중학교 영어교사의 말하기평가 채점경향 분석

백현영 · 양병곤
(부산대학교)

Hyun-young Baek and Byung-gon Yang(2011), A Study on Middle School English Teachers' Rating Patterns of Speaking Test. *Journal of Language Sciences.* The purpose of this study is to investigate middle school English teachers' rater reliability in English speaking tests operated at schools. Eight raters participated in this study. Three of them are native speakers of English and have at least one year long educational experience at Korean middle schools, and five raters are middle school Korean English teachers with different educational experiences. The Korean English teachers are divided into two groups of 'high' English speaking competence and 'medium'. All of them rated eighteen test takers' utterances, who are second year at a middle school, and their rating patterns were analyzed with FACETS, a computer software of many-faceted Rasch model. This study also interviewed the raters to make it possible to identify the causes of rater variability and rater errors in the procedures of rating, and the actual conditions of English speaking tests at a middle school. Results showed that English teachers at middle schools still have problems rating students' speaking ability appropriately with high variations in their scores and criteria. Further studies on results after training teachers or comparing native speaker's rating procedure with that of Korean teachers in detail may be desirable to resolve current issues and problems. **(Pusan National University)**

Key Words: rater reliability, speaking tests, rating patterns, rater variability, rater errors, many-faceted Rasch model

1. 서론

의사소통을 중시하는 현대사회에서는 지식으로서 영어뿐만 아니라 실제 생활 속에서 사용 가능한 언어로서 영어 습득이 필요하다. 지구촌의 다양한 언어를 구사하는 사람들의 경험과 지식을 공통의 언어인 영어를 통해 글과 인터넷 등을 통해 접근할 수 있고 또한 직접 만나서 대화를 통해 알아 낼 수도 있다. 일부에서는 보다 유창한 의사소통 능력을 갖출 수 있도록 이중언어

로서 일상생활과 학습활동에 필요한 영어 구사를 가능하게 하는 교육시스템 도입의 필요성(Jong, 2011)을 제기하기도 한다. 이러한 사회의 요구에 부응하기 위해 교육과학기술부는 영어교과에 대한 각종 교육정책을 시행하고 있다. 특히 말하기 능력 신장을 위해 원어민 보조교사 채용, 생활영어 관련 책자와 시험 개발, 영어교사의 해외파견 및 회화 관련연수 등 막대한 교육예산을 투자하고 있다. 또한, 제7차 교육과정은 영어교육의 목표를 의사소통능력을 기르는 것으로 명시하고 있으며, 이에 따라 평가항목에도 영어 말하기 영역을 반드시 포함하게 하고 있다.

의사소통능력 신장이라는 목표 하에 중학교에서는 각 학년마다 최소 1회 이상의 영어 말하기 평가가 실시되고 있다. 평가방법, 평가내용, 전체평가에서 차지하는 비율 등이 학생들의 실제 학습에 영향을 미치는 점을 감안할 때 영어 말하기 평가에서 기본점수를 제외한 실제 반영점수가 전체평가에서 차지하는 비율은 그다지 높지 않다. 또한 학생들의 상황에 맞는 자신의 의사를 표현하는 내용에 대해 평가하기보다는 교사가 제시한 틀에 박힌 대화문에 대한 단순히 암기력 테스트에 머무는 경우가 많아 평가 중요성의 인식에 반해 평가실행은 아직 초보 단계에 머물러 있음을 알 수 있다.

위와 같은 현상의 큰 원인 중 하나는 학교 영어교사에 의해 시행되는 말하기 평가 채점에 대한 학부모와 학생들의 불신이라고 볼 수 있다. 학교 내에서 이루어지는 영어 말하기는 교사의 직접적인 평가에 절대적으로 의존하게 된다(Kim, 1998). 하지만, 인력수급 문제에서 제한을 받을 수밖에 없는 학교 교사들이 공인된 영어 말하기 평가에 비해 평가와 채점에서 전문성이 떨어지므로 학부모와 학생들은 교사의 말하기 평가 채점을 신뢰하지 못하게 되며, 교사 자신도 스스로의 주관적인 판단과 해석에 의해 채점을 하는 것에 큰 부담을 느낀다. 결국, 본래의 영어 말하기 평가 취지와는 어긋나게 채점 신뢰도를 높이기 위해 특정 대화문 암기하기, 영어 팝송 부르기 등의 왜곡된 방법으로 평가를 실시하는 현상이 나타나고 있다.

위와 같은 문제점을 해결하고 영어 말하기평가 본연의 취지를 되살리기 위해서는 현재 학교현장에서 평가를 시행하고 있는 교사들의 말하기 평가 채점 경향이나 이들이 겪고 있는 어려움에 대한 자세한 연구가 필요하다. 구체적으로 본 연구에서는 다음의 연구문제를 탐구한다.

1. 중학교 영어교사의 말하기 평가 채점경향은 어떠한가?
2. 중학교 교사의 영어 말하기능력에 따른 말하기 평가 채점경향은 어떠한가?

3. 중학교 영어교사의 말하기 평가 채점경향에서 나타난 개인별 특성은 어떠한가?

2. 이론적 배경

2.1. 채점자 신뢰도

말하기평가와 같은 언어수행능력을 평가하는 시험은 평가영역과 평가기준에 대한 채점자들의 주관적인 판단에 의존하므로 채점자간 신뢰도 확보의 문제를 안게 되며, 채점자내 신뢰도도 애매한 채점기준, 피로, 수험자에 대한 사전지식 등에 의해 쉽게 영향을 받는다.

언어수행평가의 채점에서 발생하는 차이는 채점자들이 평가하는 언어수행능력에 대한 판단이 서로 다른데서 나타난다(Stansfield & Ross, 1998). McNamara(1996)와 Savignon(1985)은 말하기능력 판단시 채점자들이 평가영역 중 문법의 정확성을 중요한 요소로 여긴 반면, Halleck(1992)은 의사소통전략이 더 크게 작용한다고 하여 연구에 참여한 채점자에 따라 결과에 차이가 있음을 알 수 있다. 또한, 전체적인 채점경향에서 극단적인 점수를 부여하기를 꺼려하는 채점자와 중간 점수를 부여하기를 꺼려하는 채점자가 있어(McNamara & Adams, 1991), 언어수행능력 평가시 평가영역과 채점기준 자체보다도 그것을 해석하는 채점자의 개인적 성향이 더 큰 비중을 차지함을 알 수 있다.

2.2. 채점경향 분석을 위한 다국면 Rasch 모형의 적용

고전검사이론에 의한 문항분석은 주로 개인의 능력에 대한 평가가 수험자 집단의 능력수준에 의존하게 된다. 수험자가 높은 능력집단 내에서 평가를 받았다면 낮은 능력집단에서 평가를 받을 때보다 개인의 능력수준이 낮게 추정될 것이고, 이는 역으로도 마찬가지이다. 즉, 수험자집단의 상대적인 능력에 따라 특정문항에 정답을 한 수험자의 비율인 문항난이도의 값이 달라진다. Linacre(1989)가 개발한 다국면 Rasch 모형(many-faceted Rasch model)은 고전검사이론이 가지고 있는 위와 같은 한계를 해결한다(McNamara, 1996; North, 1993). Rasch 모형에 의한 분석에서는 수험자의 능력과 문항난이도의 관계를 확률함수를 활용하여 일반화하기 때문에 검사에 따라 수험자들의 점수가 달라지거나 수험자집단에 따라 문항난이도가 달라지

지 않는다는 장점을 지닌다.

다국면 Rasch 모형은 채점자료 분석에서 문항난이도 이외에도 측정상황에 따라 수험자의 점수에 영향을 미칠 수 있는 다양한 변인들을 국면(facet)으로 설정하여 모형 안에 추가할 수 있다. 예를 들어, 채점자의 엄격성을 새로운 국면(facet)으로 보고 추가하면 2국면 Rasch 모형이 된다. 다국면 Rasch 모형에서는 분석이 필요한 변인들을 모두 국면으로 설정하여 각각의 국면을 독자적으로 분석하거나 각 국면들의 상호작용결과를 분석하는 것이 가능하다.

Linacre(1989)가 다국면 Rasch 모형을 적용하여 개발한 컴퓨터 프로그램이 FACETS이다. FACETS는 수행평가와 같은 주관식시험에서 각 채점자의 엄격함과 관대함을 근거로 채점의 신뢰도를 측정한다. 신동일과 설현수(2005)에 의하면 FACETS로 채점자료를 분석하여 다음과 같은 정보를 얻을 수 있다. 첫째, 채점자단면 출력정보를 얻을 수 있다. 어떤 채점자가 가장 인색한 점수를 주었고, 어떤 채점자가 가장 관대한 채점경향을 보이는지와 채점자가 채점의 내부 일관성을 보이는지를 알 수 있다. 또한, 각 채점자의 채점에 대한 실제일치도와 모형에 의해서 추정된 기대일치도가 제공되므로 각 채점자의 채점경향이 서로 일치하는지의 정보를 얻을 수 있다. 채점자간의 채점일치도가 Rasch 측정치에 근거한 기대일치도와 같을 경우 일치도는 '0'로짓으로 산출된다. '−'값은 모형에 의해서 예측된 채점자간 일치도에 비해 실제 채점자 사이의 채점 불일치가 더 크다는 것을 의미하며 '+'값은 채점자간 일치도가 기대이상으로 높다는 것을 의미한다.

둘째, 평가과제 및 평가영역단면 출력정보를 얻을 수 있다. 평가과제단면 적합도와 평가영역단면 적합도에서 과적합이나 부적합을 보이는 문제는 채점자의 채점일관성에 문제가 있는 것이므로 평가과제 측면이든, 평가영역 측면이든, 그 이유를 반드시 확인해볼 필요가 있다. 여기서 과적합이라는 것은 채점경향이 너무 한 쪽에 집중되어서 채점편차가 거의 나지 않는 것을 의미하고, 부적합은 일관성이 없고 채점편차가 너무 큰 것을 의미한다.

셋째, 채점자와 평가과제간 상호작용분석 정보를 얻을 수 있다. 특정 채점자와 평가과제간의 상호작용 혹은 편향적 채점경향을 파악할 수 있다. 표준화된 Z값이 +2.0과 -2.0 범위 밖에 위치한 항목은 각각 채점자가 특정과제를 전체 모형에서 이해하는 정도보다 훨씬 관대하게, 또는 훨씬 엄격하게 채점하는 것을 의미한다.

넷째, 수험자, 채점자, 평가문항, 평가영역 간 상호작용분석 정보를 얻을 수 있다. 일반적으로 채점자료가 측정모형에 적합하다면 95%의 자료가 ±2.0 표

준잔차[(관찰값-기댓값)/표준오차)] 범위 내에 위치하고, 99%의 자료가 ±3.0 범위에 위치하게 된다. 기댓값과 관찰값을 비교하여 특정 채점자가 특정 평가과제에서 특정 수험자에게 모형 기댓값보다 높은 점수를 주었는지, 낮은 점수를 주었는지를 판단하는 것이 가능하다.

이처럼 FACETS 프로그램에 의해 채점자의 엄격성, 내부 일관성, 평가영역에 관한 정보, 편향된 채점경향, 각 단면들의 상호작용 등을 분석할 수 있다. 채점자 신뢰도와 채점경향 분석에 관한 국내 연구들 중 Rasch 분석을 말하기 평가에 적용한 사례는 흔하지 않으며, 대부분의 말하기 평가 채점자료 분석은 전문적인 말하기시험을 대상으로 하고 있다.

한문섭과 신동일(2004)은 YBM 시사영어사에서 개발하여 시행하고 있는 SEPT(Speaking English Proficiency Test)의 채점신뢰도와 타당도를 연구하였다. 수험자 47명의 SEPT 시험결과를 5명의 채점자가 독립적으로 채점하였고, 채점자집단은 채점자교육을 이수한 원어민이었다. 채점자료는 SPSS와 FACETS로 분석되어졌으며, 채점자들이 세부 평가영역 등급판정에서 상당한 편차를 보였다. 문법, 어휘, 유창성, 강세의 평가영역 중 강세에서 -8의 과적합 채점경향이 발견되어 SEPT가 전혀 차별적인 정보를 주지 못함을 알 수 있었고, 문법과 유창성의 경우 채점편차가 너무 불규칙적이어서 일관성을 가진 안정적 시험정보가 제공되지 못하는 것으로 나타났다.

장소영(2001)은 숙명여자대학교에서 개발한 영어 말하기 시험인 MATE(Multimedia Assisted Test of English)의 채점 자료를 분석하였다. 이 연구에서는 MATE 원어민과 비원어민 채점자 7명이 참여하였는데, FACETS 프로그램으로 분석한 결과 7명의 채점자들의 엄격성이 서로 다르게 분포하고 있음을 알 수 있었다. 특히, 채점자내 신뢰도에서 3명의 채점자들은 부적합 지수를 나타내었고 1명의 채점자는 과적합 지수를 보임으로써 변별력 없는 채점을 하고 있는 것으로 평가되었다.

지금까지 살펴본 기존 연구의 대부분은 동일 수험자에 대해 최소 두 명 이상의 채점자가 채점을 하여 평균점수를 산출하는 전문 영어 말하기시험에 관한 것으로, 다양한 수준의 영어능력을 가진 중학교현장에서는 이를 적용하기 어렵다. 따라서 중학교 영어교사의 말하기 평가 채점경향을 분석하여 영어 말하기 평가의 활용 및 정착에 도움이 될 기초자료를 도출할 필요가 있다.

3. 연구방법

3.1. 피험자

영어 말하기 평가에 참여한 피험자는 부산광역시 소재 C중학교 2학년 학생 18명으로 충분한 양의 말하기응답을 도출하기 위해 교내에서 1학년 2학기에 치른 기말고사를 기준으로 하여 편성된 상, 중, 하의 수준별 집단 중에서 '상' 수준의 학생들만을 대상으로 하였다.

3.2. 채점자

이 연구에 참여한 채점자는 부산시내 중학교에서 근무 중인 한국인 영어교사 5명과 영어원어민 보조교사 3명이며, 원어민 보조교사는 모두 1년 이상의 중학교 근무경력을 가지고 있다. 한국인 영어교사 5명은 영어 말하기능력 '상' 수준에 해당하는 2명과 '중' 수준에 해당하는 3명으로 구분된다.

한국인 영어교사의 말하기능력을 수준별로 구분하기 위해서 부산광역시교육연수원에서 주관한 '초·중등 영어(전담)교사 특별 직무연수'에서 말하기능력 상위집단에 속한 중학교 교사 2명과 중위집단에 속한 중학교 교사 3명을 선발하였다. 이 연수에 참여한 교사는 모두 연수시작 전에 국가공인 영어회화능력 평가시험인 ESPT(English Speaking Proficiency Test)에 응시하였으며, 이 시험 점수에 따라 중·고등학교 교사를 상, 중, 하의 세 집단으로 편성하여 연수를 진행하였다.

채점자 구분	교사구분	성별	교육경력	영어권국가 거주기간(국가)	말하기 수준
NA	원어민	남성	1년 6개월	~~아일랜드~~	
NB		남성	2년 3개월	~~캐나다~~	
NC		여성	2년 3개월	~~캐나다~~	
KA	한국인	여성	6년 6개월	없음	상
KB		여성	7년 1개월	1년 10개월(미국)	상
KC		남성	8년 6개월	없음	중
KD		여성	6년 11개월	없음	중
KE		여성	4년 5개월	2년 6개월(미국)	중

표 1. 채점자 개인별 정보

채점자들에 대한 정보는 <표 1>과 같으며, 교육경력은 중등학교에서 근무

한 경력만을 나타낸다. 채점자 중 원어민 보조교사 3명의 평균 교육경력은 2년이며, 한국에서 근무하기 전, 자국의 학교에서 근무한 경력은 없다. 한국인 영어교사 5명의 평균 교육경력은 6년이고, 이들 중 말하기능력이 '상' 수준에 해당하는 1명과 '중' 수준에 해당하는 1명은 미국에서 2년 정도 체류한 경험이 있다.

3.3. 말하기 평가도구

연구자는 ESPT 기출문제집을 참고하여 영어 말하기 평가문항을 개발하였다. ESPT는 8개의 과제 유형으로 나누어 평가하지만, 이 연구에서는 평가 및 채점의 용이성을 고려하여 개인정보(personal information), 선택질문(choice question), 사진묘사(picture description), 길안내(giving directions)의 4개 유형으로 나누어 총 8문항을 제작하였다.

평가문제는 수험자들에게 친숙한 것에 한정했으며, 특히 평가의 시작은 최대한 자연스러운 응답 유도를 위해 개인 신상에 관한 문항으로 하였다. 평가문제 1번부터 4번까지는 개인정보를 묻는 유형으로, 가족소개, 여가 및 취미활동, 미래계획, 친구의 생일선물준비에 관해 묻는 질문들로 구성되어 있다. 평가문제 5번은 선택질문으로, 제시된 두 가지 중 선호하는 하나를 고른 후, 그 이유에 대해서 설명하는 형식이다. 평가문제 6번과 7번은 사진을 보고 사진이 나타내는 바에 대해 설명하는 유형으로, 6번은 과거의 경험과 연결하여 응답해야 하며, 7번은 단순히 사진을 묘사하는 문제이다. 평가문제 8번은 약도를 보고 특정위치를 찾아가는 과정을 안내하는 형태이다.

말하기 평가 질문은 수험자들과 같은 학교에서 근무하고 있는 원어민 보조교사가 각각의 평가문항을 읽는 모습을 디지털 카메라로 동영상 촬영을 하였고, 이 동영상 파일을 컴퓨터에 옮겨 인터뷰 질문용 파워포인트로 제작하였다. 수험자들은 평가가 시작되면 컴퓨터 앞에 앉아 파워포인트의 화면을 클릭하여 질문을 듣고 30초 동안 생각할 시간을 가진 후에 대답을 하였으며, 대답이 끝나면 다시 클릭을 하여 페이지를 넘겨 다음 질문을 듣고 대답을 하였다. 질문을 듣고 난 후 응답 전까지 주어지는 시간은 기본적으로 30초이나 시간이 더 필요한 수험자는 최대 1분 30초까지 대답을 준비할 시간을 허용하였다. 수험자 1인당 걸린 전체 평가 응시시간은 평균 15분이었으며, 이들이 8개의 질문에 대답하는 각 장면을 디지털 카메라로 동영상 촬영을 하여 수험자마다 8개의 동영상파일을 제작하였다.

3.4. 채점방식과 기준

연구자는 ESP 평가아카데미사에 의해 개발되어 시행되고 있는 ESPT 공식 인터넷 사이트(http://www.espt.org)를 참고하여 평가영역을 4개로 나누었다. 평가영역은 효과적인 의사전달을 위한 문법의 정확성 및 단어와 어휘의 다양성을 보는 '정확성(accuracy)', 상대방의 말을 이해하는 정도를 판단하는 '이해(comprehension)', 질문에 응답하거나 자신의 의사를 표명할 때 머뭇거림 없이 보여주는 영어구사의 능숙함을 평가하는 '유창성(fluency)', 그리고 상대방이 이해할 수 있는 올바른 억양과 강세리듬의 사용을 측정하는 '발음(pronunciation)'으로 구분된다. ESPT에서는 정확성 30%, 이해 30%, 유창성 25%, 발음 15%의 비율로 총점에 반영하고 있지만, 이 연구에서는 채점자 개인이 각 영역에 부여하는 가중치를 보기 위해 각 영역의 반영 비율을 25%로 동일하게 하였다.

채점자들에게 평가영역과 채점기준표를 제공하였고 평가영역별로 7척도(1-7점)로 채점하게 되어있는 평가지를 전달하였다. 그리고 앞서 촬영된 수험자의 응답영상을 컴퓨터파일로 제공하였다. 동영상 파일명과 채점지의 수험자명은 수험자의 이름을 알 수 없도록 알파벳과 숫자로 제시하였다. 예를 들어, 첫 번째 수험자의 문제 3번에 대한 응답은 A3으로 표기하였으며, 세 번째 수험자의 문제 5번에 대한 응답은 C5로 표기하였다. 채점자들은 수험자별이 아니라 문항별로 채점을 하여 1번 문항을 모두 채점한 후 2번 문항을 채점하였다.

실제 교사들의 말하기 평가 채점과 유사한 환경을 조성하기 위해 채점자 교육은 따로 실시하지 않았는데, 이는 중학교 영어교사 30명을 대상으로 한 설문조사의 결과를 따른 것이다. 본 연구에서는 현재 중학교 말하기 채점환경을 알아보기 위해 30명의 중학교 영어교사를 대상으로 설문조사를 진행하였으며, 그 중 24명의 교사들이 채점자 교육을 받지 않고 임한다고 답하였고 교육을 받았다고 답한 응답자들도 말하기 평가에 대한 구체적이고 전문적인 사전교육은 대부분 받지 않은 것으로 응답하였다.

3.5. 자료수집 및 분석방법

자료수집은 말하기 평가 수험자의 응답을 디지털 카메라로 촬영하는 과정, 채점자들이 채점하는 과정, 채점결과를 분석하는 과정, 채점자를 인터뷰하는

과정으로 나누어진다. 수험자들의 응답은 문항별로 디지털 카메라 촬영을 하여 채점자들에게 컴퓨터 동영상파일(총 144개)로 제공하였다. 8명의 채점자들은 모든 수험자와 모든 과제에 대해서 동일하게 네 개의 평가영역의 분석적 채점방식으로 채점을 하였고, 그 결과는 수험자의 능력, 채점자의 엄격성, 과제(8가지), 평가영역(4가지)의 4국면 Rasch 모형(4 facets Rasch model)에 근거한 FACETS 프로그램을 이용하여 분석하였다. 그리고 채점자들을 인터뷰하여 채점자의 엄격성, 평가영역에 관한 적합도 지수, 모델의 기댓값에서 벗어난 값 등의 측정에서 나타난 개별적인 채점경향특성의 원인을 파악하고자 하였다.

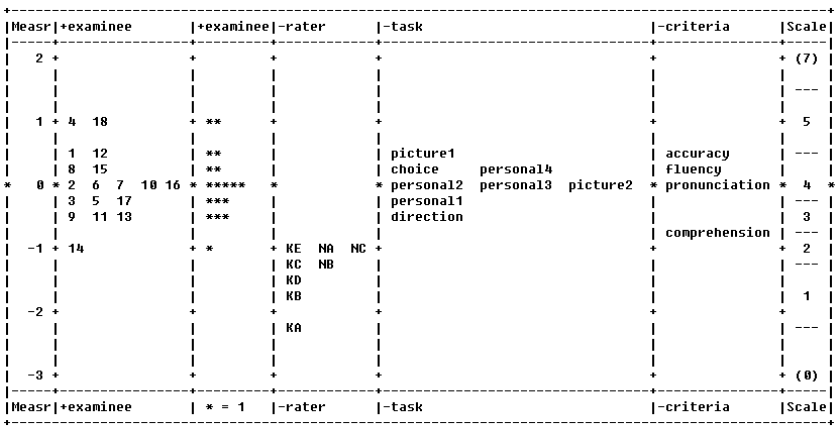
4. 연구결과 및 논의

4.1. 중학교 영어교사의 채점경향 분석결과

4.1.1. 전체평가단면 정보

FACETS의 전체출력정보를 활용하여 평가단면인 수험자의 말하기 능력, 채점자의 엄격성, 평가 과제의 변별도, 평가영역의 변별도를 전체적으로 살펴본다. <표 2>는 각 단면이 로짓값으로 환산되어 하나의 표로 제시되므로 모든 단면(facets)에 대한 총체적 정보를 보여준다.

표 2. 전체평가단면 출력정보



첫 번째 칸(Measr)은 수험자(examinee), 채점자(rater), 평가 과제(task), 평가 영역(criteria), 사용된 점수(scale)의 측정치(measure)를 보여주고 있는데, 로짓값 0을 기준으로 ±3의 범위 내로 제한하여 표시한다.

다음 칸(examinee)은 수험자 번호와 '*'로 표시된 18명의 수험자가 능력수준에 따라 분포되어 있는데, 로짓값 0에 나타나 있는 2번, 6번, 7번, 10번, 16번의 수험자가 중간정도의 능력수준을 가지고 있다고 해석할 수 있다. 그리고 로짓값 0에서 +값으로 척도의 가장 위쪽에 위치해 있는 4번과 18번의 수험자가 가장 상위의 수험자이며, -값으로 가장 아래쪽에 있는 14번의 수험자가 가장 하위의 수험자이다.

네 번째 칸인 채점자(rater) 그룹을 살펴보면 8명의 채점자가 채점 엄격성 정도에 따라 척도에 다르게 위치해 있는데, 로짓값 0 선상에 분포되어 있는 채점자는 엄격하지도 관대하지도 않은 채점자이다. 0을 기준으로 척도의 상위에 위치할수록 채점자가 보다 엄격하고 하위에 위치할수록 보다 관대함을 의미한다. 즉, 채점자 KE, NA, NC는 다른 채점자들에 비해서 엄격한 채점을 했으며, 채점자 KA는 상대적으로 관대한 채점을 하였다. 하지만, 전체적으로는 모든 채점자가 다소 관대하게 채점을 한 것으로 나타났는데, 이는 수험자가 모두 상위그룹의 학생들이기 때문인 것으로 이해된다.

다섯 번째 칸(task)은 8가지 과제가 난이도정도에 따라 분포하고 있다. 로짓값 0을 기준으로 위쪽은 어려운 과제이며, 아래쪽은 쉬운 과제이다. 다시 말해서, 'picture1(그림을 보면서 관련경험과 배경지식 설명)'이 가장 어려운 것으로 나타났으며, 'direction(약도를 보면서 길 안내)'이 가장 쉬운 것으로

나타났다. 로짓값 0선상에 있는 'personal2(여가시간활용)', 'personal3(장래희망)', 'picture2(단순그림묘사)'는 적절한 난이도로 수험자능력을 측정한다는 것을 알 수 있다.

여섯 번째 칸인 평가영역(criteria)에서는 발음이 0로짓에 위치함으로써 가장 적절한 난이도를 가지고 있는 것으로 판명되었다. 또한, 채점자들이 정확성을 가장 엄격하게 채점하고 이해는 가장 관대하게 채점함으로써 수험자의 능력수준을 측정하는데 있어 4가지의 평가영역이 다른 난이도로 작용하고 있다는 것을 알 수 있다.

마지막에 위치한 'Scale' 칸은 사용된 채점점수간의 간격을 나타낸다. 최고점수는 7점이고 최하점수는 0점으로, 2점에서 4점 사이는 간격이 다소 좁고 나머지는 간격이 다소 넓은데, 이는 각 점수들 간의 차이가 등간격(equal interval)이 되지 못한 것으로 해석할 수 있다. 또한 수험자 분포와 비교해 봤을 때, 중간점수인 4점을 받은 수험자가 가장 많은 것을 알 수 있다.

4.1.2. 채점자간 신뢰도

채점자단면 출력정보는 분리신뢰도 지수와 카이스퀘어 검증을 이용하여 채점자의 엄격성과 일관성 정보를 제공한다. 우선 분리도(separation)는 채점의 정확성(엄격성의 수준)에 상대적인 평가치의 분포를 나타내는 것으로 수치가 높을수록 채점자간의 엄격성 일치도가 폭 넓게 분포되었음을 보여주며, 신뢰도(reliability)는 채점자의 엄격함의 차이를 나타내는 지수로서 값이 낮을수록 채점자들이 보다 비슷한 엄격성을 가지고 있음을 의미한다(신동일, 2001). <표 3>의 아래에 나타나있는 엄격성 신뢰 지수를 살펴보면, 분리도 값이 10.15이고 신뢰도 지수가 .99로 이 연구에 참여한 채점자들의 엄격성편차가 대단히 심하며 서로 다른 기준을 적용하여 채점했음을 알 수 있다.

다음으로 카이스퀘어 검증으로 채점자들의 엄격성에 관한 신뢰도를 확인할 수 있다. 여기서는 자유도(d.f.) 값이 7이고 카이스퀘어(chi-square) 값이 646.5에서 p=.00으로 유의미하기 때문에, 모든 채점자들의 엄격성이 같다는 영가설을 기각하고 채점자 집단의 엄격성 차이가 크다는 것을 알 수 있다.

표 3. 채점자단면 출력정보

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	N	rater
3486	576	6.1	6.16	-2.17	.05	.71	-5.3	.77	-3.6	1.20	.53	.49	34.1	26.9	4	KA
3249	576	5.6	5.73	-1.66	.04	1.21	3.2	1.35	5.2	.66	.36	.55	27.6	28.9	5	KB
3191	576	5.5	5.63	-1.55	.04	1.24	3.7	1.20	3.1	.91	.64	.56	36.6	29.0	7	KD
3045	576	5.3	5.36	-1.29	.04	.74	-4.8	.83	-3.0	1.15	.42	.57	29.9	28.6	6	KC
2962	576	5.1	5.21	-1.15	.04	.94	-1.0	.95	-.7	1.13	.66	.58	37.8	28.0	2	NB
2863	576	5.0	5.03	-.99	.04	1.11	1.7	1.13	2.0	.94	.68	.59	34.7	27.0	3	NC
2845	576	4.9	5.00	-.96	.04	1.10	1.6	1.17	2.7	.82	.41	.59	25.3	26.8	1	NA
2839	576	4.9	4.99	-.95	.04	.90	-1.6	.93	-1.2	1.18	.77	.59	31.8	26.7	8	KE
3060.0	576.0	5.3	5.39	-1.34	.04	.99	-.3	1.04	.6		.56					Mean (Count: 8)
217.4	.0	.4	.39	.41	.00	.19	3.3	.19	3.0		.14					S.D. (Population)
232.4	.0	.4	.42	.43	.00	.20	3.5	.20	3.2		.15					S.D. (Sample)

Model, Populn: RMSE .04 Adj (True) S.D. .40 Separation 9.49 Strata 12.99 Reliability (not inter-rater) .99
Model, Sample: RMSE .04 Adj (True) S.D. .43 Separation 10.15 Strata 13.87 Reliability (not inter-rater) .99
Model, Fixed (all same) chi-square: 646.5 d.f.: 7 significance (probability): .00
Model, Random (normal) chi-square: 6.9 d.f.: 6 significance (probability): .33
Inter-Rater agreement opportunities: 16128 Exact agreements: 5199 = 32.2% Expected: 4470.8 = 27.7%

4.1.3. 채점자내 신뢰도

채점자내 신뢰도는 평균제곱값(MnSq)과 표준화값(ZStd)을 이용하여 알아볼 수 있다. 평균제곱값의 평균이 1.0에 가깝기 때문에 1.0을 기준으로 과적합(overfit)과 부적합(misfit)을 판단하게 되는데, .75이하는 과적합으로 1.3이상은 부적합으로 여겨지므로 .75에서 1.3사이의 값을 적절한 적합도 지수로 본다(McNamara, 1996).

위의 내용을 근거로 하여 채점자 일관성을 분석하면, <표 3>에서 채점자 KA와 채점자 KC의 내적합 평균제곱값이 각각 .71과 .74이고 내적합 표준화값이 각각 -5.3과 -4.8로 과적합경향을 보임으로써 채점경향이 너무 한쪽으로 집중되는 변별력 없는 채점을 하고 있다고 할 수 있다. 반면에 채점자 KB와 채점자 KD는 내적합 표준화값이 각각 3.2와 3.7로 부적합 경향을 보임으로써 일관성이 다소 부족한 채점편차가 큰 채점을 했다고 해석할 수 있다.

4.1.4. 평가영역에 관한 적합도 지수

평가영역에 대한 적합도 지수는 평가영역단면 출력정보의 내적합 표준화값으로 분석할 수 있다. 평가영역 중 발음, 유창성, 정확성의 내적합 표준화값(<표 4>에서 Infit ZStd 칸)은 각각 -7.0, -8.3, -4.4로 과적합경향을 보인다. 이는 수험자집단이 모두 영어성적이 상위 그룹인 학생들로 이루어졌으므로 수험자간 수준이 비슷했기 때문으로 추정된다.

표 4. 평가영역단면 출력정보

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N criteria
6963	1152	6.0	6.14	-.81	.03	2.12	9.0	1.93	9.0	.23	.40	.48	2 comprehension
6036	1152	5.2	5.30	.11	.03	.73	-7.0	.72	-7.5	1.25	.62	.56	4 pronunciation
5776	1152	5.0	5.07	.32	.03	.68	-8.3	.69	-8.2	1.31	.66	.57	3 fluency
5705	1152	5.0	5.01	.38	.03	.82	-4.4	.83	-4.3	1.14	.54	.57	1 accuracy
6120.0	1152.0	5.3	5.38	.00	.03	1.09	-2.7	1.04	-2.8		.56		Mean (Count: 4)
502.1	.0	.4	.45	.48	.00	.60	6.9	.51	6.9		.10		S.D. (Population)
579.7	.0	.5	.52	.55	.00	.69	8.0	.59	8.0		.12		S.D. (Sample)

Model, Populn: RMSE .03 Adj (True) S.D. .48 Separation 15.79 Strata 21.39 Reliability 1.00
Model, Sample: RMSE .03 Adj (True) S.D. .55 Separation 18.24 Strata 24.66 Reliability 1.00
Model, Fixed (all same) chi-square: 846.5 d.f.: 3 significance (probability): .00
Model, Random (normal) chi-square: 3.0 d.f.: 2 significance (probability): .22

반면에 수험자의 질문이해정도를 평가하는 '이해'는 내적합 표준화값이 9.0으로 부적합 채점경향이 발견되었는데 이는 다른 평가영역에 비해 상당히 불안정한 채점이 이루어졌음을 나타낸다. 이는 평가 영역에 대한 정보가 부적절하고 구체적이지 못했기 때문으로 채점자들의 이 평가영역에 대한 이해가 부족하다고 해석할 수 있다.

4.2. 영어 말하기능력에 따른 중학교 교사의 채점경향 비교결과

4.2.1. 채점자 신뢰도

채점자를 영어 말하기능력이 '상' 수준인 한국인 교사, '중' 수준인 한국인 교사, 원어민 보조교사의 세 집단으로 나누어 문항반응이론의 적합도 지수와 표준화값을 이용하여 분석한 결과는 <표 5>의 채점자별 엄격성과 적합도 지수 요약과 같다.

표 5. 채점자별 엄격성과 적합도 지수

채점자	교사구분	교육 경력	수준	엄격성	적합도	Z점수
NA	원어민	1년 6개월	·	-.96	1.10	1.6
NB	원어민	2년 3개월	·	-1.15	.94	-1.0
NC	원어민	2년 3개월	·	-.99	1.11	1.7
KA	한국인	6년 6개월	상	-2.17	.71	-5.3**
KB	한국인	7년 1개월	상	-1.66	1.21	3.2*
KC	한국인	8년 6개월	중	-1.29	.74	-4.0**
KD	한국인	6년 11개월	중	-1.55	1.24	3.7*
KE	한국인	4년 5개월	중	-.95	.90	-1.6

*: 부적합, **: 과적합

먼저 한국인 채점자의 데이터 분석 결과, 영어 말하기능력이 '상' 수준에 속하는 한국인 채점자들은 -2.17에서 -1.66의 엄격성의 범위를 나타냈고 '중' 수준인 한국인 채점자들은 -1.55에서 -.95의 엄격성의 범위를 나타냈다. 두 수준 모두 과적합 판정을 받은 채점자와 부적합 판정을 받은 채점자를 1명씩 포함하고 있다. '상' 수준인 채점자 KA와 '중' 수준인 채점자 KC는 엄격성에서는 다소 차이가 있었으나 적합도 지수와 Z점수는 서로 비슷한 값을 나타내면서 과적합경향을 보였고, '상' 수준인 채점자 KB와 '중' 수준인 채점자 KD는 엄격성 수준, 적합도 지수, Z점수 모두 서로 비슷한 정도를 보이면서 부적합경향을 나타냈다.

원어민 보조교사집단은 -1.15에서 -.96의 엄격성 범위로 한국인 교사들보다 채점자간 일치도가 높은 것으로 나타났고 적합도 지수와 Z점수도 적절한 수준으로 판단되는 범위 내에 위치하고 있어 채점자내 일관성도 높은 것으로 해석된다.

4.2.2. 모델의 기댓값에서 벗어난 채점경향

문항반응이론에 의해 기대되는 값에서 벗어난 채점경향을 분석하면 채점자가 어느 수험자에게 어느 영역에서 어긋난 점수를 부여하였는지를 구체적으로 알 수 있다. 채점자를 기준으로 모델의 기댓값에서 벗어난 채점 요약(<표 6> 참조)에 의하면, NA부터 KE까지의 모든 채점자가 수험자가 질문을 이해한 정도를 판단하는 영역에서 모델의 기댓값에서 벗어나는 채점경향을 보여 이해정도를 판단하는 기준에서 채점자간 편차가 큰 것으로 드러났다. 특히 채점자 KB와 KD는 이해영역에서 기댓값과 어긋나는 채점을 한 횟수가 각각 13회와 9회로 이해정도를 정확하게 측정하지 못하고 있음을 알 수 있다. 이

는 채점자 적합도 지수에서 채점자 KB와 KD가 부적합 판정을 받은 원인이기도 하다. 이에 비해 채점자 KA, KC, KE는 평가영역을 가장 정확하게 이해하고 있는 것으로 보이나 사실 채점자 KE를 제외한 KA와 KC는 과적합 채점경향을 나타내는 것으로 밝혀졌으므로 적절하게 평가를 했다기 보다는 중간 점수대를 많이 사용함으로써 잔차값이 커지는 정도가 적었다고 추정된다.

표 6. 채점자를 기준으로 모델의 기댓값에서 벗어난 채점 요약

채점자	구분	수준	평가 영역	벗어난 횟수	합계
NA	원어민	·	이해	4	5
			발음	1	
NB	원어민	·	이해	6	6
NC	원어민	·	이해	5	5
KA	한국인	상	이해	1	2
			정확성	1	
KB	한국인	상	이해	13	14
			정확성	1	
KC	한국인	중	이해	4	4
KD	한국인	중	이해	9	12
			유창성	2	
			정확성	1	
KE	한국인	중	이해	2	2

4.3. 채점자 편향분석과 인터뷰 결과

4.3.1. 편향분석

FACETS 프로그램의 편향분석(bias analysis)은 특정단면들 사이에 체계적인 패턴을 가지고 서로 상호작용이 있는지를 알아보는 분석으로서 로짓값으로 나타나 있는 편향의 크기(bias size)와 t값을 가지고 해석을 하는데 t값의 유의도에 따라 편향의 크기가 유의한지 판단한다(장소영과 신동일, 2009). 이 연구에서는 모든 단면간의 상호작용은 보지 않았고 채점자와 평가영역간의 편향성만을 분석하였다.

채점자와 평가영역간의 편향분석결과를 편향의 크기와 t값으로 정리한 <표 7>의 데이터 결과를 분석하면 모든 채점자가 역시 이해영역에서 유의한 수준의 편향성을 보이는 것을 알 수 있다. 이해영역 이외에 유의한 수준의 편향성이 있는 평가영역을 살펴보면 우선 원어민 보조교사인 채점자 NA, NB, NC는 각각 정확성, 유창성, 발음의 각기 다른 영역에서 상호작용이 있다. 다음으로 한국인 교사집단의 채점자 KB는 유창성과 발음, 채점자 KC는 발음, 채점자 KD는 정확성, 채점자 KE는 정확성과 발음에서 다소 상호작용이 있는 것으로 확인되었다.

표 7. 채점자와 평가영역간의 편향분석결과 요약

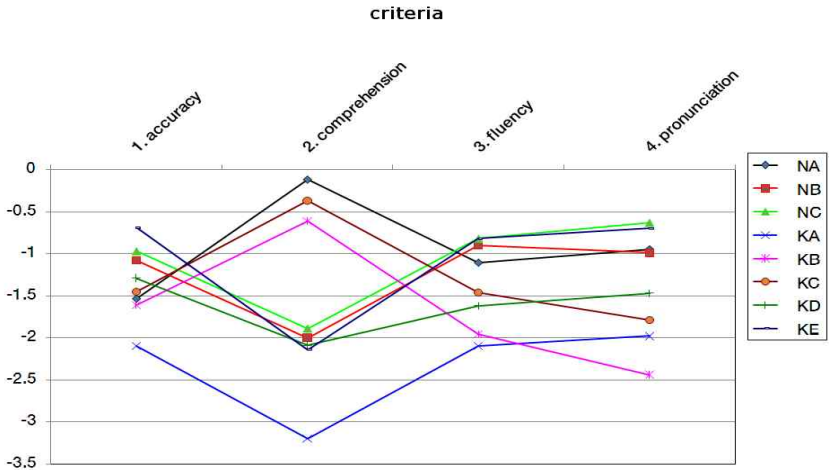
채점자	정확성		이해		유창성		발음	
	Bias Size	t	Bias Size	t	Bias Size	t	Bias Size	t
NA	0.58	7.14*	-0.84	-10.56*	0.16	2.01	0.00	-0.04
NB	-0.07	-0.87	0.85	7.06*	-0.25	-3.32*	-0.16	-1.98
NC	-0.01	-0.14	0.90	7.79*	-0.16	-2.19	-0.35	-4.66*
KA	-0.08	-0.87	1.02	5.00*	-0.07	-0.78	-0.19	-2.13
KB	-0.05	-0.62	-1.06	-12.54*	0.29	3.35*	0.77	7.53*
KC	0.16	1.96	-0.92	-11.30*	0.17	2.06	0.50	5.67*
KD	-0.26	-3.35*	0.54	4.30*	0.07	0.86	-0.08	-0.96
KE	-0.26	-3.58*	1.19	9.35*	-0.13	-1.76	-0.24	-3.21*

*p<.01

채점자와 평가영역간의 편향분석결과를 그래프로 표기한 <그림 1>은 채점자들이 평가영역에서 어느 정도의 편차를 나타내는지 보여준다. 채점자 NA, KB, KC는 전체적으로 유사한 채점경향을 보이고 있으며 정확성, 유창성, 발음영역에서 다른 채점자들에 비해 다소 엄격하게 채점을 하였으나 이해영역에서는 상대적으로 두드러지게 관대한 채점을 한 것이 확인된다. 반면에 서로 비슷한 채점기준을 적용하고 있는 것으로 보이는 채점자 NB, NC, KA, KD, KE는 앞의 채점자들과는 달리 정확성, 유창성, 발음영역에서는 다소 관대하게 채점을 하나 이해영역에서는 상당히 엄격하게 채점을 하고 있는 것으

로 해석된다.

그림 1. 채점자와 평가영역간의 편향분석결과



채점자와 평가영역과의 편향분석 결과, 채점자간 일치도가 높은 것으로 나타났던 원어민 보조교사집단이 세부 평가영역에서는 서로 일치하지 않음을 알 수 있다.

4.3.2. 채점자 인터뷰

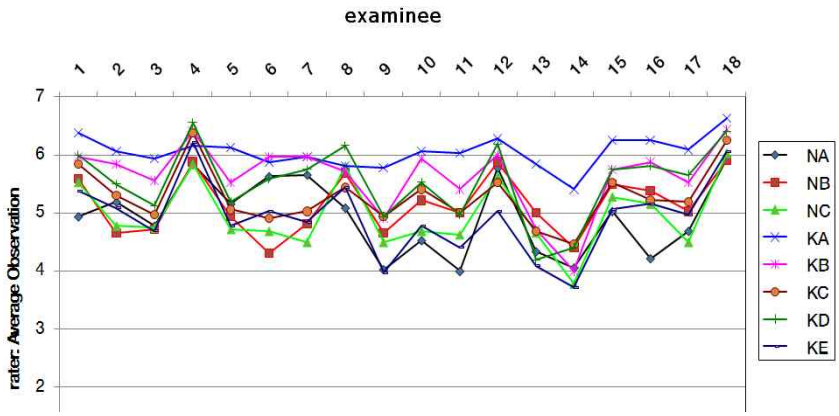
앞에서 엄격성과 적합도 지수에서 과적합과 부적합 경향을 보인 채점자 KA, KB, KC, KD는 채점과정에서 채점오류가 작용하고 있거나 채점자내 일관성이 결여된 것으로 판단되었고 채점자 NA, NB, NC, KE는 서로 유사한 채점경향을 보이는 것으로 해석되었다. 하지만, Rasch 분석은 전반적인 채점성향에서 벗어난 점수를 부적합으로 나타내는 경향이 있으므로 이들의 채점특성을 보다 구체적으로 파악하기 위해 인터뷰 방법을 이용하였다.

인터뷰는 채점자 개별면담으로 진행되었으며, 스스로의 채점에 대한 자신감, 채점시 문제점, 각각의 평가영역에 대한 채점기준, 학교에서의 말하기 평가방법과 평가시 문제점에 대한 질문을 하였다.

1) 채점자 KA

채점자 KA는 자신의 채점일관성에 의문을 가지고 있었다. 채점 중인 수험자가 앞 수험자보다 낮은 말하기능력 수준을 가지고 있으나 앞 수험자의 점수가 정확하게 기억나지 않으므로 부여점수를 결정하기가 어려웠다고 하였다. 또한 채점 시 문법적 오류와 같은 부분에는 크게 신경을 쓰지 않았으며 전체적으로 너그럽게 채점한 것 같다고 하였다. 채점자 KA는 자신의 채점에 대한 불확신으로 인해 점수의 범위를 제한하여 사용함으로써 결과적으로 가장 너그러운 채점자이면서 과적합 경향을 보이는 채점자로 판명되었다. 이러한 사실은 채점자와 수험자간의 편향분석을 그래프로 나타낸 <그림 2>를 보면 더욱 쉽게 확인할 수 있다. 채점자 KA는 대부분의 수험자에게 거의 6점 선상의 점수를 줌으로써 수험자의 능력을 차별적으로 판별하지 못하고 있었다.

그림 2. 채점자와 수험자간의 편향분석결과



채점자 KA는 학생들의 말하기능력 평가의 필요성에는 공감했지만 채점의 신뢰도에는 크게 우려를 표했다. 말하기 수행평가 채점에 학생의 말하기능력 자체보다 평소의 영어성적, 수업태도, 생활태도, 인성과 같은 주관적인 요소가 개입하는 정도가 심하다고 답하여 채점 신뢰도에 있어서 심각한 문제를 초래할 가능성이 있음을 시사하였다.

2) 채점자 KB

채점자 KB는 정확성과 일관성 있는 채점을 위해 굉장히 노력하였다. 공정한 채점을 위해 두 번, 세 번 다시 들었고, 채점의 일관성을 확보하고자 앞에 채점했던 수험자의 응답도 다시 들어보고 점수를 수정하곤 했다고 대답했다. 그럼에도 불구하고 데이터 분석결과 부적합양상을 보였는데, 그 원인은 평가영역 재해석에 있었다. 채점과정에서 평가영역에 문제가 있다고 생각하고 평가영역을 재해석하여 채점한 것으로 밝혀졌다. 수험자의 답변이 길어질수록 문법적 오류가 많이 발생하기 때문에 말하기능력 수준이 높은데도 정확성영역의 점수를 내릴 수밖에 없는 현상이 발생했기 때문이다. 따라서 이러한 불공평한 부분을 보완하기 위해서 이해영역을 글 구성능력으로 보고 채점하였다고 했다. 즉, 이해영역을 수험자가 질문의 의도를 얼마나 잘 이해하고 짜임새 있게 구성하여 대답하였는가로 재해석한 것이다. 또한, 높은 수준의 발화내용이었는데도 불구하고 정확성영역의 점수를 낮게 줘야할 때는 대신 유창성에서 좀 더 높은 점수를 주었다고 했다. 그리고 발음은 수업시간에 가르치지 않으므로 채점 시 고려하는 것이 부적절하다고 생각하여 후한 점수를 주었다고 했다. 이 때문에 유창성과 발음은 상대적으로 관대하게 채점한 것으로, 이해영역에서는 1.06로짓만큼 엄격하게 채점한 것으로 편향분석 결과가 나왔다고 생각된다.

채점자 KB는 학교에서의 말하기 수행평가방법으로 두 사람이 짝지어서 대화문 암기하기를 사용한다고 하였다. 학생들이 대화문을 암기하면서 새로운 표현을 익힐 수 있고 한 학생이 잘 못하더라도 다른 학생이 가르쳐줄 수 있는 장점이 있기 때문이다. 그리고 말하기 수행평가시 가장 어려운 점으로 채점자간 신뢰도 문제를 꼽았다. 여러 명이 공동채점을 한 적이 있었는데 채점기준에 대해서 세부사항까지 미리 합의를 하지 않았더니 각 채점자간 점수편차가 너무 심해서 채점 후에 다시 점수를 조정하기가 쉽지 않았던 경험을 털어놨다. 학교에서는 인력 및 시간부족으로 인해 교사 한 명이 채점을 하는 경우가 대부분인데 이러한 단일 채점자로 인해 발생하는 신뢰도 문제도 크지만, 공동 채점을 하더라도 채점자교육 없이 채점을 할 시 상당한 어려움이 발생할 수 있음을 알 수 있다.

3) 채점자 KC

채점자 KC는 자신의 채점에 대한 강한 자신감을 보였다. 그러나 적합도 지수에서 과적합 판정을 받고 편향분석의 발음영역에서 0.50로짓만큼 관대하

게 채점한 것으로 결과가 나왔던 원인을 인터뷰를 통해서 파악할 수 있었다. 채점자 KC는 기준 점수는 5점이지만 채점을 할 때 점수 차이를 별로 두지 않는 것 같다고 대답했다. 특별히 뛰어난 능력수준을 보이는 수험자를 제외하고는 대부분 평이한 점수를 주었으며 극단적인 점수대는 잘 사용하지 않았다고 하였다. 또한 특정과제에 대한 대답을 잘 못하더라도 다른 과제에서는 대답을 잘 했던 수험자는 평소 말하기 능력수준이 높다고 생각되어 낮은 점수를 주지 않았다고 하였다. 그리고 평가영역 중 발음을 가장 관대하게 채점한 것 같다고 대답하여 채점경향 분석결과와 일치하였다.

채점자 KC가 근무하는 학교에서는 1년에 1회 말하기 수행평가를 실시하고 있으며 대화문 외우기의 방법을 사용한다고 하였다. 그 이유로 타당한 말하기 평가방법은 시간이나 어학실 등이 뒷받침되지 않는데다가 현재 근무 중인 학교 학생들 대부분의 말하기 수준이 낮기 때문에 말하기 평가 자체가 그다지 의미가 없는 것 같다는 점을 들었다. 타당하고 신뢰로운 말하기 평가를 위해서는 평가를 실행하기 위한 환경이 먼저 조성되어야 하며, 학생들의 수준을 충분히 고려한 말하기 평가를 개발할 필요가 있음을 재확인하였다.

4) 채점자 KD

채점의 적합도 지수에서 부적합 판정을 받았던 채점자 KD는 자신의 채점에 확신을 가지지 못하고 있었다. 특히 정확성, 이해, 발음영역에서 자신감이 없다고 하였다. 정확성영역은 수험자의 발화내용이 정확하다고 하더라도 어휘의 사용범위가 좁고 풍부하지 못하면 낮은 점수를 주었다고 하였다. 또한 이해영역에서는 질문에 맞게 대답을 한 수험자의 경우는 상관없지만, 질문과 어긋난 대답을 한 경우 어느 정도의 점수를 주어야 하는지 혼란스러워 했다. 채점자 KD는 기준 점수는 4점으로 설정했지만, 본인의 영어 말하기능력에 대해 자신감이 없기 때문에 관대하게 채점을 했으며 채점과정에서 일관성이 유지되지 않는 것 같아 앞에 채점했던 동영상파일을 다시 열어보곤 했다고 답해, 엄격성, 일관성, 편향분석에서 나타난 결과의 원인을 이해할 수 있었다.

채점자 KD가 근무하는 학교에서는 매 학기 말하기 수행평가를 실시하고 있으며 전체 평가 중 10%를 반영하였다. 하지만 점수의 절반 정도를 기본점수로 부여하기 때문에 실제 반영비율은 미미하다고 하였다. 말하기 평가 반영비율이 낮은 이유로 인터뷰로 평가하게 되면 한 학생씩 평가를 하므로 시

간이 많이 소요되고 채점의 신뢰성에도 문제가 있기 때문에 주로 대화문 암기하기와 같은 타당성이 결여된 말하기 평가를 실시한다는 점과 말하기 평가 이외에도 평가해야 할 수행평가 항목이 많다는 점을 들었다. 각급 학교에서는 실제 지식을 사용하는 능력을 판단하고자 수행평가를 점점 더 강조하고 있지만 본래 수행평가가 의도하는 목적을 실제로 달성하기 위해서는 실효성 있는 수행평가연구가 반드시 이루어져야 함을 알 수 있었다.

5) 채점자 NA, NB, NC, KE

채점자 NA, NB, NC, KE의 인터뷰 결과는 이들의 서로 비슷한 채점경향에 대한 개연성 있는 설명을 제시해 주었다. 평가영역을 독립적으로 채점했다고 답했던 다른 채점자들과는 달리 이들은 수험자 발화의 전체적인 이해가능도에 중점을 두고 채점하였다. 평가영역 중 이해와 유창성에 가중치를 부여하여 이해와 유창성 점수가 높으면 정확성과 발음에서도 호의적인 점수를 주었다. 예를 들어, 발화양이 많은 수험자는 발화양이 적은 수험자에 비해 문법적 오류 빈도가 늘어날 가능성이 높고, 발화양이 적은 수험자가 모든 문장을 정확하게 말했다고 하여 발화양이 많은 수험자보다 정확성이 더 높다고 보기는 어렵다는 것이다. 즉, 다른 채점자들은 채점과정에서 각각의 평가영역을 완전히 분리하여 채점했지만 이들은 평가영역을 수험자 발화의 이해가능도를 중심으로 유기적으로 연결하여 채점했기 때문에 서로 유사한 채점기준을 적용한 것으로 생각된다. 특히, 채점자 KE는 미국에서 거주하면서 상대적으로 원어민들의 문화와 언어에 더 많이 노출됨으로써 말하기능력 채점에 있어서 원어민 보조교사들과 유사한 경향을 나타낸 것으로 보인다. 채점자 KE의 인터뷰 결과는 비원어민 채점자의 경우에는 영어에 대한 자신감이나 목표언어문화권에 대한 친숙함의 정도가 채점과정과 결과에 영향을 미친다는 선행연구(신동일, 2001)와도 일치하였다.

위의 채점자 인터뷰를 통해 채점자 KA의 낮은 채점자신감, 채점자 KC의 점수의 중앙집중경향이 과적합 채점경향의 원인이며, 채점자 KB의 평가영역 재해석, 채점자 KD의 일관성 결여가 부적합 채점경향의 원인임을 알 수 있었다. 그리고 채점자 NA, NB, NC, KE는 언어의 이해가능도를 중심으로 채점을 함으로써 서로 유사한 채점경향을 보였음이 밝혀졌다.

5. 요약 및 결론

이 연구의 목적은 중학교 영어교사들의 말하기 평가 채점경향을 분석하여 실제 채점상의 특징과 문제점을 알아보는 것이었다. 이러한 목적을 달성하기 위해 중학교 2학년 학생 18명이 4개 유형의 총 8개 문항으로 구성된 말하기 평가에 응시하였고, 말하기 능력 수준에 따라 세 집단으로 나누어진 8명의 영어교사들이 채점한 결과를 다국면 Rasch 모형에 근거한 FACETS 프로그램을 사용하여 분석하였다. 또한, 채점결과 분석과정에서 나타난 채점자들의 채점경향 특성을 보다 면밀히 파악하고자 채점자들에 대한 개별 인터뷰를 시행하였다. 이렇게 분석한 자료에서 도출된 결과는 다음과 같이 요약될 수 있다.

첫째, 채점자들은 엄격성의 편차가 대단히 심하며 서로 다른 기준을 적용하여 채점함으로써 낮은 채점자간 신뢰도를 보여주었다. 채점자내 신뢰도의 측면에서도 8명의 채점자 중 과적합이 2명, 부적합이 2명으로 확인되어 안정적이고 일관성 있는 채점이 이루어지지 못한 것으로 해석되었다. 특히 평가영역 적합도 지수에서 이해영역이 부적합 판정을 받아 채점자들이 이해영역을 제대로 이해하지 못하고 채점결과에서 큰 차이를 보임으로써 전체적인 채점자 신뢰도에 많은 영향을 미친 것으로 밝혀졌다.

둘째, 과적합 경향을 보인 채점자와 부적합 경향을 보인 채점자는 모두 한국인 채점자였으며 이들은 말하기능력에 상관없이 채점과정에서 채점오류나 일관성의 문제가 있는 것으로 드러났다. 원어민 채점자들은 엄격성이나 적합도 지수에서 상당히 일치하는 양상을 보였으나, 세부 평가영역 채점에서는 편차가 있음이 밝혀졌다.

셋째, 인터뷰 결과, 낮은 채점 자신감, 점수의 중앙집중경향 등이 과적합 채점경향의 원인이며, 평가영역 재해석, 채점의 일관성 결여 등이 부적합 채점경향의 원인인 것으로 밝혀졌다. 그리고 그 외의 채점자들은 채점기준적용에 있어서 수험자 발화의 이해가능도에 가장 큰 중점을 두었기 때문에 서로 비슷한 채점경향을 보인 것으로 해석되었다.

마지막으로, 인터뷰를 통해서 중학교 영어 말하기 수행평가로 타당한 말하기 도출방법이 사용되지 않고 대부분 암기해서 말하는 방법이 사용되고 있음을 알 수 있었다. 그리고 중학교 영어교사인 채점자들은 채점시 가장 우려되는 부분을 채점자 신뢰도에, 가장 힘든 부분을 시간 소요에 두고 있어 말하기 평가와 채점에 관한 전문성이 확보되지 않는 상황에서의 말하기 평가 시행의 어려움을 호소했다.

이상의 결과를 통해 중학교 영어교사들이 말하기 평가 채점에서 높은 편차

를 보이고 있으며, 일관성이 없거나 특정 점수대만을 사용함으로써 적절하고 변별력 있는 채점을 하지 못하고 있다고 말할 수 있다. 본 연구에서 나타난 분석 결과들과 현행 중학교 영어 말하기 평가의 여러 사항들을 고려해 볼 때, 다음과 같은 교육적 시사점을 제시할 수 있다.

첫째, 채점의 일관성을 확보하고 수험자의 능력수준을 구별할 수 있는 의미있는 채점이 이루어지도록 중학교 영어교사들을 대상으로 FACETS와 같은 프로그램을 활용한 채점자훈련을 실시할 필요가 있다.

둘째, 본 연구의 분석 대상자의 수가 적다는 한계가 있음에도 불구하고, 분석 결과 한국인 교사에 비해 원어민 보조교사 그룹의 채점경향은 서로 상당히 일치한다는 점은 주목할만한 사실이다. 따라서 원어민의 영어 말하기능력 판단기준과 한국인의 영어 말하기능력 판단기준의 차이를 연구하여 영어 말하기능력 평가에 적용해볼 필요가 있다.

셋째, 영어교육에 있어서 말하기능력은 지속적으로 강조되고 있으나 중학교에서 말하기능력을 평가하기 위한 인적, 물적 자원의 기반은 약하므로 체계적인 말하기 평가 시스템 구축이 선행되어야 할 것이다.

주제어: 다국면 Rasch 모형, FACETS, 채점자 신뢰도, 편향분석, 채점경향

참고문헌

신동일. 2001. 채점경향 분석을 위한 Rasch 측정모형 적용연구. *Foreign Languages Education,* 8(1), 249-272.

신동일, 설현수. 2005. NEW FACETS을 활용한 채점자료 분석방법. *Foreign Languages Education*, 12(2), 191-210.

장소영. 2001. 영어 말하기 능력의 이해: 채점자 변이로부터 평가 결과의 새로운 해석. 석사 학위논문. 숙명여자대학교 대학원.

장소영, 신동일. 2009. 「언어교육평가 연구를 위한 FACETS 프로그램: 기초 과정편」. 서울: 글로벌콘텐츠.

한문섭, 신동일. 2004. SEPT 채점 신뢰도와 타당도 연구. 「영어교육연구」16(2), 285-297.

Halleck, G. 1992. The oral proficiency interview: Discrete point test or a

measure of communicative language ability? *Foreign Language Annals*, 25, 227-231.
Jong, Y. K. 2011. Preparing Bilingual Pre-service Teachers for Bilingual Students in the US. 「언어과학」 18(1), 153-175.
Kim, Y. S. 1998. Conversation Analysis in Second Language Classroom Talk. 「언어과학」 5(2), 295-311.
Linacre, J. M. 1989. *Many-facet Rasch Measurement*. Chicago, IL: MESA Press.
McNamara, T. F. 1996. *Measuring Second Language Performance*. London: Pearson Education.
McNamara, T. F. & Adams. 1991. Exploring Rater Behavior with Rasch Techniques, ERIC Document Reproduction Service ED 345498.
North, B. 1993. *The Development of Descriptors on Scales of Language Proficiency*. NFLC Occational Papers. Washington DC: National Foreign Language Center at the Johns Hopkins University.
Savignon, S. 1985. Evaluating of Communicative Competence: The ACTEL Provisional Proficiency Guidelines. *Modern Language Journal*, 69, 129-134.
Stansfield, C. W., & Ross, J. 1998. A Long-term Research Agenda for the Tests of Written English. *Language Testing*, 5(2), 160-186.

백현영
609-735 부산광역시 금정구 부산대학로63번길 2
부산대학교 사범대학 영어교육과
전화번호: 010-4110-9997
전자우편: p-dogs@hanmail.net

양병곤
609-735 부산광역시 금정구 부산대학로63번길 2
부산대학교 사범대학 영어교육과
전화번호: 010-9618-7636
전자우편: bgyang@pusan.ac.kr
홈페이지: http://fonetiks.info/bgyang