에, 0점 등급과 4점 등급은 가장 낮은 것으로 나타났다. 이 검사에서 전문가 집단의 척도 범주가 적절하게 기능하는 것으로 나타났다. 이는 척도 등급이 높을수록 평균 능력 추정치가 증가하고 있음을 확인할 수 있다. 경계 모수치 분석에서도 평균 능력 추정치와 마찬가지로, 각 범주 경계 간 모수치도 척도 점수가 증가할수록 외국인 한국어 말하기 능력 범주도 서열 순으로 모두 감소하는 경향을 보였다. 단, 4점 등급대에서 상승하는 것은 해석상 제한적인 양상임을 보여준다. 적절한 척도 반응은 척도 간 간극이 넓은 소수의 등급 영향으로 보이는데, 이는 전문가 집단에서도 최고 등급인 4점대를 주는데 주저하고 있음을 보여준다. 앞으로 평가자 훈련은 등급 배분에 관심을 두어야 한다는 점을 보여주고 있다.

5. 결론

Rasch 모형에 근거한 FACETS 프로그램을 이용한 외국인 한국어 말하기 평가를 채점자, 수험자, 평가과업, 평가영역 등 여러 요인으로 나누어 살펴보았다. 본 연구 결과는 요약하면 다음과 같다. 분석 결과 전문가 집단은 상대적으로 비전문가 집단에 비하여 전반적으로 내적 일관성을 유지하고 있었고, 엄격성 측면에서도 타당한 정도의 형태를 보이고 있었다. 그럼에도 두 집단 모두 등급 간 척도가 적절함을 유지하고 있었다. 그러나 문법 영역과 발음 영역에서 두 집단 간 차이를 보이고 있었다.

전반적으로 두 집단 모두 평가의 일관적인 형태를 유지하고 있지만, 평가자가 전문가이냐 비전문가이냐에 따라 평가의 준거가 달라지는 경향이 일부 발견되었다. 상당한 수의 비전문가 평가자들은 채점 일관성을 측정하는 내, 외적합 제곱 평균값이 과적합 혹은 부적합 채점 경향을 보이고 있었다. 평가 편향성 측정에서도 비전문가 평가자가 비교적 과적합

반응 비율이 높게 나왔다. 이는 과업/영역과 집단 영역 변수가 심각한 정도의 관계를 가진다는 의미이며, 이들의 평가가 매우 관대하게 이루어진다는 것을 뜻한다. 결론적으로, 평가자 역량은 외국인 한국어 말하기 평가에서 일정 부분 영향을 미친다고 볼 수 있으며, 문법 영역과 담화영역에서 두드러지게 나타난다.

▌말하기 평가 샘플▐

수험자 대표 샘플 <수험자 3, 중국인, 거주 기간 3.3년>

1. 개방형 반응 과업: 한국에서 명절때는 무엇을 합니까?

가족과 모임을 하고, 식사를 함께 합니다. 그리고 텔레비전을 봅니다. 명절때는 특집을 많이 합니다. 추석 특집, ...흠, 명절 특집을 봅니다. 저는 특히 ... 아이돌이 나오는 노래와 춤을 좋아합니다. .. 정말 좋아요.

2. 지시 반응 문항: 당신이 가고 싶은 장소를 설명해 주시오.

서울에 있는 남산 타워에 가고 싶습니다. 여자 친구랑 거기서 사랑의 열쇠를 잠그고 싶습니다. .... 흠, 사랑을 영원히 걸고 싶습니다. 여자 친구가 무척 가고 싶어합니다.

3. 사진 설명 과업: 사진을 설명하세요.

여기는 수영장입니다. 많은 사람들이 수영을 합니다. 흠.. 큰 우산이 보입니다. 나무도 많습니다. 사람들이 그늘에서 쉬고 있습니다.... 날씨가 무척 덥습니다.

4. 의견제시 문항: 누가 가장 똑똑한가?

친형입니다. 친형은 나에 대하여 모든 것을 알고 있습니다. 형제이기 때문에 모든 것을 알고 있습니다. 공부도 학교에서 1등을 합니다. 정말 똑똑합니다. 부럽습니다.

<참고 문헌>

김정숙·원진숙(1993). 한국어 말하기 능력 평가기준 설정을 위한 연구, <이중언어학> 10집, 이중언어학회. 24쪽～33쪽.

김정숙·이동은·이유경·최은지(2007). 한국어 표준 말하기 시험 측정 도구를 위한 기초 연구, <한민족어문학> 51집, 한민족어문학회. 229쪽～258쪽.

노대규(1983). 외국어로서의 한국어 시험과 평가, <이중언어학> 1집, 이중언어학회. 139쪽～170쪽.

신동일(2006).『한국의 영어 평가학』, 서울: 한국문화사.

이석재·박전규(2003). 한국인의 영어 문장 발음에 대한 한국인/원어민/ILT 평가 점수 사이의 상관관계, <대한음성학회 2003 가을 학술대회 발표집>. 83쪽～87쪽.

이영식(2006). 영어 말하기 평가의 개발, <영어교육연구> 11호 2권, 글로벌영어교육학회. 1쪽～18쪽.

이은혜·김성아(2007). 영어 말하기 평가 도구 IELTS와 MATE 비교 분석, <언어연구> 24호 1권, 1-23쪽.

이호(2011). 귀납적 이분법 척도를 활용한 예비 영어교사의 영어 말하기 평가 신뢰도와 효용성에 관한 연구, <한국교원교육연구> 27호 3권, 한국교원교육학회. 215쪽～237쪽.

이호·김혜영·이진화(2011). 영어논술평가에서의 일반 척도, 세부과업별 척도, 이분식 척도의 비교 연구, <영어학> 11호 3권, 한국영어학회. 601쪽～626쪽.

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

전나영·한상미·윤은미·홍윤혜·배문경·정혜진·김수진·박보경·양수향(2007). 한국어 말하기 능력 평가 도구 개발 연구, <외국어로서의 한국어교육> 32호, 연세대학교 언어연구교육원. 259쪽～338쪽.

전은주(1997). 한국어 능력 평가 - 말하기 능력 평가범주 설정을 위하여, <한국

어학> 6호, 한국어학회. 132쪽～172쪽.
정광·고창수·김정숙·원진숙(1994). 한국어 능력 평가 방안 연구, <한국어학> 1호, 한국어학회. 1쪽～22쪽.
최숙기(2011). Rasch 모형을 활용한 요약문 평가 준거 개발및 타당도 분석, 25호, <독서연구>, 한국독서학회. 415쪽～445쪽.
Bachman, L. F., & Palmer, A. F(1996). *Language testing in practice.* Oxford: Oxford University Press.
Chalhoub-Deville, M.(1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing*, 12, 16-33.
Hadden, B. L.(1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning*, 41, 1-24.
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.
Knoch, U.(2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing,* 26(3), 275-304.
Lane, J, L.(2008). The basics of rubrics. htttp://www.schreyerinstitute. psu.edu/ pdf.
Lincare, J. M.(1989). *Many-facet Rasch measurement.* Chicago, IL: MESA Press.
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. F.(1996). *Measuring second language performance*. London: Longman.
North, B., & Schneider, G.(1998). Scaling descriptors for language proficiency scales. *Language Testing*, 15(2), 217-262.
Weigle, S. C.(1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263-287.

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

안현기(Hyunkee Ahn)
서울대학교 영어교육과

151-015 서울시 관악구 관악로 1번지
전화번호: 82-2-880-7673
전자우편: ahnhk@snu.ac.kr

접수일자: 2014년 4월 20일
심사(수정)일자: 2014년 5월 21일
게재확정: 2014년 6월 4일