말하기수행평가에서 발음범주 채점방식에 따른 채점신뢰도 분석[1]

이향 (이화여자대학교)

1. 서론

1990년대부터 수행 평가라는 용어는 종래의 평가 방식과는 구별되는 새로운 대안적인 평가라는 의미로 매우 포괄적으로 사용하기 시작하였다. 이로 인해 수행 평가라는 용어는 각 분야마다의 특성과 평가 목적에 따라 다양한 의미로 사용되어 대안적 평가(alternative test), 실제적인 평가(authentic test), 직접적 평가(direct test), 실기평가(performance-based test), 포트폴리오, 과정 중심의 평가 등의 용어와 혼용되어 사용되고 있다(Rothman, 1995; Reeves & Okey, 1996; 이준호, 2009). 이는 수행 평가가 가지는 성격이나 평가 방식이 이들 평가들과 공유하는 성격을 지녔기 때문일 수 있다. 그러나 사실 이들 용어는 수행 평가의 한 단면에 초점을 맞춘 용어들로 이들 용어가 수행 평가라는 용어를 대신할 수 있다고 보기에는 미흡하다. 한국어 말하기 수행 평가란 평가 도구의 과제를 통하여 유도된 수험자의 말하기 수행을 평가자가 평가 척도를 사용하여 채점하는 일련의 과정으로 정의할 수 있다(이향, 2012). 즉 수행 평가는 수험자가 목표어로 주어진 과제에서 요구하는 발화를 하는 과정과, 채점자가 주어진 채점 척도를 사용하여 점수나 등급을 부여하는 과정으로 이루어진다. 수행 평가는 평가 과제, 수험자의 수행과정 및 결과, 채점자, 채점 척도라는 모든 구성 요소가 평가 결과에 영향을 미칠 수 있으므로 수행 평가는 그 특성상 평가의 모든 단계에 걸쳐 신뢰도와 타당도를 갖춰야 한다. 그러나 이러한 신뢰도 확보 문제와 실용성 문제는 교육 현장에서 말하기 수행 평가를 망설이게 하는 요인으로 작용되어 왔다. 특히 말하기 수행 평가에 있어서의 과제의 신뢰도, 채점자간 신뢰도, 평가 영역의 신뢰도, 채점 방식의 신뢰도의 문제는 수행 평가에서 해결해야 하는 문제로 지적되어 왔다.(Bachman & Palmer, 1996; Fulcher, 1996, 2003; Hughes, 2003; Norris et al, 1998; Skehan, 1998)

본 연구는 한국어 말하기 수행 평가 범주 중 발음 능력 평가에 초점을 맞춘 연구이다. 이를 위하여 한국어 학습자들에게 컴퓨터 기반의 말하기 수행 평가를 실시하여 총체적 채점과 분석적 채점이라는 두 가지 방식으로 채점하도록 하였다. 그리고 그 채점 결과를 발음 평가 결과에 영향을 미치는 평가 과제, 평가 영역, 채점자, 채점 척도 국면으로 다국면 라쉬 모형(many-facet Rasch measurement)을 적용하여 채점 신뢰도를 분석하였다. 본 연구는 말하기 능력 중 일부인 발음 능력에 대한 평가 방식에 대한 연구이나, 궁극적으로 말하기 수행 평가를 실시하는데 있어서 신뢰성과 타당성을 확보하기 위한 객관적인 자료로 활용할 수 있을 것이라고 본다.

2. 다국면 라쉬 모형

다국면 라쉬모델은 언어 수행 수행 평가에서의 채점자나 과제 등의 측정에서의 영향을 측정하기 위하여 활발하게 적용되고 있는 모델로(Bachman, 2000) 이미 영어 교육에서는 이 모델을 사용하여 언어 수행 평가에서의 신뢰도를 분석하고 이를 활용하여 수행 평가에서의 신뢰도과 타당도를 높이기 위한 방안들에 대한 연구들이 활발히 진행되고 있다

1 본 연구는 이향(2012)에서 '일반화가능도 이론'을 사용하여 분석하였던 것과 같은 실험 데이터를 사용하여 '다국면 라쉬 모형'으로 분석한 것이다. 이향(2012)에서는 말하기 수행 평가에서의 발음 범주 평가의 채점 방법별 최적화 조건을 탐색해 보았으며, 본 연구를 통하여 평가에 영향을 미치는 국면들의 채점 신뢰도를 구체적으로 살펴보고자 하였다.

(신동일, 2002). 주로 언어 수행 평가에 있어서의 다국면 라쉬모델은 이미 사용되고 있는 평가 실행에 있어서의 채점자들의 간의 혹은 채점자 내의 신뢰도를 확인하는 연구(Brian et al, 1998; Kenyon, 1992, Brown & Abeywickrama, 2010; 신동일, 2001) 와 새로운 시험 개발이나 평가 모형 개발 과정에서 신뢰도와 타당도를 갖춘 채점 척도 설정을 위한 연구(Brian et al, 1998; Council of Europe, 2000; North, 2000; McNamara, 1996 ; 신동일, 2002)를 위하여 사용되어 왔다. 그러나 아직까지 한국어 교육에서는 이은하(2007)에서 학문 목적의 쓰기 수행 평가를 위한 분석적 채점척도를 개발하기 위하여 다국면 라쉬모델을 사용한 분석을 실시한 것이 거의 유일한 연구이다.

다국면 라쉬 모형(Linacre, 1989)은 문항반응이론(IRT: item response theory)을 바탕으로 일모수 로지스틱 모형인 라쉬 모형(Rasch,1980)을 확장하여 수행 평가의 결과에 영향을 미칠 수 있는 여러 가지 국면들을 분석하는 데 적합하도록 변형한 것이다. 이는 평가 도구의 등급 조절, 평가 영역 및 과제의 유형이나 수의 조정을 위한 타당화 근거 자료를 제시해 준다(장소영 & 신동일, 2009).

다국면 라쉬 모델에서는 모든 국면을 동시에 분석하여, 학습자, 채점자나 과제가 어떤 수준에 놓이게 되는지에 대한 확률을 로가리듬(logarithm)의 등간 척도(interval scale)를 사용한 로짓 범주(logit scale)로 나타내 주는데, 이는 '0'로짓을 중앙으로 상하 혹은 좌우로 +, − 값으로 나타난다. 또 각각의 국면들은 표준 오차(standard error)와 적절 통계치(fit statistics)를 고려하여 측정된다. 측정 결과는 점수가 아닌 객관적인 측정 범주인 로짓값으로 표현된다. 여기서 표준 오차는 평가와 관련된 있을 법한 오류들을 의미하며, 적절 통계치란 어느 정도까지 각각의 국면[2]들이 평가와 맞으며 다국면 라쉬 모델과 부합하는지를 나타낸다. 충분하지 못한 적절 통계치는 일관성에 결함이 있거나 혹은 평가의 결과가 부적합함을 나타낸다. 특히 이 모형은 채점자의 엄격성(severity)이나 관대함(leniency)을 로짓값으로 나타내 줌으로써 주관적인 채점의 신뢰도를 측정할 수 있게 해준다. 이 모델의 특성을 정리하면 다음 <표1>과 같다.

<표 1> 다국면 라쉬 모형의 분석(김성숙, 2001, p. 308)

	다국면 라쉬 모형
측정 이론 근거	잠재특성 모형(latent trait model)
연구 질문	· 측정상황의 오차를 배제한 후 개별 피험자 능력 추정값
연구 목적	· 개별 피험자 능력추정에 미치는 오차 효과 통제함 · 각 국면 효과 배제 후 능력을 추정하고자 함 · 다국면 측정 상황을 효율적으로 설명하는 모형을 탐색하고자 함
연구 설계/모형	· 국면: 모형(공식) 내 고유 변인으로 정의 · 모형: 주효과와 상호작용 효과
분석 결과	· 모형 내 모수치 추정값 · 각 효과의 신뢰도 및 카이제곱검정 · 모형 모수추정값 분포도
결과 활용	· 개별 피험자에 대한 독립적 선형정보 제공 · 오차 배제 후 피험자 능력 추정 정보 제공
제한점	· 다국면 모형의 상호작용에 대한 해석 불충분 · 분리신뢰도 계수 해석의 오해 가능성

2 여기서의 국면은 수험자의 수행에 영향을 주는 모든 측면 조건을 의미한다.

다국면 라쉬 모형의 이론적인 근거는 다음과 같다.

<표 2> 다국면 라쉬 모형의 이론적인 근거(Linacre, 1989)

log(Pnijk/Pnijk−1) = Bn − Di − Cj − Fk
Pnijk: 피험자 n이 말하기 과제 i에 대하여 j 평가자로부터 k의 점수를 받을 확률
Pnijk−1: 피험자 n이 말하기 과제 i에 대하여 j 평가자로부터 k−1의 점수를 받을 확률
Bn: 피험자 n의 능력 수준
Di : 문항 i의 난이도
Cj : 평가자 j의 엄격성
Fk : 점수 k−1에 비해 점수 k를 받기가 어려운 정도

이러한 다국면 라쉬 분석 모형을 통하여 수험자의 능력, 과제의 난이도, 채점자 엄격성 등 여러 요인들이 다국면으로 설정될 수 있으며 이들 요인들이 미치는 영향을 분석하여 통제할 수 있다. 특히 이 방법은 다양한 국면들은 동시에 또한 상호 동립적으로 분석하여 '동일한 척도'에 수량화 시킨 준거의 틀을 제공해 줌으로써 국면들 간의 비교를 가능하게 해 준다. 이러한 모수치에 대한 추정은 FACET 컴퓨터 프로그램을 사용하여 구할 수 있다. FACET 프로그램은 각각의 국면들의 종합적인 추정 결과뿐만 아니라 각각의 국면들에 대한 적합도 지수와 분리 신뢰도(reliability of separation) 등과 같은 타당도와 신뢰도를 추정할 수 있는 상세한 정보를 제공해 준다. 본 연구에서는 Facet 프로그램[3]을 사용하여 발음 채점 방식에 따른 각각의 국면들에 대한 신뢰도와 타당도를 비교해 보았다.

3. 실험 방법

3.1. 실험 참가자 및 채점자

본 연구에는 국내 언어 교육원에서 한국어를 배우고 있거나 졸업한 유학생 33명(N=33)이 참가하였으며 가급적 초급(한국어 학습 경험 3개월 이내)에서 고급 학생들(한국 대학 4학년 유학생)까지 고르게 구성되도록 하였다. 채점자는 국내 4년제 대학 내 언어교육원 이상의 기관에서 한국어 교육 경력 7년~ 15년인 한국어 교사들로 구성하였다.

<표 3> 채점자 배경

	한국어 교육 경험	학부	석사	박사
1	10년	국어국문	한국어 교육	통사론(국어학)
2	9년	국어국문	한국어 교육	없음
3	12년	불어불문	한국어 교육	없음
4	12년	일어일문	한국어 교육	음운론(국어학)
5	7년	국어국무	음운론	음운론(국어학)
6	15년	일어일문	한국어 교육	통사론(국어학)

3 Minifac(Facets Student/Evaluation) Version No. 3.68.1 Copyright ⓒ(c) 1987−2011, John M. Linacre. All rights reserved.

3. 2. 평가 도구와 평가 과제

본 실험은 'Brigham Young University(1999, 2000)'에서 제작한 'Enhance Oral Testing Software window version 1.1'을 사용하여 컴퓨터 기반 말하기 평가를 제작하였다. 본고에서 사용한 과제는 다음 <표 4>와 같다.

<표 4> 실험에 사용된 과제 유형

	과제 유형	내용	준비시간	수행시간
1	대화문 낭독하기	친구간의 다툼 대화 낭독하기	1분	2분
2	그림 보고 묘사하기	그림 보고 상황 묘사하기	1분	2분
3	서술하기	자기 가족 소개하기	1분	3분

말하기 평가 과제의 난이도가 높을 경우 피험자의 발음 능력을 평가하기에 충분한 양과 질을 갖춘 발화를 이끌어 낼 수 없으므로 본고에서는 일반적인 한국어 교육에서 초급에서 고급까지의 학습자들에게 적용할 수 있는 과제를 선정하였다. 과제1은 '낭독하기'과제로 이화여자대학교 「말이트이는 한국어1」의 대화문을 선택하였으며[4], 과제2는 '묘사서하여 말하기'로 이화여자대학교 「말이트이는 한국어1」의 하루일과 삽화하여 그림을 보고 가능한 자세히 하루일과와 그림 속의 상황을 자세히 묘사하도록 하였다. 과제3은 '서술하기'과제로 본인과 본인의 가족에 대하여 본인의 한국어 수준에 맞게 가능한 자세하게 발화하도록 하였다.[5]

3. 3. 채점 방식

채점자들은 위의 세 가지 과제들을 과제 별로 다음 두 가지 채점표를 사용하여 채점을 진행하도록 한다.

<표 5> (방법1)에 의한 분석적 채점 기준표

발음		
평가 영역	채점 기준	채점 척도[6]
분절음	정확도: 모국어 화자와 같은 정도	1 2 3 4 5 6
초분절음	정확도: 모국어 화자와 같은 정도	1 2 3 4 5 6

<표 6> (방법2)에 의한 총체적 채점 기준표

발음

4 피험자들이 이 교재를 사용한 경험이 있을 경우 평가 결과에 영향을 미칠 수 있으므로 피험자들이 본 교재를 사용한 적이 없음을 확인하였다.
5 이 모든 과정은 'Enhance Oral Testing Software window version 1.1'에서 평가자 툴을 사용하여 평가 문항과 녹음 파일을 입력하면, 피험자들은 수험자용 툴을 사용하여 평가 상황에 맞게 상호작용하도록 되어 있는 프로그램에 의하여 혼자 평가에 응하게 되어 있다.
6 6점이 가장 모국어 화자와 같은 정도가 큼을 의미한다.

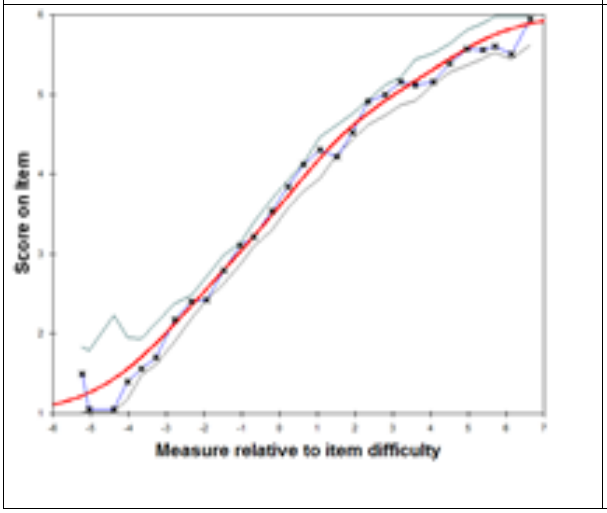
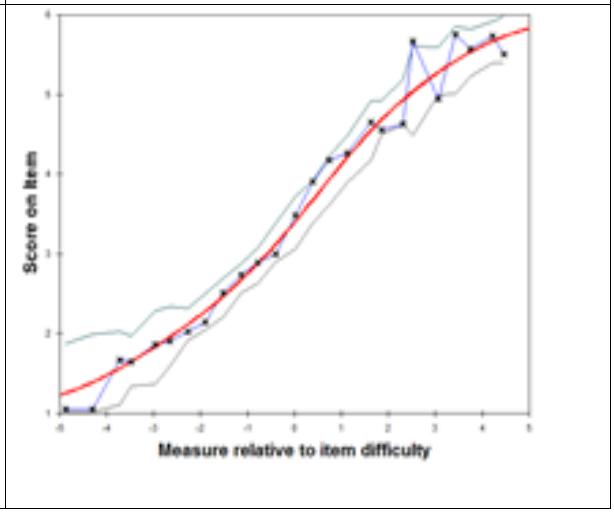
항목	채점 기준	채점 척도
발음	정확도: 모국어 화자와 같은 정도	1 2 3 4 5 6

첫 번째 채점표를 사용한 (방법1)은 분절음(개별음소, 음운변화)과 초분절음(억양), 두 영역으로 나누어 채점을 하는 분석적 방식이며, 두 번째 채점표를 사용한 (방법2)는 발음(분절음과 초분절음)영역, 한 영역으로 채점을 하는 총체적 방식이다.

4. 다국면 라쉬 모형에 근거한 채점신뢰도 검증

4. 1. 문항 특성 곡선에 의한 모형 적합도 분석

문항특성곡선은 다국면 라쉬 모형 이론에 근거한 문항반응 곡선과 실제 관찰된 자료에서 나타난 반응 빈도 간의 관계를 나타낸 곡선이다. 채점 자료가 라쉬 모형에 적합하다면 수집한 채점 자료가 라쉬 분석 모형에 적합하다는 것을 의미하며 이후의 분석 및 논의가 의미를 갖게 된다(신동일, 2006:147).

(방법1)	(방법2)
	

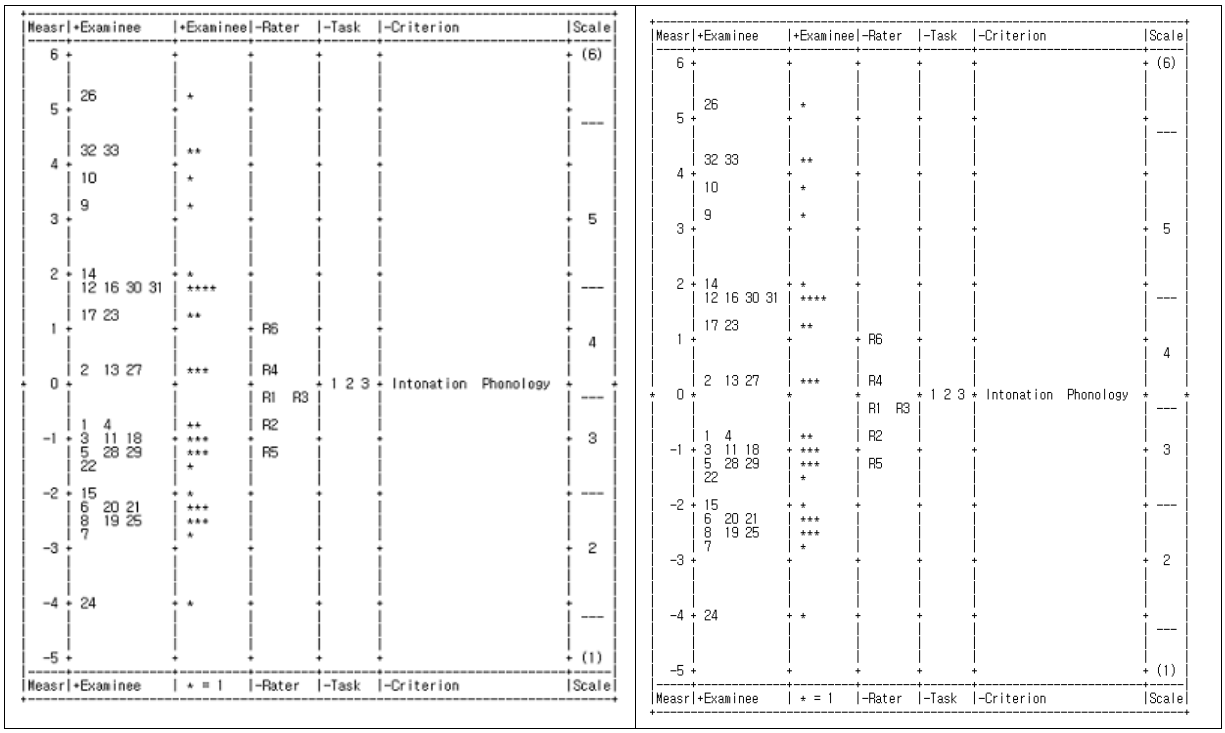
<그림 1> 문항 특성 곡선

<그림 1>을 보면 두 평가 방법 모두 모든 관찰치가 신뢰구간 안에 위치하고 있는 것으로 수집된 채점 자료가 다국면 라쉬 모형을 적용한 분석에 적합한 것으로 볼 수 있다.

4. 2. 측정 단면의 분포도 분석

본 연구에서는 말하기 평가에 영향을 미치는 국면으로 수험자, 채점자, 과제, 평가 영역의 네 가지 국면을 설정하였다. 먼저 수험자(Examinee)와 관련된 정보를 살펴보면 (방법1)은 −4~5로짓에 걸쳐 학생들이 분포하고 있는 반면 (방법2)은 −4~4로짓에 걸쳐 학생들이 분포되어 있는 것을 볼 수 있다.

(방법1)	(방법2)



<그림 2> 측정단면 분포도

구체적으로 살펴보면 (방법1)과 (방법2) 모두 같은 수험자의 말하기 수행 데이터를 채점한 결과임에도 불구하고 각각의 학습자 능력에 대한 추정치가 평가 방법에 따라 다른 것을 볼 수 있다. 예를 들어 (방법1)의 경우 26번 수험자가 가장 높은 능력을 지닌 수험자인 반면 (방법2)로 평가할 경우 32번 학생이 상대적으로 가장 높은 능력을 지닌 수험자로 평가되는 것을 볼 수 있다 . 이는 '발음' 능력을 측정하는데 있어서 사용된 평가 방법에 따라서 학습자의 능력이 다르게 평가될 수 있음을 보여 주며, 이는 발음 능력의 서로 다른 측면을 채점하거나 아니면 어느 한 방식이 신뢰할 수 없음을 의미하는 것이기도 한다.

다음으로 채점자 엄격성에 대한 정보인 4열을 살펴보면, 두 방법 모두에 '채점자6'은 가장 엄격한 채점자이며 '채점자5'는 가장 관대한 채점자로 나타났다. '채점자3'과 '채점자2'의 경우 (방법1)로 채점할 때와 (방법2)로 평가할 때 엄격성에 있어서의 차이를 보여 준다. 이 두 방법은 모두 채점자 엄격성에 있어서 2로짓 정도의 상대적인 차이를 보이는데 이는 가장 관대한 채점자가 지금 점수를 줄 확률이 50%라고 가정할 경우 가장 엄격한 채점을 하는 채점자가 이를 채점해서 똑같은 점수를 받을 확률이 12%밖에 되지 않음을 의미한다.[7]

다음으로 5열에는 과제의 난이도에 대한 정보가 척도 상에 나타나 있다. (방법1), (방법2) 모두 '과제1(낭독하기), 과제2(그림보고 말하기), 과제3(서술하기)'이 동일한 난이도를 갖는 것으로 나타났다. 이는 발음 능력을 측정하는데 있어서 세 과제들에 대하여 수험자들이 느끼는 어려움의 정도가 같음을 의미한다. 즉, 학생들이 '과제1'을 수행하는데 있어서 발음에 더 어려움을 느끼거나, '과제3'을 수행하는데 발음을 더 쉽게 하는 것과 같은 경우는 없음을 의미한다. 이는 본 실험에서 사용한 세 가지 유형의 과제가 발음

7 이 추정치는 McNamara(1996)의 책 6장에 제시된 로지트-확률 변환표(Wright와 Linacre1991)를 참고하였다. McNamara(1996)에 의하면 이는 심각한 결과이긴 하나 희귀한 사례는 아니며 실제 거의 대부분의 상황에서 이러한 평가 유형이 발생한다고 한다(참고:136).

능력을 채점하는 데는 난이도의 차이가 없음을 의미하는 것으로 볼 수 있다.

다음은 (방법1), (방법2)의 6열로 평가 영역의 곤란도에 대한 정보를 볼 수 있다. 이 곤란도는 수험자의 내재된 능력의 편차일 수 있으나 채점자가 각각의 영역에서 보이는 상대적인 엄격성에서의 차이 혹은 수험자가 느끼는 상대적인 수월함의 정도의 차이로 해석할 수 있다. 6열을 통하여 분절음과 초분절음 두 평가 영역 모두 같은 정도의 곤란함을 느낌을 알 수 있다. 이 두 영역 중 어느 것 하나가 더 수험자가 감당하기 쉽거나 어려운 차이는 없음을 알 수 있다.

각각의 방법에 마지막 열의 평가 척도(Scales)는 평가자가 채점할 때 사용한 척도의 관찰 범위를 수직 척도 상에 나타낸 것이다. 예를 들어 (방법1)의 경우 −2～0 로짓에 해당하는 학생들이 척도3에 해당하는 점수를 받았음을 의미하며, 0～+2로짓에 해당하는 학생들이 척도4의 점수를 받았음을 보여 준다. 이를 보면 '척도1, 2, 3, 4'는 상대적으로 동간의 척도로 평가한 반면 '척도5'는 좀더 넓게 사용되었음을 볼 수 있다.

FACET 프로그램은 이와 같은 기본 정보 뿐 아니라 각각의 국면의 단면에 대한 세부 정보도 제공되는데, 다음 절에서는 국면 별로 분리된 정보를 살펴봄으로써 본 평가의 신뢰성과 타당성에 대하여 더 자세히 살펴보도록 하겠다.

4. 3. 각 단면의 적합도 분석

4. 3. 1. 수험자 단면

적합도란 라쉬 모형에 의해 기대되는 점수와 실제 관찰된 점수를 비교한 것을 의미한다.[8] 각각의 평가 방법에 따른 수험자 국면의 출력 정보를 보면 과적합과 부적합에 해당하는 수험자들이 있음을 볼 수 있다.

<표 7> (방법1)의 수험자 국면 출력 정보

피험자 번호	능력 추정치	표준 오차	내적합		외적합	
			MnSq	ZStd	MnSq	ZStd
21	−2.18	0.23	0.56	−2.1	0.57	−2.1
23	1.15	0.24	3.30	6.4	3.18	6.0
16	1.71	0.26	1.82	2.7	1.78	2.6
9	3.27	0.30	1.59	2.1	1.60	2.1
평균(n=33)	0.00	0.25	0.98	−0.2	0.98	−0.3
표준편차	2.36	0.04	0.50	1.7	0.49	1.6
	신뢰도 지수	0.99				

8 일반적으로 내적 적합도 제곱 평균값이 1.0로짓을 기준으로 0.5～1.5 로짓 범위 안에 위치하면 적절한 수험자 반응으로 간주하고 1.5보다 크면 모형에 부적합(misfit)한 것으로 보며, 0.5보다 작은 지수의 수험자 반응은 과적합한 것으로 본다. 이는 표준화된 내적합도 값으로도 판단할 수도 있는데 +2이면 부적합, −2 이하이면 과적합한 것으로 판단할 수 있다. 그러나 본고에서는 적합도 판단 기준을 McNamara(1996)가 제안한 산출 방법을 사용하였다. 그는 적절한 적합도 지수는 n크기가 30명이나 이보다 많을 경우 FACET 출력 정보표 하단에 제시된 Mean Square 통계값에 대한 Standard Deviation 값의 2배의 ±값이라고 했다(p181). 이에 따르면 표 하단에 제시된 값들의 Mean±[2×S.D.]의 식으로 산출할 수 있다. 본 산출 방법을 사용하여 각각 (방법1)은 0.02～1.98, (방법2)는 −0.5～2.46를 기준으로 하였다.

<표 8> (방법2)의 수험자 국면 출력 정보

피험자 번호	능력 추정치	표준 오차	내적합		외적합	
			MnSq	ZStd	MnSq	ZStd
7	−2.58	0.37	0.29	−2.7	0.30	−2.7
19	−1.60	0.33	2.35	2.9	2.36	2.9
22	−1.50	0.33	0.24	−3.1	0.24	−3.2
5	−1.09	0.31	0.35	−2.5	0.40	−2.2
28	−0.55	0.28	0.34	−2.7	0.34	−2.6
23	1.38	0.29	0.39	−2.3	0.40	−2.3
9	2.02	0.32	3.25	4.3	3.41	4.5
26	2.80	0.35	3.36	4.5	3.53	4.8
평균(n=33)	.00	0.33	0.98	−0.4	1.0	−0.3
표준편차	1.94	0.04	0.73	1.8	0.76	1.8
	신뢰도 지수	0.97				

(방법1)은 '수험자23'이 3.3으로 부적합(misfit)한 수험자이며, (방법2)의 '수험자9, 수험자 26'이 부적합한 수험자로 확인된다. 이는 평가 과정이 이들 수험자들에게 맞지 않거나 검사 결과로 도출된 결론이 맞지 않을 수 있다는 정보로 이 수험자들에 대해서는 원 수집 자료를 분석해서 그 원인을 찾아 필요에 따라 재채점의 필요성 여부를 평가자들이 결정할 수 있다. 수험자 국면의 정보 중에 주로 이용되는 정보는 표 하단에 제공되는 신뢰도(Reliability)값으로 이는 과제별 난이도 분포가 수험자의 능력 분포를 적절하게 설명하고 있는지를 수치로 나타낸 값이다(신동일 2006, p. 153). 각각의 방법은 0.99, 0.97값으로 나타나고 있다. 이는 수험자의 능력 편차가 높은 것으로 볼 수 있으며, 수험자 관찰 변량의 대부분이 측정 오차가 아닌 수험자의 능력에 의한 편차로 발생했다 의미로 볼 수 있으다. 이들 각각의 평가 방법 중에서는 (방법1)을 사용했을 경우 상대적으로 수험자간의 능력 편차가 많고, (방법2)를 사용했을 경우 수험자간 편차가 상대적으로 적게 채점되었다는 것으로 이해할 수 있다.

4.3.2. 채점자 단면

다음으로 채점자 국면에 대한 정보로 채점자의 엄격성과 일관성에 대한 정보를 살펴보면 다음과 같다.

<표 9> (방법1)의 채점자 국면 출력 정보

채점자 번호	능력 추정치	표준 오차	내적합		외적합		Exct obs %	Agree Exp %
			MnSq	ZStd	MnSq	ZStd		
R5	−1.34	0.11	0.82	−1.8	0.88	−1.1	37.4	34.7
R2	−0.77	0.11	0.60	−4.5	0.69	−3.4	43.6	38.2
R1	−0.28	0.10	0.81	−2.0	0.80	−2.1	41.2	39.1
R3	−0.27	0.10	1.27	2.5	1.27	2.5	39.6	39.1
R4	0.16	0.10	1.10	0.9	1.08	0.8	40.3	38.0
R6	1.07	0.10	1.14	1.4	1.14	1.4	28.8	30.6
평균	−0.24	0.10	.96	−.6	.98	−.3		

표준편차	0.75	0.00	.23	2.4	.20	2.1		
카이제곱	308.0		d.f.	5	유의도		.00	
분리도	7.12				신뢰도 지수		0.98	
실제 일치도			38.5%		기대 일치도		36.6%	

<표 10> (방법2)의 채점자 국면 출력 정보

채점자 번호	능력 추정치	표준 오차	내적합		외적합		Exct obs %	Agree Exp %
			MnSq	ZStd	MnSq	ZStd		
R5	−.48	.13	.62	−3.0	.62	−3.0	40.2	36.5
R3	−.41	.13	.56	−3.6	.62	−3.0	42.6	36.9
R2	−.33	.13	.48	−4.5	.51	−4.2	44.8	37.1
R1	−.21 .	.14	2.59	7.8	2.63	8.1	29.9	37.4
R4	32	.14	.81	−1.3	.80	−1.5	39.6	36.9
R6	0.92	0.14	0.81	−1.3	0.84	−1.1	30.7	32.9
평균	−0.03	0.14	0.98	−1.0	1.00	−0.8		
표준편차	0.50	0.00	0.73	4.1	0.73	4.1		
카이제곱		79.1	d.f.	5	유의도		.00	
분리도		3.54			신뢰도 지수		0.93	
실제 일치도			38.0%		기대 일치도		36.3%	

능력 추정치(Measure)는 관찰된 점수를 다국면 라쉬 모형 이론에 적용하여 나온 값으로 채점자들의 엄격성을 나타내 값이다. 앞의 측정 단면 분포도에서도 확인했듯이 '채점자3'을 제외하고는 '채점자6, 채점자4, 채점자1, 채점자2, 채점자5' 순으로 채점자의 엄격성의 차이가 있는 것을 볼 수 있다. '채점자6'은 가장 엄격한 채점자이며 '채점자5'는 가장 관대한 채점자인 것이다. '채점자3'의 경우 (방법1)로 채점할 때 (방법2)로 평가할 때 보다 더 엄격한 채점을 했음을 볼 수 있다. 표의 하단에 제공되는 분리도(Separation)는 채점의 엄격성의 정도의 분포를 보여 주는 값으로 (방법1)은 7.12, (방법2)는 3.54, 값으로 나타났다. 이는 평가(방법1)로 채점할 때가 (방법2)로 채점할 때 보다 채점자들 간의 엄격성의 차이가 크다는 것을 의미한다. 신뢰도(Reliability)는 채점자들의 엄격함에 있어서의 차이를 나타내는데 각각 0.98, 0.93으로 나타났다. 신뢰도(Reliability) 값으로 보듯 채점자들 간의 엄격성에 차이가 있음을 다시 한 번 확인할 수 있다. 카이스퀘어(Chi−square) 검증으로도 채점자들의 엄격성이 모두 같다는 영가설(null hypothesis)를 검증한 결과 모두 p=.00으로 영가설을 기각하고 채점자 집단의 엄격성 차이가 존재한다는 것을 알 수 있다. 그리고 여기서 하나 더 주목해 볼 것이 추정 오차(Model S.E.)이다. 이는 평가자의 엄격성에 존재하는 실제적인 변량인데 (방법2)가 (방법1)보다 상대적으로 추정 오차가 늘어남을 볼 수 있다. 이는 높은 수치는 아니나 상대적으로 다른 평가 방법에 비하여 (방법2)를 사용하였을 때 엄격성의 차이 이외의 다른 추정오차에 의하여 설명되어야 하는 부분이 늘어난다고 볼 수 있다.

채점자 국면 출력 정보로는 평가가 타당성을 갖추는데 필요 조건이 되는 채점에 있어서의 일관성 즉 신뢰도에 대한 정보를 알 수 있다. 이는 적합도(Fit)에 대한 내적 적합도 제곱평균(Infit MnSq), 표준화된 내적 적합도(Infit ZStd), 외적합도 제곱평균(Outfit MnSq), 표준화된 외적 적합도(Outfit ZStd)에 대한 정보로 알 수 있는데, 이는 채점된 관찰값과 라쉬 모형으로 예측한 값 사이의 적합도로 채점의 일관성을 판단할 수 있게 해준다.

여기서는 McNamara(1996)가 제안한 적합도 산출방법을 사용하여 각각 (방법1)은

0.5~1.42, (방법2)는 −0.48~2.54 사이의 값을 그 기준으로 보았다. 그 결과 (방법2)에서 '채점자1'이 2.59로 부적합으로 나타났다. 부적합 채점자는 채점을 하는데 있어서 채점자의 채점 경향이 모형이 예측한 것과 다르다는 것을 의미하며 일관성이 결여 됐음을 의미한다. 그러므로 이 경우 세부적인 편향 분석이나 원 자료와의 비교 분석을 통하여 그 원인을 찾아볼 필요가 있으며 채점자 훈련을 통하여 일관성을 갖도록 교육을 시키는 것도 고려해 봐야 할 것이다.

또한 이 제곱평균 값은 1의 기대값을 갖기 때문에 평균 1로부터 얼마나 떨어져 있는지를 통하여 모형이 예측한 값과의 차이를 알 수 있는데 (방법1), (방법2)의 '채점자4'와 '채점자6'이 상대적으로 더 모형이 예측한대로 일관성 있는 채점을 하고 있다고 볼 수 있다. 이러한 각각의 채점자들의 일관성에 대한 구체적 일치도를 실제 관찰 일치도(Exact Obs)와 모형에 의해 기대되는 일치도(Agree Exp)의 차이로 살펴보면 모두 모형에서 기대한 일치도 보다 높음을 알 수 있다. 예를 들어 (방법1)로 '채점자1'에게 기대되는 일치도는 39.1%인데 실제 관찰값에 의한 일치도는 41.2%임을 볼 수 있다. 모든 평가 방법에 있어서 다른 채점자들도 마찬가지로 기대되는 일치도 보다도 실제 관찰값 일치도가 높음을 확인할 수 있다. 각각의 평가 방법들의 모형 전체에 대한 모든 채점자들의 일치도를 확인해 보면 (벙법1)은 36.6%, (방법2)는 36.3%의 일치도를 라쉬 모형 이론에 근거하여 기대하였는데 실제 관찰 일치도가 (방법1)은 38.5%, (방법2)는 38%의 일치도를 보인 것을 볼 수 있다. 이 수치를 활용한 라쉬모형 전체 일치도 통계치를 산출해 보면[9] (방법1)은 0.03로짓, (방법2)은 0.3로짓으로 산출된다. 이는 채점자간 일치도 통계로서 개인의 엄격성 혹은 일관성 지수와는 구별되는 정보이다(신동일, 2006: 155). 여기서는 (방법1)이 (방법2)보다 0에 가까운 수치를 보여 (방법1)이 모형이 기대한 값과 실제 채점자들 간의 채점이 상대적으로 더 일치하고 있는 것으로 나타났다.

4. 3. 3. 과제 단면

다음으로 과제 국면을 살펴보도록 하겠다.

<표 11> (방법1)의 과제 국면 출력 정보

과제 번호	능력 추정치	표준오차	내적합		외적합		Correlation	
			MnSq	ZStd	MnSq	ZStd	Pt Mea	Pt Exp
1	−0.07	0.07	0.91	−1.2	0.94	−0.8	0.86	0.85
3	0.02	0.07	1.05	0.7	1.05	0.6	0.85	0.85
2	0.05	0.07	0.92	−1.2	0.94	−0.8	0.86	0.86
평균	0.00	0.07	0.96	−0.6	0.98	−0.3	0.86	
표준편차	0.05	0.00	0.06	0.9	0.05	0.7	0.01	
카이제곱	1.6		d.f.	2	유의도		0.45	
분리도	0.00				신뢰도 지수		0.00	

<표 12> (방법2) 의 과제 국면 출력 정보

과제 번호	능력 추정치	표준오차	내적합		외적합		Correlation	
			MnSq	ZStd	MnSq	ZStd	Pt Mea	Pt Exp

9 출력 정보 하단의 [(실제 일치도 퍼센트−기대 일치도 퍼센트)/(100 − 기대 일치도 퍼센트)]의 공식으로 산출한다(신동일, 2006:155).

3	−0.07	0.10	0.98	−0.1	1.00	0.0	0.84	0.84
2	0.02	0.10	0.94	−0.5	0.98	−0.1	0.86	0.84
1	0.05	0.10	1.02	0.2	1.03	0.3	0.84	0.84
평균	0.00	0.10	0.98	−0.2	1.00	0.1	0.84	
표준편차	0.05	0.00	0.03	0.3	0.02	0.2	0.01	
카이제곱	0.9		d.f.	2	유의도		0.62	
분리도	0.00				신뢰도 지수		0.00	

과제별로 난이도에 있어서의 큰 차이는 없으나 세밀하게 본다면 (방법1)로 분절음과 초분절음 두 가지 영역으로 채점할 경우 '과제1, 과제3, 과제2' 순으로 과제 곤란도가 높아지는 것을 확인할 수 있다. 또한 (방법2)의 경우 '과제3, 과제2, 과제1' 순으로 과제 난이도가 높아지고 있다. 그러나 표 하단에 제시된 분리도와 신뢰도 두 방법 모두 편차가 0.00으로 과제 난이도에 있어서 편차가 없는 것을 볼 수 있다. 이는 두 방법으로 평가한 세 가지 과제가 미세한 차이는 있으나 라쉬 모형 전체로 볼 때 거의 난이도에 있어서의 차이가 없음을 보여주는 것으로 수험자가 세 가지 과제에서 상대적으로 점수를 받기 쉽거나 어려운 과제는 없고 모두 비슷하게 어려운 과제들임을 알 수 있다.
또한 오차(Model S.E.)는 (방법2)가 (방법1) 보다 상대적으로 변량이 큰 것을 보여 준다. 과제 적합도를 살펴보면 (방법1)과 (방법2) 모두 기준 범위 사이에 위치하고 있는 것으로 나타났다. 이는 라쉬 모형으로 예측된 자료와 실제 관찰된 자료 사이의 불일치가 되는 경우가 없음을 의미한다. 또한 이로 각각의 과제들이 변별력 있게 평가에 사용되고 있음을 의미하는 것으로 볼 수 있으며 각각의 과제들이 평가하고자 하는 요인에 대한 서로 다른 정보를 제공해 주고 있음을 의미하는 것으로 볼 수 있다.

4. 3. 4. 평가 영역 단면

다음으로 평가 방법 별로 평가 영역이 어떻게 작용하고 있는가를 살펴보면 다음과 같다.

<표 13> (방법1)의 평가 영역 단면 출력 정보

평가 영역	능력 추정치	표준 오차	내적합		외적합		Correlation	
			MnSq	ZStd	MnSq	ZStd	Pt Mea	Pt Exp
초분절음	−0.10	0.06	1.01	0.2	1.02	0.3	0.86	0.85
분절음	0.10	0.06	0.91	−1.6	0.93	−1.2	0.86	0.86
평균	0.00	0.06	0.96	−0.7	0.98	−0.4	0.86	
표준편차	0.10	0.00	0.05	0.9	0.04	0.8	0.00	
카이제곱	5.6		d.f.	1	유의도		0.02	
분리도	1.35				신뢰도 지수		0.65	

<표 14> (방법2)의 평가 영역 단면 출력 정보

평가 영역	능력 추정치	표준 오차	내적합		외적합		Correlation	
			MnSq	ZStd	MnSq	ZStd	Pt Mea	Pt Exp
발음	0.00	0.06	0.98	−0.3	1.00	1.03	0.84	0.84
평균	0.00	0.06	0.98	−0.03	1.00	0.1	0.84	

표준편차	0.00	0.00	0.00	0.0	0.00	0.0	0.00

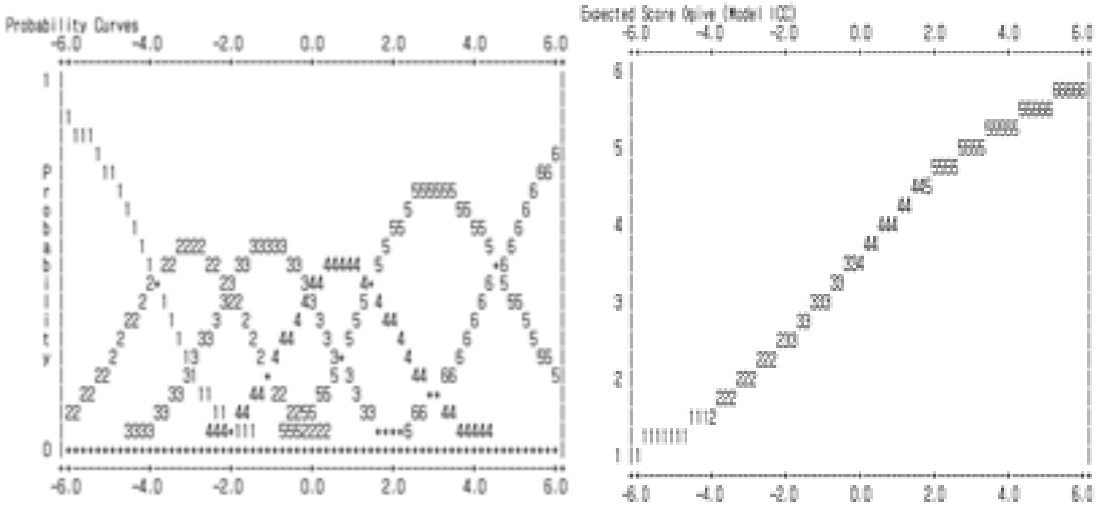
(방법1)의 능력 추정치로 두 영역의 곤란도를 비교해 보면 '분절음' 영역이 0.1, '초분절음' 영역이 −0.1로 초분절음 영역 보다 분절음 영역이 상대적으로 점수를 받기 곤란한 영역임을 보여 준다. 이는 '분절음' 영역이 좀 더 엄격하게 채점되고 '초분절음' 영역이 좀 더 관대한 채점 경향을 보인다는 것을 의미하기도 한다. 평가 영역에 있어서의 적합도를 살펴보면 (방법1)의 '분절음' 영역은 0.91, '초분절음' 영역은 1.01로, (방법2)의 '발음' 영역은 0.98로 나타났다. 이는 라쉬 모형이 예측하는 대로 평가 영역이 작용하고 있음을 의미한다. 이들 영역이 모두 적합도 기준에 적절한 것으로 보아 평가하고자 하는 평가 영역이 설득력이 있음을 보여 주어 타당한 구인으로 작용하고 있음을 볼 수 있다.

4. 3. 5. 평가 척도 단면

채점에서 사용한 6개의 평가 척도의 분석 결과를 살펴보면 다음과 같다.

<표 15> (방법1)의 평가 척도 단면 출력 정보

채점 척도	척도별 데이터			QUALITY CONTROL			
	관찰치	%	누적	Avge Meas	Exp. Meas	OUTFIT Meas	Cat Peak Prob
1	57	5	5	−3.18	−3.13	1.0	100%
2	200	17	22	−2.16	−2.11	1.0	54%
3	324	27	49	−1.07	−1.04	0.8	56%
4	249	21	70	0.55	0.47	0.8	50%
5	263	22	92	2.64	2.53	1.0	71%
6	95	8	100	4.23	4.53	1.3	100%



<그림 3> (방법1)의 6점 척도의 확률 곡선과 모형 특성 곡선

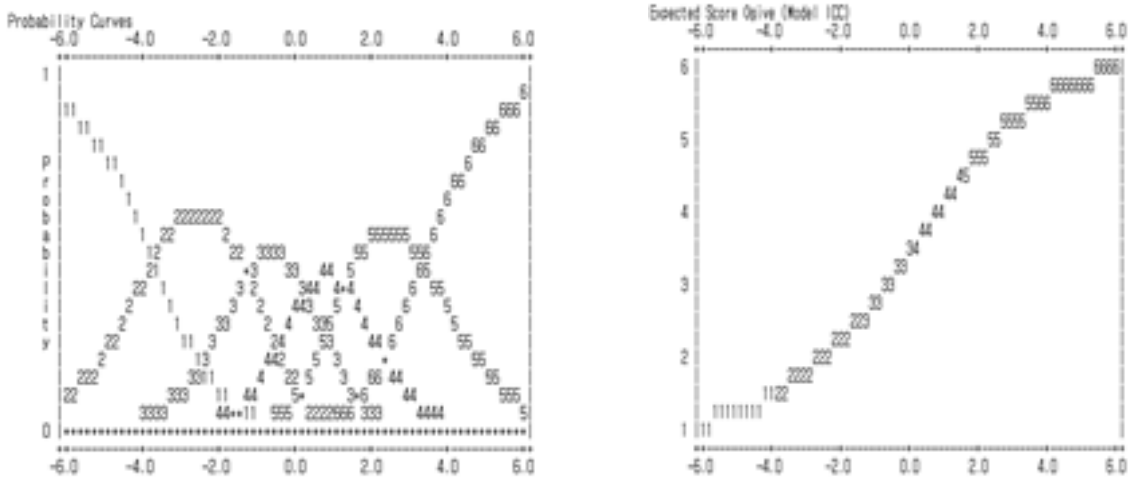
(방법1)의 6등급 척도 사용을 분석해 보면 3점 척도 27%, 5점 척도를 22%, 4점 척도를 21%, 2점 척도를 17%, 6점 척도를 8%, 1점 척도를 5% 사용한 것으로 나타났다. 또한 높은 척도일수록 평균 측정값(Avge Measure)이 증가하면 각각의 척도가 적절하게

기능하고 있는 것으로 볼 수 있는데, 1에서 6점 척도로 갈 수록 각각 '−3.18, −2.16, −1.07, 0.55, 2.64, 4.25'로 증가하는 것을 볼 수 있다. '6점 척도의 확률 곡선'을 보면 Y 축 맨 위에서 시작하는 1점 곡선부터 마지막 6점 곡선까지 각각의 점수를 받을 수 있는 수험자들의 확률 곡선이 겹치지 않고 이동되고 있는 것을 확인할 수 있는데 이는 각각의 척도가 적절하게 기능하고 있음을 보여 주는 것이다. 또한 측정치 변화에 따른 6점 척도의 기대 점수를 '모형특성곡선'으로 확인해 보면, Y축은 등급 척도이며, X축은 관찰값으로 측정치가 오른쪽으로 증가할수록 척도가 증가하고 있는 것을 볼 수 있다. 1에서 6점 척도까지 척도가 겹치지 않고 이동되고 있음을 볼 수 있는데 이는 척도간의 등급차이가 타당하게 존재하고 있다는 것을 의미한다. 그러나 1−4점은 수험자들의 능력이 비교적 등간으로 측정되고 있는 반면 5점 척도와 같은 경우 수험자 상대적으로 더 넓은 구간으로 사용되고 있음을 확인할 수 있다. 이는 5점 척도를 조금 더 세분해 볼 수 있음을 시사해 주는 것으로 볼 수 있다.

<표 14> (방법2)의 평가 척도 단면 출력 정보

채점 척도	척도별 데이터			QUALITY CONTROL			
	관찰치	%	누적	Avge Meas	Exp. Meas	OUTFIT Meas	Cat Peak Prob
1	31	5	5	−2.87	−2.82	1.0	100
2	141	24	29	−1.68	−1.67	1.4	62
3	154	26	55	−0.69	−0.66	0.8	50
4	110	19	73	0.76	0.60	0.7	43
5	108	18	92	1.85	1.96	1.2	57
6	50	8	100	3.32	3.26	0.9	100

<그림 4> (방법2)의 6점 척도의 확률 곡선과 모형 특성 곡선



총체적 방법인 (방법2)를 사용한 <표 14>와 <그림 14>를 보면 (방법2)의 6등급 척도 사용을 분석해 보면 3점 척도 26%, 2점 척도를 24%, 4점 척도를 19%, 5점 척도를 18%, 6점 척도를 8%, 1점 척도를 5% 사용한 것으로 나타났다. 또한 1에서 6점 척도로 갈수록 평균 측정값(Avge Measure)도 각각 '−2.87, −1.68, −.069, 0.76, 1.85, 3.32'로 증가하는 것을 볼 수 있다. <그림 4> '6점 척도의 확률 곡선'으로 척도의 타당성을 살펴보면 수험자들의 확률 곡선이 겹치지 않고 이동되고 있는 것을 확인할 수 있는데 이는 각각의 척도가 적절하게 기능하고 있음을 보여 주는 것이다. '모형특성곡선'을 보면 1에서 6점 척도까지 척도가 겹치지 않고 이동되고 있음을 볼 때 척도간의 등급차이가 타당

하게 존재하고 있다는 것을 알 수 있다.

5. 결론

본고에서는 세 가지 유형의 과제로 컴퓨터 기반의 말하기 수행 평가를 실시하여 6명의 채점자가 피험자의 발음 능력을 '분절음과 초분절음'이라는 두 가지 영역으로 채점하는 분석적 방식과 '발음(분절음과 초분절음 포함)'영역 하나로 채점하는 총체적 방식으로 채점하여 채점 신뢰도를 비교해 보았다.

그 결과 평가 방법에 따라 수험자의 능력이 다르게 추정되는 것을 발견하였다. (방법1)이 (방법2)보다 상대적으로 신뢰도가 높았으며 이는(방법2)의 경우 두 가지 측면 중 어느 측면에 채점자가 더 초점을 맞추는가에 따라 채점자간 채점 결과가 달라진 것으로 추정해 볼 수 있다. 그러나 이는 추정일 뿐 보다 향후 근본적인 원인 분석을 통하여 평가 방법을 선택하는데 도움을 줄 수 있는 연구가 필요하다. 다음으로 평가 과제 유형은 발음 능력을 채점하는데 있어서 난이도의 차이가 없는 것으로 나타났다. 평가 영역의 경우도 (방법1)의 경우 분절음과 초분절음 두 영역 간에 대한 수험자가 느끼는 곤란도에는 차이가 없는 것으로 나타났으며 분절음이 초분절음 보다 상대적으로 엄격한 채점을 하는 것으로 나타났다. 또한 (방법1)의 두 영역과 (방법2)에서의 '발음' 영역 모두 발음 능력을 채점하는데 적절한 구인으로 작용하고 있음을 확인하였다. 또한 채점자 국면을 살펴본 결과 두 방식 모두 채점자 내, 채점자 간 신뢰도는 높은 것으로 나타났으며, (방법1)이 (방법2)보다 상대적으로 엄격성의 차이가 크고, (방법2)가 (방법1)보다 추정 오차가 큰 것으로 나타났다. 또한 채점자 엄격성에 있어서의 차이는 큰 것으로 나타났다. 그러므로 보다 객관적인 발음 평가를 위해서는 채점자들 간의 엄격성의 차이의 원인을 규명하고 엄격성의 차이를 고려한 적절한 조치를 취하기 위한 연구들이 진행되어야 할 것이다. 마지막으로 채점에 사용한 6점 척도 또한 각각의 점수가 의미 있게 작용하고 있음을 확인하였다. 그러나 (방법1)의 경우 5점 척도가 사용되는 구간에서 상대적으로 넓은 수험자의 능력을 채점하고 있는 것으로 나타나 이를 좀 더 세분할 수 있음을 확인할 수 있다.

지금까지 본고에서는 말하기 수행 평가 중 '발음 범주'의 채점에 한하여 연구를 진행하였으나 향후 한국어 교육에서의 말하기 수행 평가의 실시와 보급을 위해서는 말하기 능력의 모든 범주에 대한 여러 국면의 분석과 그 의미를 도출하는 연구가 활발히 진행되어야 할 것이다.

참고문헌

김성숙. 2001. 채점의 변동요인 분석방법에 대한 고찰: 일반화가능도 이론과 다국면 라쉬 모형의 적용과 재해석. 교육평가연구14(1), 303−325.

신동일. 2001. 채점 겨향 분석을 위한 Rasch 측정모형 적용 연구. 외국어 교육8(1), 249−272.

신동일. 2002. Rasch 모형을 이용한 고등학교 영어과 말하기 및 쓰기능력 등급 기술표 개발. 영어교육57(4), 469−499.

신동일. 2006.「한국의 영어 평가학2: 말하기 시험편」. 서울: 한국문화사

이은하. 2007. 학문 목적 한국어 학습자의 쓰기 수행 평가를 위한 분석적 채점척도 개발. 석사학위논문. 이화여자대학교.

이준호. (2009). *한국어 수행 평가의 원리 및 방안 연구.* 박사학위논문, 고려대학교.

이화여자대학교 언어교육원. 2009. 「말이 트이는 한국어1」, 서울: 이화여자대학교 출판부.

이향. 2012. 말하기 수행 평가에서 발음 범주 채점의 최적화 방안 연구−일반화가능도 이론을 활용하여−. 한국어교육23(2), 301−330.

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

Bachman, L. F., & Palmer, A. S. 1996. *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F. 2000. Modern language testing at the turn of the century: assuring that what we count counts. Language Testing17(1), 1-42.

Brian, K., Lynch, B. K., & McNamara, T. 1998. Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. Language Testing15(2), 158-180.

Brown, H. D., & Abeywickrama, P. 2010. *Language assessment : principles and classroom practices*. White Plains, NY: Pearson/Longman.

Europe, C. o. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Press Syndicate of the University of Cambridge.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge Univ Press.

Fulcher, G. 1996. Invalidating validity claims for the ACTFL oral rating scale. System24(2), 163-172.

Fulcher, G. 2003. *Testing second language speaking.* Harlow: Longman.

Hughes, A. 2003. *Testing for language teachers*. Cambridge: Cambridge Univ Press.

Kenyon, D. M. 1992. Introductory remarks at symposium on Development and use of rating scales in language testing. Paper presented at the 14th Language Testing Research Colloquium, Vancouver.

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

Norris, J. M., Brown, J. D., Hudson, T. D., & Yoshioka, J. 1998. *Designing second language performance assessment*s. Honolulu: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

North, B. 2000. *The development of a common framework scale of language proficiency*. NewYork: Peter Lang.

Rasch, G. 1980. *Probabilistic models for some intelligence and attainment tests* Chicago: University of Chicago Press

Reeves, T. C., & Okey, J. R. (Eds.). 1996. Alternative assessment for constructivist learning environments. New Jersey: Educational Technology.

Rothman, R. 1995. *Measuring up: Standards, assessment, and school reform*. San Francisco, CA: Jossey-Bass

Skehan, P. 1998. *A cognitive approach to language learning*. Oxford: Oxford University Press.

Wright, B. D., & Linacre, J. M. 1991. *A user's guide to BIGSTEPS Rasch-model computer programs version 2.2*. Chicago IL: MESA Press.